New Edge-to-Cloud Orchestrator for Intelligent Task Allocation and Efficiency Optimisation

Sergio Comella
Data & Analytics R&I
Engineering Ingegneria Informatica
Palermo, Italia
sergio.comella@eng.it

Delia Milazzo
Data & Analytics R&I
Engineering Ingegneria Informatica
Palermo, Italia
delia.milazzo@eng.it

Salvatore Cipolla
Data & Analytics R&I
Engineering Ingegneria Informatica
Bologna, Italia
salvatore.cipolla@eng.it

Ons Aouedi
Associate Researcher
University of Luxemburg
Luxemburg, Luxemburg
ons.aouedi@uni.lu

Susanna Bonura
Data & Analytics R&I
Engineering Ingegneria Informatica
Palermo, Italia
susanna.bonura@eng.it

2025 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC) | 979-8-3315-8534-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICE/ITMC65658.2025.11106606

***Abstract*—Modern industrial systems face several challenges in resource allocation, operational efficiency, and scalability, driven by the need for real-time data processing, and sustainable automation. The integration of advanced technologies, including edge-to-cloud (E2C) Artificial Intelligence (AI), digital twins, and blockchain, has created opportunities for innovation while increasing the complexity of orchestration and task distribution. This paper presents a novel E2C orchestrator that optimises task distribution across hybrid infrastructures, addressing the limitations of existing solutions such as static resource allocation and energy inefficiency. By leveraging AI-driven dynamic policies, the orchestrator increases efficiency, reduces latency and ensures robust data protection, in line with Industry 5.0's paradigm which emphasises human-centric and sustainable industrial ecosystems. Key outcomes include improved scalability, adaptability to dynamic workloads, and integration of advanced technologies for real-time responsiveness and operational visibility.**

E2C Orchestrator, AI, Industry 5.0, scalability, resources optimisation, energy efficiency, sustainability.

I. Introduction

The industrial scenario is experiencing significant change due to the ever-increasing demand for greater efficiency and scalability from a sustainable perspective. Challenges such as the need for real-time responsiveness, significantly reduced energy consumption and increased security in increasingly complex environments are emerging. The accelerated adoption of advanced technologies, such as edge-to-cloud (E2C) computing, digital twins, and blockchain, is a contributing factor to this transformation. However, it is important to note that while these innovations hold great potential, they are also introducing new obstacles, particularly regarding resource allocation, task coordination, and operational visibility (1).

Current orchestration solutions, such as Kubernetes and Docker Swarm, contribute to provide support for managing containerised applications even if their static resource allocation methods and limited adaptability often fall short of meeting the dynamic demands of hybrid systems that integrate edge and cloud computing (1). As industries aim to adopt solutions granting sustainability, human-centered design and ethical use of technology, the need for more advanced orchestration frameworks becomes increasingly urgent.

This paper presents a new E2C orchestrator, designed to address the abovementioned challenges by facilitating intelligent task allocation and optimising resource management. The proposed solution has been designed for dynamic industrial environments with the aim of guaranteeing efficiency, scalability and security while supporting the transition to more sustainable and human-centred industrial practices. Moreover, this work aims to enable smarter, more adaptive and trustworthy operations, aligning innovation with the broader goals of sustainability and ethical responsibility.

The next session will present industry challenges in resource allocation.

II. Industry Challenges in Resource Allocation and Efficiency

Rapid modern industrial systems development poses various challenges in the areas of resource allocation and improvement of operational efficiency. Integrating emerging technologies such as E2C Artificial Intelligence (AI), blockchain, and digital twins into industrial systems necessitates designing new strategies for optimisation of performance, scalability, and security. With the diversification of industrial applications, the ability to reconcile computational capabilities, energy efficiency, and the need for explainable artificial intelligence (XAI) becomes critical for sustainable automation. Most challenges in this regard are the most critical factors that will drive innovation and assure sustainable advanced industrial ecosystems in the long term.

A. Towards Intelligent Industrial Systems

Industries are quickly adopting new technologies like edge computing and distributed AI, among others, in an attempt to deal with increasing requirements concerning the processing of real-time data and quicker decision-making capabilities that answer to increased responsiveness. In their strategy for positioning computation closer to the point where data originates, priority is given to solutions; as a result, more advanced AI-enabled edge devices, which are not necessarily constrained by conventional storage or processing limitations, become feasible.

Contemporary industrial platforms, therefore, are developing from the traditional centralised systems toward more responsive and adaptive architectures by incorporating

intelligent sensors and adaptive systems. This will give them the ability to analyse and respond to data in real-time, hence making operations agile and efficient.

