Received: 24 October 2023 | Revised: 27 March 2024 | Accepted: 24 April 2024

DOI: 10.1002/spe.3340

RESEARCH ARTICLE

WILEY

Sustainable edge computing: Challenges and future directions

Patricia Arroba[1] | **Rajkumar Buyya**[2] | **Román Cárdenas**[1] | **José L. Risco-Martín**[3] | **José M. Moya**[1]

[1]Department of Electronic Engineering. Center for Computational Simulation, Universidad Politécnica de Madrid, Madrid, Spain

[2]Cloud Computing and Distributed Systems (CLOUDS) Lab, The University of Melbourne, Melbourne, Australia

[3]Department of Computer Architecture and Automation, Universidad Complutense de Madrid, Madrid, Spain

Correspondence
Patricia Arroba, Department of Electronic Engineering, ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, Spain.
Email: p.arroba@upm.es

Funding information
University of Melbourne; HiPEAC Network; Agencia Estatal de Investigación, Grant/Award Number: PID2019-110866RB-I00/AEI/10.13039/501100011033; Ministerio de Ciencia e Innovación; Centro para el Desarrollo Tecnológico Industrial, Grant/Award Numbers: IDI-20171194, RTC-2017-6090-3

Abstract

The advent of edge computing holds immense promise for advancing the digitization of society, ushering in critical applications that elevate the overall quality of life. Yet, the practical implementation of the edge paradigm proves more challenging than anticipated, encountering disruptions primarily due to the constraints of applying conventional cloud-based strategies at the network's periphery. Increasingly influenced by sustainability commitments, industry regulations currently view edge computing as a potential threat, primarily due to the energy inefficiency of solutions situated in close proximity to data generation sources and the rising density of computing. This paper presents a proactive strategy to transform the perceived threat into an opportunity, steering the sustainable evolution of future edge infrastructures to make them both environmentally and economically competitive for accelerated adoption. The vision outlined addresses key challenges associated with edge deployment and operation, emphasizing energy efficiency, fault-tolerant automation, and collaborative orchestration. The proposed approach integrates two-phase immersion cooling, formal modeling, machine learning, and federated management to effectively harness heterogeneity, propelling the sustainability of edge computing. To substantiate the efficacy of this approach, the paper details initial efforts towards establishing the sustainability of an edge infrastructure designed for an Advanced Driver Assistance Systems application.

KEYWORDS

collaborative resource management, edge federations, energy efficiency, fault-tolerant automated operation, formal modeling, two-phase immersion cooling

Abbreviations: ADAS, Advanced driver assistance systems; AI, Artificial intelligence; CAGR, Compound annual growth rate; CHF, critical heat flux; CNN, Convolutional neural network; CONSR, smart grid consumer model; DC, Data centers; DL, Deep learning; EDC, Edge data center; FaaS, Function-as-a-service; GAN, Generative adversarial network; HIL, Hardware-in-the-loop; ICT, Information and communications technology; IoT, Internet of things; ISP, Internet service providers; IT, Information technology; KPI, Key performance indicator; M&S, Modeling and simulation; M&S&O, Modeling, simulation, and optimization; MBSE, Model-based system engineering; MEC, Multi-access edge computing; ML, machine learning; MSOBSE, Modeling, simulation, and optimization-based systems engineering; NRMSD, Normalized root-mean-square deviation; PROVR, energy provider; PUE, Power usage effectiveness; QoS, Quality of service; RAN, Radio access network; SE, Systems engineering; SoS, Systems of systems.

1 | INTRODUCTION

The emergence of the edge computing paradigm is driven by the need to address escalating data rates stemming from data-intensive applications. The proliferation of data, particularly from Internet of Things (IoT) applications, aimed at enhancing the efficiency across diverse sectors such as healthcare, transportation, energy, and agriculture, underscores the urgency of this challenge. The influx of substantial data into the network has the potential to saturate it, resulting in significant delays. This poses a critical issue, especially considering that many IoT applications necessitate strict latency requirements for real-time data processing and decision-making. For instance, IoT-enabled devices employed for remote patient health monitoring rely on prompt detection and notification of any anomalies, facilitating faster and more accurate treatment.

Edge computing serves as a solution to this problem by strategically positioning computing resources in proximity to the data sources. This approach minimizes the volume of data requiring transmission to the cloud for processing. The localized processing and analysis of data offer numerous advantages: (i) mitigation of user-perceived delays and acceleration of processing speed, (ii) reduction in network costs and bandwidth utilization, and (iii) provision of local reliability in the event of connectivity issues.[1]

Based on Gartner's report on edge computing,[2] 19% of the surveyed industries have already deployed edge computing, and an additional 32% expect to do so within the next three years. According to Cisco, edge technology ranges from multi-access edge computing (MEC), cloudlets and micro data centers, to fog computing (between the cloud and edge devices).[3] Despite these theoretical concepts, how are cloud providers reaching the edge today? Hyperscale operators, managing massive clouds often situated in remote areas, are extending their presence closer to users through various means.

Primarily, they adopt a less risky approach by utilizing local DCs, such as colocation DCs, to host edge nodes. These nodes consist of Information Technology (IT) equipment, ranging from small racks to entire data rooms. This integration blurs the boundaries between cloud, hybrid-cloud, and edge computing to some extent. However, this strategy alone may prove inadequate to support the proliferation and commoditization of IoT services, necessitating the deployment of an edge infrastructure with finer granularity and enhanced computing capabilities.