These technological advancements are implemented in applications such as autonomous robotics, virtual reality, and the Internet of Things (IoT). Industries can now create greater flexibility, scalability, and efficiency while simultaneously reducing latency and saving energy thanks to decentralization. All this has been happening around the ideals of Industry 5.0, where a human-centered approach to design and environmental sustainability in solutions are central.

The emphasis on edge-based intelligence brings not only improved performance but also energy efficiency, contributing to greener industrial practices. However, implementation of such a system calls for novel solutions in order to fulfill diverse needs and ensure long-term sustainability (2).

B. Resource Allocation Challenges

Resource allocation in Industry 5.0 paradigms is complex and challenging because of the different, often conflicting needs of modern industrial applications. These demands include the need for high accuracy, robustness, scalability, and security and privacy protection. A major challenge lies in how to properly divide the computational load between the edge and cloud resources to maximise the system's performance subject to the different design constraints for each application. Attaining this objective necessitates the agile selection of datasets, algorithms, and models that are most appropriate for the specific requirements of each scenario, thereby guaranteeing maximum efficiency without compromising the overall performance of the system (2).

Moreover, in addition to enhancing resource allocation, it is essential to address the challenges associated with reducing learning latency and augmenting system adaptability. Recent advances in semantic AI, multi-task learning, and transfer learning have been identified as promising approaches to improve reusability and reduce computational requirements, thereby increasing energy efficiency and shortening model training time (3). These techniques enable industrial systems to adapt rapidly to new situations and optimise processes without compromising performance. Besides, it has become integral to incorporate blockchain to ensure complete security and privacy based on a powerful framework that guarantees protection of the most sensitive info to enable greater trust across the industrial networks. Finally, there is a need to ensure the transparency and explainability of AI-driven systems if they are to be widely adopted in industrial environments. Solutions such as digital twins and human-in-the-loop methods are increasingly being seen as effective in explaining the decision-making processes of AI, thus making the activities of AI more understandable and increasing system transparency. These are very important strategies aimed at building trust and ensuring ethical use of AI technologies within industrial environments (4).

III. Emerging Role of Edge AI

Edge AI is one of the innovations in modern computing characterised by the ability to perform data processing and AI inference on edge devices, closer to where the data is generated. This critical shift is driven by the need for real-time processing, increased security and reduced bandwidth costs. The proliferation of IoT devices, coupled with the advancement of 5G networks, has increased the importance of edge AI, creating opportunities for industries to improve their operational methods.

A. Characteristics of Edge AI Networks

Edge AI networks display distinct features that set them apart from conventional centralised frameworks:

1) ***Low Latency****: Proximity-based processing reduces delays in data transmission. Zeng L. et al (5) illustrated how this functionality benefits applications requiring prompt responses, such as autonomous drones.*

2) ***Energy Efficiency****: Distributed intelligence mechanisms significantly reduce the energy footprint, as shown by Ramo et al. (6), which is in line with global sustainability goals.*

3) ***Scalability****: Effective orchestration frameworks allow for dynamic resource allocation under varying workloads (7).*

4) ***Security and Privacy****: Local processing reduces vulnerabilities by limiting data transmission and enhancing user privacy (8).*

B. Industry Needs: Real-Time Responsiveness, Security, and Scalability

The shift to Edge AI addresses directly some of the most important industrial requirements by enhancing responsiveness, safety, and flexibility. Among the most striking advantages is real-time responsiveness, which considerably reduces decision latency in applications such as autonomous vehicles and predictive maintenance. Processing sensitive data locally, edge AI enhances security by reducing the risks associated with data transmission. What's more, The operational scalability achieved through edge-based solutions enables dynamic adaptation to fluctuating workloads while leveraging customised AI models optimised for specific requirements. The concept of an 'orchestrator' is the solution for the efficient management of resources and the optimisation of performance in such systems.

C. Analysis of Existing Solutions

This section aims to depict the state of the art by presenting selected findings from review studies, along with an analysis of existing orchestrators and their limitations.

a) Prior Research Review (State of the Art)

Progress in edge AI depends on the investigations associated with Next-Generation Software-Defined Networks (NG-SDNs) and intelligent orchestrators, which are supposed to deal with resource optimisation, scalability, and low-latency processing issues in distributed systems. M. Babar et al., 2018 (9) highlighted the integration of AI into software-defined networks and their possible significant effect on enhancing QoS, primarily in terms of real-time adaptability and effective resource distribution. This work underscores the role of AI-driven optimisation in managing network traffic and improving service reliability. Similarly, S. Guo et al. (2020) (10) investigated the optimisation of resources in edge-cloud ecosystems, where novel algorithms balancing computational loads under constraints of minimum energy consumption and data transmission cost are proposed.

These studies epitomise the growing interest in using AI for more efficient orchestration in complex environments. Recent research also highlights the importance of transfer learning and semantic AI in addressing the computational challenges of distributed systems. The potential of transfer learning has been investigated to reduce model training time and energy consumption to improve the sustainability and scalability of AI applications on different edge and cloud nodes. Moreover, the application of semantic AI allows for better reusability of models and algorithms, which reduces the overhead of traditional training methods and enables faster adaptation to changing workloads. Another major factor discussed in the literature is that of reduced latency. W. Liang et al., in 2020 (11), have used hybrid models of edge computing in the applications of IoT by showing that processing happens closer to where the data is, hence reducing latency in the transmission of the data. This makes it vitally important for applications that require ultra-low latency, such as autonomous systems and real-time analytics. Traditional orchestrators face challenges related to dynamic scalability and energy efficiency, as noted by S. D. A. Shah et al. (2020) (12). Their dependence on static policies and lack of adaptive strategies makes them inefficient in dealing with dynamic workloads. Moreover, although NG-SDNs provide a good foundation for network resource management, they usually lack advanced AI techniques that enable prediction and adaptation to sudden changes in system demands. The findings underline the requirement for further development in the design of orchestrators, with a special focus on the integration of AI-driven decision mechanisms, in order to achieve improved efficiency, scalability, and adaptability in edge-cloud ecosystems. This creates a basis upon which challenges in modern distributed intelligence systems can be dealt with.

b) Existing Orchestrators and Their Limitations

While current orchestrators, such as Kubernetes and Docker Swarm, have proved effective in the management of containerised applications, they also reveal limitations in some fundamental aspects. For example, Rahimi et al. (2014) (13) highlighted that resources optimisation remains one of the significant challenges as these systems mainly rely on static allocation methods, which are not suited to deal with dynamic and uncertain workload fluctuations. This rigidity can lead to inferior resource utilisation, particularly in E2C settings where demand undergoes frequent fluctuations.

In addition, the capability of current orchestrators to scale efficiently in hybrid environments consist of both edge and cloud parts hindered by their inability to provide seamless transitions of resources and tasks. S. Wang et al. (2017) (14) noticed that the combination of these two layers often results in latency and inefficiency, further increasing the orchestration problems faced in real-world cases.

Another critical weakness relates to the lack of energy efficiency strategies in such systems. Orchestrators are mainly designed to focus on functionality and performance, most of the time disregarding energy consumption reduction, one of the key ingredients for sustainable computing. More recent frameworks try to overcome these issues by adopting dynamic, AI-driven policies coupled with adaptive resource allocation strategies, as highlighted by T. Taleb et al. (2017) (15), who presented frameworks designed especially for E2C environments. These enhancements pave the way for more resilient, energy-efficient orchestration while keeping high performance and scalability.

IV. Mitigating allocation resources problems with proposed E2C Orchestrator

Efficient resource allocation in E2C environments remains a critical challenge due to the dynamic nature of workloads, varying network conditions, and limited computing capacity at the edge. The E2C Orchestrator presented in this article addresses these issues in the scope of the European Commission-funded TALON project, under the Grant Agreement N. 101070181, by introducing a novel architecture designed to optimise resource utilisation, minimise energy consumption, and ensure seamless task execution across distributed systems. This section describes the architecture and core components of the E2C Orchestrator, focusing on its role in solving resource allocation problems.

A. Orchestrator Architecture within TALON project

The TALON project introduces an E2C AI orchestrator designed to optimise resource utilisation, ensure seamless task execution, and improve energy efficiency across distributed systems. Its architecture is based on modular and dynamic components, enabling adaptability to changing workloads and heterogeneous environments.

1) Key Architectural Goals

The architecture of the E2C Orchestrator is designed to achieve multiple objectives that enhance its functionality and adaptability. It supports zero-touch deployment, enabling automated configuration and management of AI workloads without requiring manual intervention. This capability is powered by Kubernetes[1], which provides robust container orchestration. The orchestrator also ensures dynamic scalability by efficiently managing computational and communication resources to adapt to fluctuating demands. Additionally, it incorporates AI-based profiling mechanisms to optimise energy consumption while maintaining high levels of performance. Enhanced security and privacy are integral to the design, with advanced mechanisms implemented to safeguard AI models, data, and communication across nodes.

B. E2C Orchestator's Core Components and Functions

The E2C Orchestrator is composed of several key components (Figure 1) that work together to optimise task management and resource allocation. At its core is the Service Orchestrator, which oversees task deployment, real-time monitoring, and the efficient distribution of tasks across edge and cloud nodes. Supporting this is the Smart Policy Manager, which leverages AI models to analyse critical metrics such as latency, resource utilisation, and energy consumption, enabling data-driven decisions for optimal deployment strategies. The Resource Manager further enhances efficiency by managing task scheduling and offloading based on resource availability and predefined service-level agreements. Additionally, the User Interface

1 https://kubernetes.io/it/

includes a Monitoring System which plays a crucial role by providing real-time insights into system performance, detecting anomalies, and enabling proactive adjustments to maintain reliability and efficiency.