1.1 | Sustainability implications of transitioning to edge computing

For prominent cloud providers like Google, Amazon, or Microsoft,[4] we encounter highly sustainable and efficient cloud DCs that they own and operate. These companies currently assert carbon neutrality and aspire to achieve Net Zero Carbon by 2030–2040, intending to offset as much carbon as they emit.[5] Additionally, they exhibit efficiency in energy utilization. To gauge this efficiency, we often refer to Power Usage Effectiveness (PUE), a metric that delineates how effectively energy is utilized in a DC (calculated as the total facility energy divided by the energy consumed by the IT devices). Consequently, PUE provides insights into the energy overhead, primarily utilized for cooling the IT resources, for each unit dedicated to computing. According to the Uptime Institute, the average PUE of the DC industry in 2022 hovered around 1.55,[6] indicating that roughly 35% of the total energy budget is utilized by the cooling systems.

Google, Amazon, and Microsoft assert a PUE around 1.1[7,8,9] signifying that their energy overhead is six times lower than the industry average. How do they achieve such efficiency? For example, examining the locations of Google's DCs in 2022, as depicted in Figure 1A, reveals that most of them are situated in: (i) cold areas, where cooling is significantly more efficient, leading to a better PUE, and (ii) regions supporting renewable energy production, thereby reducing the carbon footprint. Hence, they leverage cleaner energy sources and employ them more efficiently.

Analysis of the latest DCs constructed by these cloud providers uncovers commonalities. They occupy extensive areas, not only to accommodate data rooms but also for renewable energy production (wind turbines and solar panels) and cooling equipment, necessitating substantial financial investment. Thus, when the cloud is located in an optimized, cool, and spacious environment and is adequately funded, it can achieve high sustainability and efficiency. However, challenges arise when the IT infrastructure needs proximity to the data sources.

Figure 1B depicts Google's edge nodes, numbering over 1,300, strategically positioned closer to users. Edge nodes are servers owned by cloud providers and deployed within the networks of local operators and internet service providers. Many of these nodes are housed in colocation DCs, ranging from individual racks to entire data rooms, spanning over 200 countries and territories. The colocation market experienced significant growth in 2020, expanding by 22% and 9%

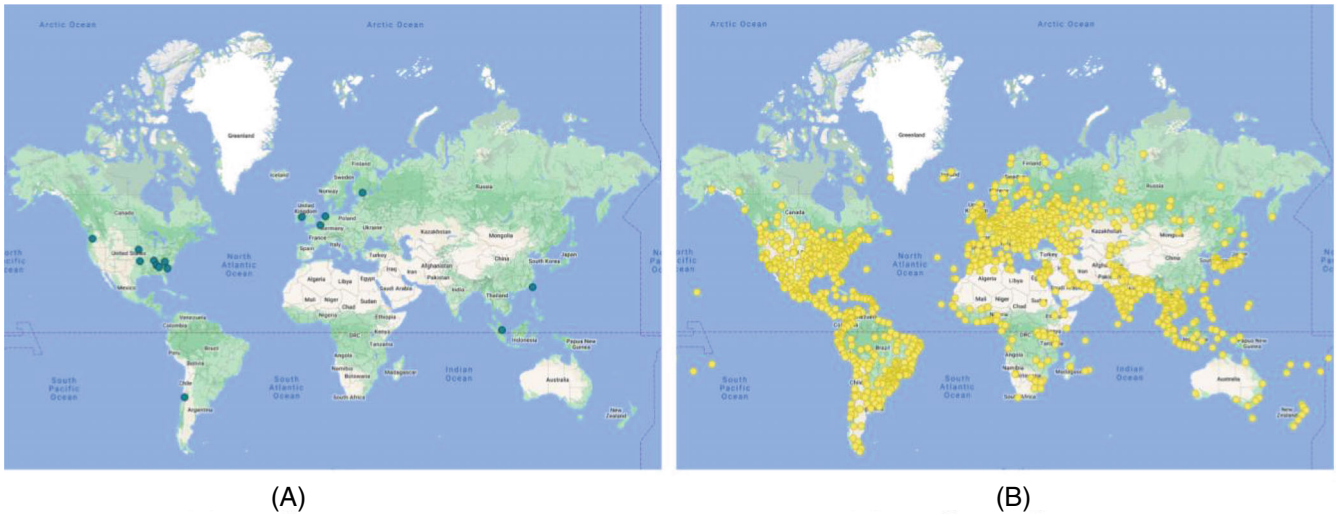

FIGURE 1 Google's cloud DCs and edge nodes (https://peering.google.com//#/infrastructure). (A) Google's cloud DCs, (B) Google's edge nodes.