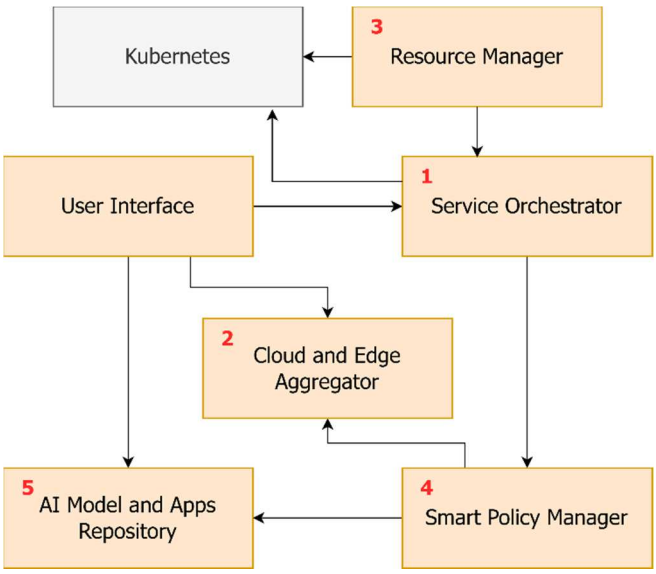

Figure 1. E2C Orchestrator Components

1) Service orchestrator

The Service Orchestrator (box numbered with 1 in the figure) is a central component, responsible for deployment, execution, and lifecycle management of tasks on edge and cloud resources. It ensures, based on AI-driven policies and real-time system metrics, the optimal mapping of tasks in a way that balances performance outcomes with energy efficiency. It determines the most appropriate nodes for task execution, skilfully assigning the edge devices to latency-sensitive tasks and cloud nodes to computationally demanding workloads. These, in turn, are supported by machine learning models that are trained on historical and real-time data to forecast resource demands, helping to flag potential performance bottlenecks.

During task execution, the Service Orchestrator actively monitors performance indicators and system states. In case of anomalies such as node overloads or failures, it actively reschedules tasks to maintain high system reliability and operational efficiency, thus minimizing disturbances and ensuring a smooth execution. Together with Kubernetes, the Service Orchestrator automates the deployment processes, hence eliminating the need for manual intervention and taking advantage of Kubernetes' functionalities in container orchestration, resource scaling, and self-healing mechanisms against potential failures.

The orchestrator relies on predictive models, especially Long Short-Term Memory (LSTM) networks, to estimate key parameters such as CPU usage, latency, and energy consumption. This forecasting ability allows it to make anticipatory adjustments of resource allocations while ensuring that application tasks are executed within pre-specified performance and energy bounds. The Service Orchestrator works very closely with both the Resource Manager and the Monitoring System to achieve the goal of complete task optimisation, where real-time insight from the Monitoring System meets fine-grained control over resources by the Resource Manager, aligning decisions with broader system goals. The Service Orchestrator enhances the reliability and sustainability of the TALON project through energy-efficient deployment strategies, scalability for varying workloads, and resilience by means of self-healing mechanisms and proactive anomaly detection.

2) Cloud and Edge aggregator

Cloud and edge aggregator supports the monitoring framework within the E2C Orchestrator of the TALON project. It is designed to provide real-time insights into the performance and health of edge and cloud resources. By integrating tools such as Prometheus [2], InfluxDB [3], and Kepler [4], it collects granular metrics including CPU and memory usage, energy consumption, task performance, and network stability. These metrics are aggregated into a central database to enable dynamic adjustments to task deployments, ensuring both resource efficiency and system reliability. For instance, Kepler enables precise energy tracking, allowing energy-aware orchestration across edge and cloud nodes.

Beyond real-time data collection, the monitoring system incorporates AI-driven anomaly detection to identify irregularities such as unexpected spikes in resource usage or network delays. These insights empower the orchestrator to proactively redistribute workloads, scale resources, or trigger self-healing mechanisms, ensuring uninterrupted service. Additionally, the system leverages historical data to support predictive analytics, enabling the anticipation of future demands and the preparation of resources to handle fluctuating workloads effectively.

The monitoring component's role extends to enhancing user transparency through the TALON dashboard, which visualises key performance indicators in real-time. This capability allows stakeholders to track task execution, resource utilisation, and energy efficiency. By integrating these monitoring capabilities, the TALON E2C Orchestrator achieves robust and adaptive resource management, meeting the demands of dynamic and distributed workloads with high reliability and sustainability.