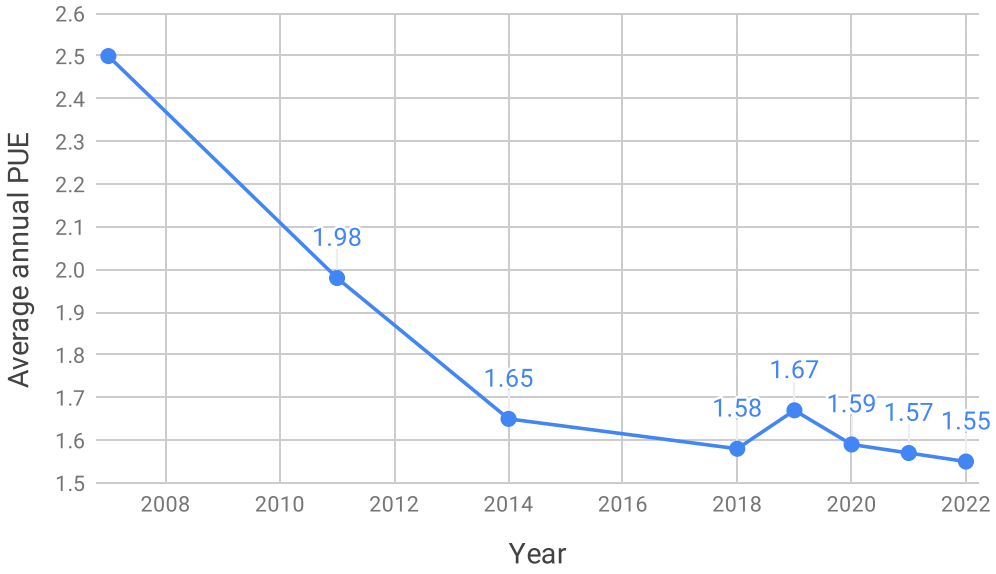

FIGURE 2 Evolution of PUE in the DC industry (data sourced from the Uptime Institute[6]).

for wholesale and retail colocation, respectively, driven by the revenues generated by these hyperscale operators.[10] These edge nodes deliver services tailored to the preferences of the local user base, thereby reducing latencies and alleviating network congestion.

Given the prevailing global average PUE of approximately 1.55, a noticeable decline in energy efficiency becomes imminent as IT infrastructure migrates closer to the network's edge. While local DC operators have implemented various energy efficiency measures, achieving the low PUEs of hyperscalers is often unattainable due to economic, technical, or geographical constraints. This limitation results in the average PUE plateauing, as illustrated in Figure 2. Furthermore, since air cooling continues to be the predominant technology even in new DCs, the adoption of new processors with higher thermal power will adversely impact cooling efficiency. Consequently, the industry-average PUE is anticipated to increase in the coming years before resuming a downward trend.[6] Thus, the closer we position dense IT infrastructure to the network's edge, the more obstacles we encounter in terms of sustainability and energy efficiency.

According to Gartner,[11] the number of IoT devices is expected to triple from 2020 to 2030, leading to a growth in data injected into the network. The fastest growing areas will be manufacturing and natural resources, with a Compound

Annual Growth Rate (CAGR) of 18%, healthcare providers (13%), smart buildings (12%), and automotive (10%). The adoption of IoT applications in nearly every industry will drive an increase in compute density at the edge. To accommodate this growth, new infrastructure must be deployed in the near future as colocation DCs reach their limits.

The design of this new infrastructure is still undetermined and must be customized to meet the requirements of the applications that can gain the most from edge computing, while also ensuring its environmental and economic sustainability. In this context, this paper presents our vision and our early steps toward ensuring the sustainability of future edge computing infrastructures equipped with robust and dense IT capabilities. *Our efforts are focused on achieving three main goals*: (i) an energy-efficient and economically viable deployment, (ii) a fault-tolerant, learning-centric automated operation, and (iii) an energy-aware collaborative resource management to improve resource efficiency.

The remainder of this article is organized as follows. In Section 2, we delineate the primary limitations associated with implementing cloud-based approaches in close proximity to data generation sources. We identify the research challenges posed to edge sustainability by these constraints. Sections 3–5 introduce two-phase immersion cooling, formal modeling, machine learning (ML), and energy-centric federated management as edge-enabling technologies, explaining the significance and innovations of our proposed vision. Section 6 discusses the current state-of-the-art vision on challenges that still need to be solved and require further study in different aspects of energy efficiency. In this section, we present our integrated approach and detail how it helps to explore the global impact of these relevant energy-related dimensions. To illustrate the potential benefits of our integrated approach that combines these technologies, Section 7 presents a use case for advanced driver assistance systems. Finally, in Section 8, we draw the main conclusions and outline future directions.

2 | LIMITATIONS OF CURRENT APPROACHES AND KEY OPEN CHALLENGES

When designing the deployment and operation of infrastructure to support edge computing, it is important to consider three factors derived from its nature. First, this emerging edge infrastructure will be strategically deployed in proximity to end-users, and its placement cannot be determined solely by optimizing for energy efficiency conditions. Second, given the global average PUE, it is anticipated that the energy efficiency will markedly decline as IT converges towards the edge. Third, as edge computing will enable the digitization of society, we should aim for a deployment that meets the IT needs of future applications. Driver assistance systems and personalized medicine, and online gaming are just a few examples of applications that edge computing will facilitate and enhance. These applications require intensive computing for massive data processing through ML. This must be done close to the users due to the large data rates and critical latencies.

Edge computing technology ranges from small edge devices to micro DCs. However, determining which infrastructure suits emerging applications poses a challenge. Resource-constrained and battery-powered edge nodes typically benefit from computation offloading to improve their lifetime. However, the computation has traditionally been offloaded directly to the cloud.[12] Therefore, it is reasonable to consider future edge deployments that include small, compute-intensive micro DCs located closer to the data sources. These edge Data Centers (EDCs) can partially or completely process the computation offloading, connecting cloud DCs to edge devices for demanding tasks.

EDCs can facilitate a more scalable and flexible model. However, as they host a significant amount of IT resources, they become more susceptible to higher temperatures. Since EDCs are deployed in close proximity to users, traditional cooling solutions may be less efficient in terms of power consumption and size. Therefore, this infrastructure would benefit the most from novel cooling solutions.

In this context, edge computing proves to be significantly more disruptive than initially anticipated, representing not merely an incremental extension from the cloud. The current technological advancements that facilitated highly efficient developments in cloud computing exhibit several limitations when extended to the deployment, operation, and management at the network's edge.

2.1 | Infrastructure deployment

The edge computing paradigm has unique characteristics that prevent direct application of cloud-based solutions. EDCs must be located closer to data sources and cannot be deployed in areas with cool or cold weather, as preferred by big hyperscaler clouds. The network edge may reduce energy efficiency, particularly in cooling processes, resulting in a notable elevation in the average PUE.

Second, in order to be close to data sources, the size of EDCs needs to be limited, especially considering their predominant urban location. Currently, the most efficient clouds typically require expansive spaces to host massive renewable resource generators (such as wind turbines or solar farms) and cooling equipment, which may not be readily accessible to EDCs.

Third, the economic viability of this infrastructure depends heavily on the cost of EDCs, given the expected massive deployment of these facilities. We face the task of reducing costs while maintaining energy efficiency and minimizing physical footprint in suboptimal locations. However, it's worth noting that the technological solutions required to enhance efficiency and diminish the spatial requirements often incur high costs.

Consequently, the *first key challenge* at hand is devising strategies for an environmentally and economically sustainable deployment in the realm of edge computing.

2.2 | Distributed dynamic operation

Cloud operations are conducted on-site, benefitting from dedicated personnel managing the day-to-day functioning and decision-making within these expansive, centralized infrastructures. In contrast, EDCs are a widely distributed computing infrastructure, making on-site operation impractical. However, the edge computing paradigm will support applications critical to health and safety, such as personalized medicine and driver assistance. In this context, service disruptions are unacceptable as they pose a risk to the physical safety of users.

This underscores the significance of learning-centric automatic operation as a pivotal research area for dynamic distributed infrastructures. However, the question arises: how do we respond to unforeseeable anomalous situations beyond our control, such as security attacks, hot spots in data rooms, system malfunctions, or service outages while maintaining its sustainability? In the cloud, operators manage these anomalies, but in unattended EDCs, there is a pressing need for tools to prevent irreparable equipment damage and service disruptions.

Therefore, the *second key challenge* is to develop mechanisms that enable fault-tolerant automatic operation integrated into the evolving energy-aware landscape of edge computing.

2.3 | Efficient use of available resources

Hyperscale cloud operators consolidate user requests across regions, resulting in smoother DC demand profiles that are minimally affected by user mobility. This facilitates scaling of DC resources, improving resource management and utilization efficiency. In contrast, edge computing processes user demand locally, resulting in handovers between small DCs within the same region. These handovers occur, for instance, at the Radio Access Network (RAN) level or between RANs across the backbone. Consequently, the processing demand perceived by each EDC is significantly influenced by the mobility patterns of the data generation sources. This results in pronounced variability, including daily or seasonal fluctuations.

Scaling EDCs to accommodate peak demand is impractical as it would result in underutilized and prohibitively expensive infrastructure. However, supporting critical applications at the edge requires an uncompromised Quality of Service (QoS) for end users. Then, maintaining QoS in high-demand scenarios and intelligently distributing workloads requires collaboration between EDCs as well as with the cloud. Therefore, the *third key challenge* in ensuring the sustainability of EDCs is implementing energy-aware, effective collaborative resource management.

In the following sections, we present our three approaches to address each of these key challenges. Specifically, we explain the advantages for the deployment and operation of the future edge using (i) Two-phase immersion cooling technology for sustainable deployments, (ii) Model-Based Systems Engineering together with Machine Learning for fault-tolerant automatic operation, and (iii) edge federations for collaborative resource management. Additionally, we present the significance and innovation of our proposed vision for each of them in comparison to the current state of the art.

3 | TWO-PHASE IMMERSION COOLING FOR SUSTAINABLE EDGE DEPLOYMENTS

Two-phase immersion cooling emerges as a technology poised to address numerous challenges at the edge while enhancing the sustainability of existing DCs. In this cooling method, the IT components are submerged in a sealed tank filled with

an engineered fluid that undergoes cyclic evaporation and condensation processes at notably high operating temperatures, exceeding 60°C.