3) Resource Manager

Effective resource management is essential in E2C environments, where multiple workloads must be dynamically allocated to meet performance demands, minimise costs, and reduce energy consumption. Traditional static approaches can struggle to handle the complexity and variability of these distributed systems, motivating the use of machine learning techniques to anticipate resource demands, adapt policies in real-time, and optimise overall efficiency. In the E2C Orchestrator, the Resource Manager leverages a multi-task LSTM model to analyse container-level metrics from the Kepler dataset. LSTM models are particularly well-suited for resource management because it captures both short- and long-term temporal dependencies in time-series data, making it effective at understanding patterns in CPU usage, memory usage, and energy consumption over extended periods.

[2] https://prometheus.io/docs/introduction/overview/
[3] https://www.influxdata.com/
[4] https://sustainable-computing.io/

Our multi-task LSTM architecture encodes these metrics into a shared representation before branching into two specialised heads. The classification head is used to identify the optimal policy (e.g., scale up, scale down, no action, or task offloading), while the regression head determines the numeric magnitude associated with that policy. This multi-task approach exploits complementary signals from classification and regression, improving accuracy, consistency, and the overall capability to learn from limited data. Once trained, the model can adapt to unlabelled datasets by employing threshold-based labelling or utilising pre-labelled data directly. To facilitate real-time integration into larger resource-management pipelines, the Resource Manager offers a lightweight RESTful API. Through this interface, users can upload container metrics, invoke predictions on both policy and magnitude, and seamlessly incorporate energy-aware decisions into the broader E2C Orchestrator architecture. This design ensures that the system remains flexible, data-driven, and capable of meeting the evolving demands of distributed edge-to-cloud workloads.

4) Smart Policy Manager

An additional feature of the orchestrator is its ability to automatically determine the most suitable implementation to deploy for a given task. Each task may have multiple available code implementations, which vary in terms of performance metrics such as result accuracy, execution speed, energy efficiency, and computational cost. The smart orchestrator within the TALON project systematically evaluates these implementations based on predefined priority criteria specified during task configuration. This evaluation ensures the selection of the optimal algorithm to meet the specific requirements of the task. The presence of multiple implementations characterised by distinct performance metrics introduces a substantial degree of flexibility in addressing heterogeneous user requirements. For instance, while one user may prioritise result accuracy as a critical factor, another may prioritise reduced execution time or minimal energy consumption. By maintaining a repository of implementations optimised for various performance metrics, the orchestrator facilitates a tailored selection process that aligns with diverse operational objectives. This strategy not only enhances system performances but also ensures that tasks are executed in accordance with specific operational priorities. Such an approach proves particularly advantageous in scenarios where trade-offs between metrics are necessary, thereby enabling the orchestrator to adapt dynamically to varying user-defined criteria and constraints. These capabilities underline the utility of the orchestrator in environments demanding precision, efficiency, or a combination thereof.

Below we illustrate how task implementation selection works:

For the selected task, the orchestrator performs an evaluation of the available implementations by first normalising their premeasured performance metrics onto a unified scale, ensuring comparability across diverse algorithms. These normalised metrics are subsequently transformed to align with the user's specified priorities, reflecting the relative importance of each criterion. Coefficients are then computed to quantify the degree to which each algorithm satisfies the user-defined criteria. Based on these coefficients, the algorithms are ranked in descending order, and the implementation with the highest coefficient is selected for execution.

5) AI Model and Apps Repository

The AI Model and Apps Repository serves as a centralised repository for managing AI models and applications within the TALON architecture. It provides access to pre-trained and custom models optimised for E2C environments, ensuring they are readily available for deployment across distributed systems. The repository supports version control, enabling efficient updates and rollbacks while maintaining consistency across all components. Integrated with the TALON orchestrator, it facilitates streamlined selection and deployment of models tailored to dynamic workloads and system requirements.

C. E2C Orchestrator Key Innovations and Differentiators

The E2C Orchestrator within the TALON project introduces a suite of innovations and differentiators that set it apart from conventional orchestration frameworks. These advancements enable efficient resource management, enhanced scalability, and energy-efficient task allocation in E2C environments.

A core innovation of the E2C Orchestrator is its integration of AI for dynamic and predictive resource management. By leveraging machine learning models, such as LSTM networks, the orchestrator predicts resource demands based on real-time and historical data. These predictions allow it to allocate resources adaptively, minimising task delays and energy consumption. The E2C Orchestrator continuously refines its policies using real-time feedback, ensuring optimal task placement across edge and cloud nodes.

Energy efficiency is a key differentiator of the E2C Orchestrator. By utilising tools like Kepler for precise energy consumption tracking, it dynamically adjusts workloads to minimise power usage without compromising performance. For instance, the orchestrator can offload energy-intensive tasks to cloud nodes when edge devices are at risk of overloading or prioritise edge nodes for latency-sensitive tasks while maintaining low energy consumption. This focus on sustainability aligns with global initiatives for reducing the environmental impact of IT infrastructure.

The orchestrator achieves seamless integration between edge and cloud environments through advanced NG-SDN and a modular design. Features like virtual logical networks, traffic replication, and dynamic policy enforcement ensure smooth transitions of tasks and data. By supporting programmable interfaces, including eBPF[5], the E2C Orchestrator enables fine-grained control over network resources, improving data flow and reducing bottlenecks.

The E2C Orchestrator's monitoring and self-healing capabilities are central to its resilience. It continuously collects metrics from edge and cloud nodes, enabling anomaly detection and proactive failure mitigation. If a node or task encounters an issue, the orchestrator automatically redistributes workloads or restarts affected processes,

[5] https://ebpf.io/

ensuring uninterrupted operations. This real-time adaptability enhances reliability and reduces downtime, critical for applications in industries like healthcare and manufacturing.

The orchestrator also stands out for its user-friendly interface and comprehensive visualisations. The TALON dashboard provides real-time metrics on task performance, resource utilisation, and energy consumption, empowering stakeholders with actionable insights. Its zero-touch deployment feature automates the configuration and scaling of tasks, reducing operational complexity and the need for manual intervention.

By combining AI-driven policies, energy-efficient operations, seamless edge-cloud integration, and real-time adaptability, the E2C Orchestrator redefines the standards for task orchestration in dynamic environments. These innovations position TALON as a player in advancing the capabilities of distributed computing for Industry 4.0 and beyond.

V. Experimentation with TALON and Analysis of Results

This section provides a detailed analysis of the outcomes derived from our proposed multi-task LSTM model for energy optimisation in distributed systems. Our experiments demonstrate the model's ability to accurately predict resource management policies and magnitudes while optimising energy consumption. Additionally, the insights gained from this study underscore the potential for AI-driven solutions in addressing dynamic resource allocation challenges in E2C environments. The case studies presented here illustrate how the model performs under real-world conditions, providing a foundation for future advancements in sustainable orchestration systems.

A. Experimental Setup

The experimental setup involved a local deployment of the TALON E2C Orchestrator prototype. Task workloads were synthetically generated based on container-level monitoring data extracted from the Kepler dataset, simulating varying resource demands over time (CPU, memory, energy consumption). Stress scenarios were introduced by creating fluctuating workloads of different intensities, mimicking dynamic industrial conditions. System monitoring tools ensured precise tracking of computational and environmental metrics.

B. Results and Case Studies

The objective of this case study is to evaluate the performance of the proposed multi-task LSTM model in optimising energy consumption while ensuring efficient resource allocation in distributed systems. Leveraging the Kepler dataset, the study focuses on predicting resource management policies and their associated magnitudes based on sequential container metrics, such as CPU usage, memory usage, and energy consumption. The goal is to demonstrate the feasibility of AI-driven solutions in addressing energy-aware resource management challenges in E2C environments.

The experiment was conducted on a MacBook Pro equipped with the Apple M3 Pro chip, utilising macOS 14.3 as the operating system and Python 3.11.8 as the primary programming language. Implementation was achieved through the PyTorch 2.3.5 deep learning framework. To assess the environmental impact of the computational processes, CodeCarbon was employed, ensuring precise tracking of energy efficiency and sustainability metrics.

The architecture implemented for this experiment is a multi-task LSTM network, specifically tailored for energy optimisation tasks in time-series data. Its design leverages the sequential nature of container metrics from the Kepler dataset, which includes features such as CPU usage, memory usage, and energy consumption metrics (core and DRAM joules). The architecture consists of an input representation layer for ingesting these metrics, followed by a shared multi-layer LSTM encoder that captures both short- and long-term dependencies. This shared representation feeds into two distinct task heads: a classification head that predicts optimal resource management policies (e.g., scaling actions) using cross-entropy loss, and a regression head that determines the magnitude of the actions using mean squared error (MSE) loss. A dropout mechanism with a 20% rate is integrated to mitigate overfitting risks and ensure robust model generalisation.

The model underwent a rigorous training regimen spanning 100 epochs, with a focus on minimising both classification and regression losses. A structured split of the dataset into training and validation sets facilitated a systematic evaluation of the model's performance.