When the fluid comes into contact with high-temperature surfaces, such as CPU or GPU chips, it undergoes evaporation. The condenser then circulates glycol water in a secondary circuit to exchange heat with the external environment, condensing the vapor back into the liquid phase, as shown in Figure 3. These phase transitions prove highly efficient in extracting heat from the system, resulting in a noteworthy reduction in cooling consumption, up to 95% when compared to conventional methods.[13]

3.1 | Advantages of two-phase immersion cooling for edge computing

Two-Phase Immersion Cooling Technology has several advantages to meet the emerging requirements for deploying edge computing infrastructures. These are EDCs that are energetically efficient, resilient to hot climates, small in size, and affordable.

3.1.1 | Enhanced energy efficiency and weather flexibility

The transition between gas and liquid phases in this technology proves highly effective in extracting heat from immersed systems, leading to a substantial reduction in cooling consumption, with reported gains of up to 95% and attaining PUE values of 1.02–1.03, an achievement previously observed only in free cooling. Moreover, its low global warming potential (below 1) ensures compliance with stringent environmental regulations.[13]

Two-phase immersion technology confers a notable advantage in edge computing by consistently delivering benefits regardless of external temperature and humidity conditions. Operating at temperatures exceeding 60°C, this technology allows for the configuration of cooling systems with elevated setpoint temperatures. Consequently, an impressive PUE of approximately 1.03 can be achieved, surpassing not only the industry average (PUE = 1.55) but also outperforming the average for hyperscale operators (PUE = 1.1). This advantage holds true even in edge locations with warmer climates.

3.1.2 | Increased power densities and reduced physical footprint

Power density is a crucial parameter when it comes to designing computing and cooling infrastructures in DCs.[6] Despite advancements, there has been limited growth in rack power over recent years, with the industry primarily working with cabinets that support typical power ranges of 4–6 kW. This lack of substantial progress can be attributed to the

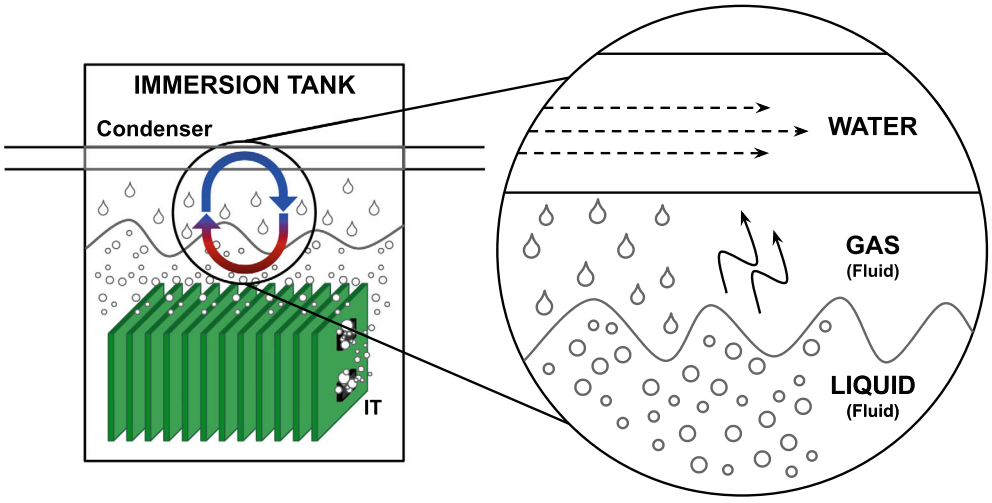

FIGURE 3 Two-phase immersion cooling operation.

predominant use of air cooling in the industry, which is constrained by the relatively low heat transfer coefficient of air. Additionally, this traditional technology requires the implementation of hot and cold aisle configurations, as well as a significant amount of equipment (e.g., chillers, computer room air conditioner and handler units, etc.). Consequently, these requirements impose considerable limitations on the physical footprint of DCs utilizing air cooling.

Two-phase immersion cooling revolutionizes power density in DC racks by enabling a remarkable 60-fold increase. This substantial enhancement is achievable due to its exceptional heat transfer capabilities, allowing for power densities of up to 250 kW per rack.[13] By adopting this technology, DCs can reduce their floor space requirements by up to 10 times. This efficiency translates into a higher computing power-to-area ratio, increasing from the traditional 10 kW/m2 with conventional cooling methods to an impressive 100 kW/m2 with two-phase immersion technology. Consequently, this allows EDCs to be downsized and situated in urban environments near end users, enabling them to provide high computing capabilities to support applications with demanding requirements.

3.1.3 | Cost-effectiveness to address adoption concerns

Two-phase immersion cooling significantly reduces infrastructure costs by improving heat extraction performance and efficiency. It minimizes cooling consumption by up to 95%, resulting in a mere 1% energy overhead. Additionally, it requires less cooling equipment and floor space compared to traditional cooling methods. However, what is causing the industry's reluctance to adopt this technology? The primary reasons for the low adoption of two-phase immersion cooling are concerns about reliability, maintenance, and fluid leakage.[14] Additionally, there is a lack of practical information regarding relevant aspects, such as adapting IT infrastructure and equipment needs to support this cooling technology.

The main capital expenditures include dielectric liquids, immersion tanks, piping, and pumps for recirculation. Dielectric fluids require filtering and careful handling to prevent spills, but they do not degrade and can be used indefinitely. Additionally, the use of fans throughout the infrastructure is no longer necessary since air does not have to be recirculated throughout the facility as in air-based cooling. This results in cost savings in the range of 15%–25%.[14] The operating expenses related to equipment support and maintenance are similar to those of air cooling.[14] Furthermore, two-phase immersion cooling results in a significant reduction of approximately 50% in energy consumption without compromising computing performance.