Training ss exhibited a steady decrease, signalling effective model learning, while validation loss stabilised early, indicating strong generalisation without overfitting. The classification performance (Figure 2) achieved over 95% accuracy on the training set and maintained approximately 90% accuracy on the validation set, demonstrating its efficacy in predicting resource management policies. Meanwhile, the regression performance achieved significant reductions in MSE during initial epochs (Figure 4), stabilising at values indicative of accurate magnitude predictions (approximately 0.5 for training and 1.0 for validation). These results highlight the model's competence in both policy prediction and action magnitude estimation. Additionally, the environmental impact was commendably low, with CO_2 emissions of just 0.0012 kg, underscoring the sustainable nature of the computational process.

This work underscores the advantage of an AI-driven approach to real-time decision-making in resource management, a marked improvement over static policy. This case study will be additionally considered to extend the model by incorporating features like network latency and task priority, alongside exploring transfer learning techniques to enhance adaptability across diverse datasets and applications. This forward-looking approach will ensure broader applicability and greater efficiency in managing distributed systems.

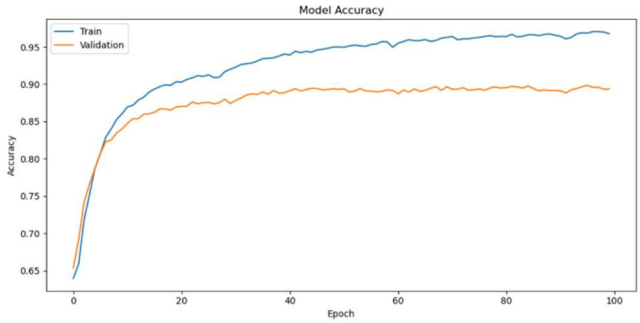

Figure 2. Model Accuracy Diagram

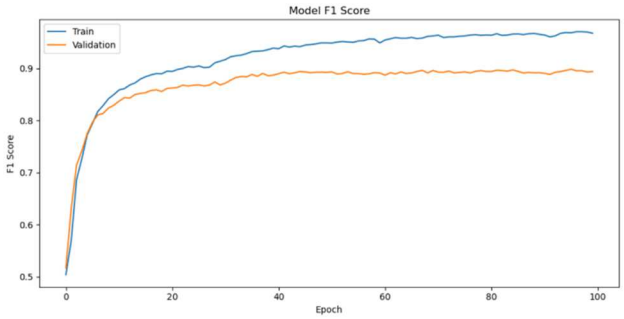

Figure 3. Model F1 Score Diagram

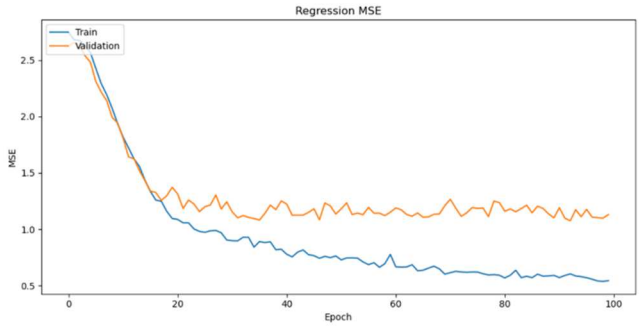

Figure 4. Model Regression MSE Diagram

VI. Limitations and Future Work

While the TALON E2C Orchestrator shows strong potential in optimizing task allocation and supporting sustainable industrial operations, several limitations of the current study should be acknowledged. First, the paper does not include a comparative evaluation with existing orchestration frameworks such as Kubernetes or OpenFaaS[6]. Although the orchestrator is built on top of Kubernetes and introduces advanced AI-driven, zero-touch orchestration logic, future work will include benchmarking against these platforms to validate performance improvements. Second, initial experiments were conducted on a single MacBook Pro, which limits the generalizability of the results. However, the orchestrator is also deployed in a cloud-based Kubernetes environment and is designed for scalable operation across distributed edge and cloud infrastructures. Moreover, while the integration of blockchain technology aims to enhance data integrity and model protection, the current version lacks formal security validation. Future work will include threat modeling, attack simulations, and empirical assessment of security features. Lastly, although the orchestrator aligns with Industry 5.0 values by enabling human-centric and adaptive decision-making, the current study does not explore user interaction in depth. Planned usability studies and evaluations in industrial settings will help assess its effectiveness in real-world human-in-the-loop scenarios.

VII. Conclusions

The TALON E2C Orchestrator represents a paradigm shift in resource allocation for distributed computing environments, combining state-of-the-art AI techniques with a solid architectural design to meet the growing demands of modern industrial systems. The orchestrator's modular and adaptive approach has been observed to successfully balance computational efficiency, energy consumption, and operational reliability, thus addressing the complexities inherent in dynamic workloads and heterogeneous infrastructures.