While retrofitting conventional air-cooled DCs with two-phase immersion cooling may not be economically feasible, this cost reduction is particularly important for new edge computing deployments, which require a large number of compute-intensive EDCs to meet local application requirements.

3.2 | Significance and innovation of our approach in two-phase immersion cooling

Commercial two-phase immersion cooling solutions are currently under development, although only a few companies have begun to leverage this technology. Cryptocurrency mining has been the primary application driving the development of market-ready products. Companies like Bitfury and LiquidStack have deployed a 160MW infrastructure, achieving an impressive power density of 252 kW in 48U DataTanks.[15] Some hyperscale operators have also recently shared their progress in implementing this technology in their own cloud infrastructures.[16] For example, Microsoft has deployed 36 Open Compute server-class blades in Azure.[17,18]

However, the industry has not yet fully harnessed the potential of this technology, and with our approach, we aim to take it a step further. Two-phase immersion cooling allows for a dissipated heat flux exceeding 20 W/cm2 and has reached a critical heat flux (CHF) of 53 W/cm2 and 59 W/cm2 for Novec 7100 and 7200 coolants, respectively.[19] This capability has the potential to cool about 2 to 3 MW in a standard-size DataTank,[15] but the current technology is still far from reaching this limit.

Operating in proximity to the CHF is more feasible for constant loads, such as those encountered in cryptocurrency mining applications. However, in applications driven by user demand, such as cloud and edge computing, fluctuations in workload can potentially push the system into hazardous regions beyond the CHF. Hence, comprehending the behavior of electronic systems within the fluid is imperative for optimizing performance without compromising their physical integrity.

In this context, our objective is the development of highly accurate predictive power and thermal models for submerged IT systems. These models are crucial to ensure peak performance while safeguarding the physical integrity of the

systems. Given the inherent turbulence in two-phase cooling systems, ML emerges as a promising alternative for crafting precise models suitable for environments characterized by highly complex fluid dynamics.

In our research, our objective is to advance beyond the current state of the art by predicting the behavior of these systems, particularly in terms of power consumption and thermal performance. This proactive approach empowers us to harness the full potential of two-phase immersion cooling by anticipating and addressing abnormal situations, minimizing failures, and mitigating various operational challenges within future edge computing environments.

4 | MODEL-BASED SYSTEMS ENGINEERING AND ML FOR EDGE AUTONOMOUS OPERATION

Today's DCs already integrate elements of Artificial Intelligence (AI) to optimize the management of computing and cooling systems. These systems actively monitor the environment and inject collected data into algorithms, which in turn learn to optimize performance decisions. According to the Uptime Institute,[6] over half of DC owners and operators express a reliance on appropriately trained ML models for making operational decisions. While automated operation is a feasible choice for DCs managed on-site by operators, the distributed nature of edge computing complicates the deployment of staff in the infrastructures. Consequently, in edge computing, the automation of operation becomes indispensable, and the standardization of the environment must be an integral part of the operational model to control and manage edge systems.[1]

In this context, our aim is to amalgamate Model-Based Systems Engineering and Systems Optimization through Simulation with formal specification and ML. The objective is to create a formal edge computing model with two primary goals. Firstly, standardizing the design and operation of the infrastructure using mathematical formalisms. Secondly, facilitating optimization algorithms to explore system behavior and make improved decisions automatically.

4.1 | Formal modeling for system design, implementation, and digitalization

In this article, we present a novel approach that integrates Modeling, Simulation, and Optimization (M&S&O) methodologies into the Systems Engineering (SE) technical process for the development of complex systems. Our Modeling, Simulation, and Optimization-Based Systems Engineering (MSOBSE) strategy accentuates the use of formal modeling tools to construct more robust and reliable systems. Simulation serves as the primary tool for assessing the alignment of the proposed model with system requirements. Additionally, it incorporates optimization techniques to automate design decisions that enhance compliance with Key Performance Indicators (KPIs) during the system development phase. Figure 4 illustrates the proposed MSOBSE workflow.

In the initial stages of the SE process a thorough examination of the requirements from all system stakeholders (e.g., customers, contractors, and developers) is imperative. This scrutiny leads to the definition of KPIs, constituting the system specification. Once the target system is delineated, the process commences with system modeling. Initially, a conceptual model of the solution is established, followed by the translation of this conceptual model into a computational model. Simulations on the computational model are used to verify its conformity to the conceptual model's specifications.

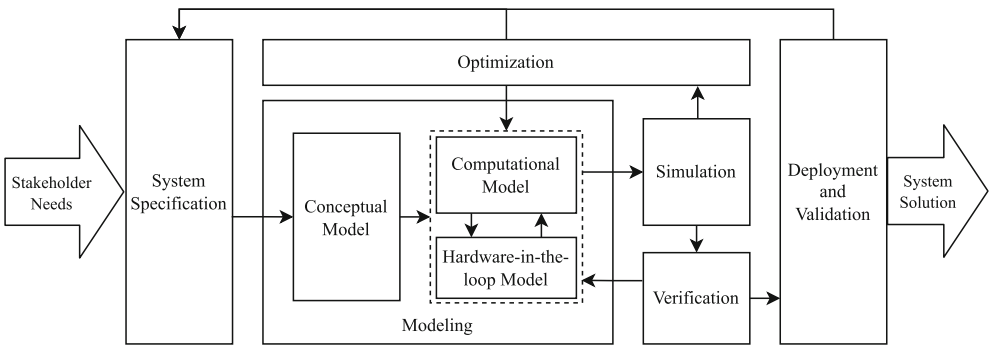

FIGURE 4 Workflow of the proposed MSOBSE technical process.

Typically, this process adopts an incremental approach, where the conceptual model is decomposed into multiple submodels. Each subcomponent is translated into a computational model, simulated, and independently verified. As these submodels are validated, they are merged, and the resulting model undergoes further verification. Ultimately, a computational model is achieved that effectively captures the behavior outlined in the formal model of the system.

In the realm of complex systems, such as DCs, endowed with numerous degrees of freedom, designing all elements optimally becomes non-trivial. To address this complexity, we propose the integration of automated optimization algorithms into the system development process. These algorithms adjust the verified computational model to attain a solution aligning with system specifications while reducing development costs. In cases where an ideal design meeting all system requirements is unattainable, specifications are revisited, initiating a iterative process of modeling, simulation, and verification.

Following the optimization of the computational model, the system implementation phase commences. Hardware-In-The-Loop (HIL) models progressively substitute parts of the computational model with actual implementations of the modeled system components.[20] Subsequently, testing and verification of implemented elements are conducted through real-time simulations over the HIL model. If a component deviates from its corresponding model, a review of the conceptual model is undertaken to rectify potential errors. Ultimately, the process culminates in a complete implementation of the system under study, a solution derived from a validated conceptual model, wherein components are optimized to meet all system specifications while minimizing development expenses. After the design and implementation of the system, the MSOBSE strategy can function as a digital twin, aiding in decision-making for the real-world complex system.

Integrating analytical models with ML brings unique challenges, particularly in ensuring that the combined model accurately reflects the system's behavior. To tackle these challenges, we have established a systematic process that includes the following steps: First, we ensure that the data used for training and validating both ML and analytical models are consistent and representative of the same real-world conditions. This is crucial for maintaining a uniform understanding of the system's behavior across different modeling approaches. Second, we employ an iterative refinement methodology, where the models are continuously improved based on discrepancies between their outputs and actual system data. This iterative process allows for the fine-tuning of both analytical and ML components, enhancing the overall model accuracy. Third, we conduct extensive validation of the integrated model by comparing its predictions with observed data, ensuring that the model's performance is reliable and robust. These strategies are essential for maintaining the transparency and interpretability of the integrated model, which is particularly important for decision makers who rely on the model's outputs.

4.2 | ML for fault-tolerant decision making

ML-based solutions have become a reality in DCs, offering highly accurate dynamic behavioral models of infrastructure elements, including temperature, humidity, and power consumption. ML aids operators in optimizing control decisions such as temperature setpoints, workload allocation, and scheduling. Effective training of ML models requires a substantial volume of data collected by monitoring systems, which works well for capturing the standard behavior of real systems in production. But what occurs when anomalous situations arise, such as equipment or monitoring system failures?

Failures in DCs can result in outages with severe economic consequences. Over the past three years, approximately 60% of surveyed operators[6] experienced outages primarily attributed to power issues (37%), computer systems (22%), and cooling (13%). The financial impact of the last outage for 45% of these DCs ranged between $100,000 and $1 million, while another 25% faced costs exceeding $1 million. Operators generally acknowledge that these outages could have been prevented through better procedures, management, configuration, and training. The challenge becomes more pronounced for distributed, unattended EDCs, where potentially longer downtimes translate to higher costs.

Our objective is to enhance fault-tolerant management in these infrastructures, enabling predictive, preventive, or automatic mitigation of outages and equipment failures. Given the rarity of these situations, only a small percentage of the collected data includes anomalies. Consequently, ML algorithms trained on such datasets exhibit bias, hindering management systems from effectively learning to make decisions during failures.

Reproducing an anomaly in an operational DC is potentially risky for the equipment and may entail significant time and energy consumption. In this paper, we propose leveraging ML to optimize system management and generate balanced datasets by generating realistic synthetic data that replicate anomalous situations. This presents a significant challenge due to the complex, multivariate nature of DC scenarios. For synthetic datasets to be realistic, they must maintain nonlinear correlations between variables when generating on-demand anomalies.

4.3 | Significance and innovation of our integrated MSOBSE and ML approach for the edge

Edge computing infrastructures are complex Systems of Systems (SoS) that serve multiple applications with diverse technical requirements and are closely linked to other complex systems. Relevant stakeholders for edge infrastructures include customers seeking low-latency services, Information and Communications Technology (ICT) companies looking to expand their offerings, and Internet Service Providers (ISPs) that can benefit from optimized network traffic flow. KPIs for evaluating edge computing solutions include perceived QoS for customers, energy efficiency for infrastructure owners, and network resource utilization for ISPs.