Experimental results demonstrate the effectiveness of the TALON framework, with the multi-task LSTM network showing accuracy and adaptability in predicting optimal resource management strategies. The incorporation of energy-aware policies, proactive anomaly detection, and seamless edge-to-cloud integration not only enhances performance but also aligns with global sustainability goals.

The findings of this paper highlight the transformative potential of AI-driven orchestration in distributed environments, paving the way for further advances in scalable, efficient, and sustainable computing systems. As industries increasingly adopt these solutions, systems such as the E2C Orchestrator are poised to play a pivotal role in shaping the future of intelligent, energy-efficient resource management. Future research should focus on extending the orchestrator's applicability across diverse industrial domains, enhancing security protocols, and integrating emerging technologies to further optimise performance and sustainability.

[6] https://www.openfaas.com/

MATCH & CONTRIBUTION

This paper aligns closely with the ICE IEEE 2025 conference theme of "AI-driven Industrial Transformation: Digital Leadership in Technology, Engineering, Innovation & Entrepreneurship." It presents a novel edge-to-cloud (E2C) orchestrator developed within the TALON project (Horizon Europe GA No. 101070181), designed to enhance intelligent task allocation, operational efficiency, and energy sustainability in industrial environments. By leveraging AI-driven policies, real-time monitoring, and predictive analytics, the orchestrator supports zero-touch deployment and dynamic scalability across heterogeneous edge and cloud infrastructures. The work contributes to digital leadership through the integration of advanced technologies such as LSTM-based forecasting, semantic AI, and energy-aware scheduling, ensuring robust, secure, and adaptive orchestration. This approach directly addresses the conference's call for innovations in AI-driven engineering practices and entrepreneurship by promoting explainability, interoperability, and sustainability in industrial automation. The paper offers both theoretical and applied value, demonstrating how AI-centric orchestration architectures can be foundational to the next generation of Industry 5.0 systems.

ACKNOWLEDGEMENT

The research leading to the results presented in this paper has received funding from the European Union's funded Project TALON under grant agreement no 101070181.

REFERENCES

[1] Orchestration in the Cloud-to-Things compute continuum: taxonomy, survey and future directions. Amjad Ullah, et al. 2023, Journal of Cloud Computing 12(5), pp. 1-20.

[2] Edge Computing for Industry 5.0: Fundamental, Applications, and Research Challenges. M. Sharma, A. Tomar, A. Hazra. 2024, EEE Internet of Things Journal, vol. 11, no. 11, pp. 19070-19093.

[3] A Joint Energy and Latency Framework for Transfer Learning over 5G Industrial Edge Networks. Yang, B., Fagbohungbe, O., Cao, X., Yuen, C., Qian, L., Niyato, D., & Zhang, Y. 2021, arXiv:2104.09382.

[4] Digital Twins for Transparent and Explainable AI Systems in Industry 5.0. P. Nguyen, L. Wang, and K. Zhang. 2023, International Journal of Automation and Computing, vol 20, no 3, pp. 365-378.

[5] A3D: Adaptive, Accurate, and Autonomous Navigation for Edge-Assisted Drones. Zeng, L., Chen, H., Feng, D., Zhang, X., & Chen, X. 2023, IEEE Transactions on Network Science and Engineering.

[6] Ambient intelligence—the next step for artificial intelligence. Ramos, C., Augusto, J. C., & Shapiro, D. 2020, Journal of Ambient Intelligence and Humanized Computing, 11(10), pp. 3833–3849.

[7] A resource orchestration framework for adaptive application deployment in cloud environments. Zhou, P., Wang, Y., Wang, H., & Li, J. 2023, Journal of Cloud Computing 12(3), pp. 245–260.

[8] Federated machine learning: Concept and applications. Yang, Q., Liu, Y., Chen, T., & Tong, Y. 2021, ACM Transactions on Intelligent Systems and Technology 10(2), pp. 1-19.

[9] A secured data management scheme for smart societies in industrial internet of things environment. M. Babar, et al. 2018, IEEE Access.

[10] Trusted cloud-edge network resource management: DRL-driven service function chain orchestration for IoT. S. Guo, et al. 2020, IEEE Internet of Things Journal.

[11] Deep Reinforcement Learning for Resource Protection and Real-Time Detection in IoT Environment. W. Liang, et al. 2020, IEEE Transactions on Industrial Informatics.

[12] SDN enhanced multi-access edge computing (MEC) for E2E mobility and QoS management. S. D. A. Shah, et al. 2020, IEEE Access.

[13] Mobile Cloud Computing: A Survey, State of Art and Future Directions. M. R. Rahimi, et al. 2014, Mobile Networks and Applications.

[14] A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications. S. Wang, et al. 2017, IEEE Access.

[15] On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration. T. Taleb, et al. 2017, IEEE Communications Surveys & Tutorials.