In this context, numerous research studies delve into enhancing fault tolerance within edge infrastructures. Wang et al.[21] propose a fault-tolerant real-time messaging architecture aimed at effectively addressing varying levels of message loss tolerance requirements and mitigating latency penalties arising from fault recovery in a local edge computing test-bed. Tuli et al.[22] develop an AI model utilizing a Generative Adversarial Network (GAN) to predict preemptive migration decisions for proactive fault-tolerance in containerized edge deployments. Their approach involves training a few-shot anomaly classifier to anticipate migration decisions, ensuring reliable computing in a Raspberry-Pi based edge environment. Mudassar et al.[23] introduce a fault-tolerance methodology based on checkpointing and replication for edge computing, augmenting the system's overall reliability.

Moreover, edge computing, in addition to complementing cloud infrastructures, holds the potential to enhance energy efficiency through smart grid technologies. Feng et al.[24] explore the synergistic effects arising from the integration of edge computing with smart grid. The reciprocal relationship between these two domains enables mutual benefits, supporting the sustainability of both technologies. However, the relative youth and immaturity of edge computing technology give rise to open questions in its integration with other systems, which necessitate further research and development. Various aspects offer opportunities for improvement, encompassing system design, algorithms, resource management, and hardware accelerations within the combined environment. As their implementation advances, it is expected that both promising opportunities and technical challenges will emerge. Currently, the forefront of research primarily revolves around enhancing fault tolerance in edge device infrastructures characterized by constrained resources and limited energy consumption evaluation, primarily concentrated on IT power. However, certain critical aspects, such as the infrastructure's cooling requirements, its physical footprint, and the integration of other systems as Smart Grids, have been overlooked in this state-of-the-art approach. To improve these strategies, we can perform a comprehensive assessment of their impact. This can be achieved by integrating them into a system model that considers advanced hardware, accounts for IT and cooling consumption and thermal behavior, and accommodates complex network, energy storage, and energy generation infrastructures.

The adoption of MSOBSE strategies plays a pivotal role in simplifying the design and implementation of these SoS, facilitating their integration, and optimizing various approaches. We propose an M&S&O standardization framework based on formal methods that offers an interface for seamlessly integrating models and optimization strategies into more intricate infrastructures. This approach enables iterative refinement of the system through simulation as it evolves during development. Our vision aims to optimize the deployment and automatic operation of future infrastructures, thereby reducing costs, mitigating risks, and minimizing time to market. The primary advantages of our approach can be summarized as follows.

4.3.1 | Formal foundation and model modularity

A modeling approach that relies on a mathematical formalism provides greater completeness and robustness to the system model. Consequently, it facilitates the timely detection of errors in the conceptual model. This strategy enables the development of the system at an earlier stage and with reduced costs.

In order to accurately capture the real physical behavior of the system under study and to incorporate it into the mathematical formalism of the conceptual model, different strategies can be used, such as analytical modeling or ML-based solutions. A critical challenge for the success of the MSOBSE strategy is the development of a robust and reliable model. However, validating complex models against how real systems behave can be difficult to achieve, and the accuracy and reliability of simulation results are highly dependent on the correctness of the underlying models.

Therefore, our research relies on characterizing prototypes that operate under real-world workload conditions. We collect data by monitoring the equipment, curate it, and use it to feed the appropriate modeling algorithms. The accuracy

of the obtained behavioral models is tested against the monitored data from the real prototype. The verified models, whether they are analytical or expressions for inferring neural networks, are then used to infer or predict the value of different parameters in the computational model of the MSOBSE strategy.

Moreover, modeling a complex system is an iterative process. System components are modeled and verified separately. Then, we progressively merge the verified models and prove the resulting model again. In such scenarios, a modular modeling technique proves advantageous as it simplifies the integration and verification process of the edge infrastructure with pre-existing technology (e.g., clouds, IoT devices, cooling systems, and network elements).

4.3.2 | Hardware-in-the-loop integration

The proposed workflow integrates HIL modeling techniques in the system implementation and validation phases. Our goal is to establish a modeling framework that facilitates a seamless transition from a computational model to a HIL model without introducing additional development time. This approach enables the replacement of computational models with real hardware, allowing the MSOBSE framework to operate in pure simulation contexts, real-world scenarios, or hybrid environments. Consequently, the framework can comprehensively explore the entire system's behavior, taking into account its impact on both formal models and actual equipment (e.g., resource usage, energy consumption, and QoS perceived by the users).

4.3.3 | Fault-tolerant decision making

Finally, the MSOBSE workflow must provide a complete interface for optimizing the system under development. One of the biggest challenges for the industry to include AI-driven optimizations is to get massive amounts of data to train the algorithms properly (especially for DL-based approaches),[25] as it is expensive in terms of costs, time, and resources. This issue is aggravated when ML-driven optimizations must make decisions in anomalous situations since these are sporadic and the amount of data is limited.

As shown in Figure 5, ML can be seamlessly integrated with the MSOBSE framework to produce unbiased/biased training environments. The ML-based synthetic data generator allows the MSOBSE framework to train with sufficient data from standard system behavior and anomalies. This approach ensures that the resource management optimization algorithms learn to make decisions in scenarios of system failure, drawing insights from the impact observed in the formal model. Generative Adversarial Networks (GANs) allow a high degree of control of the generated scenarios through the latent space, which makes them especially interesting for augmenting biased datasets with on-demand

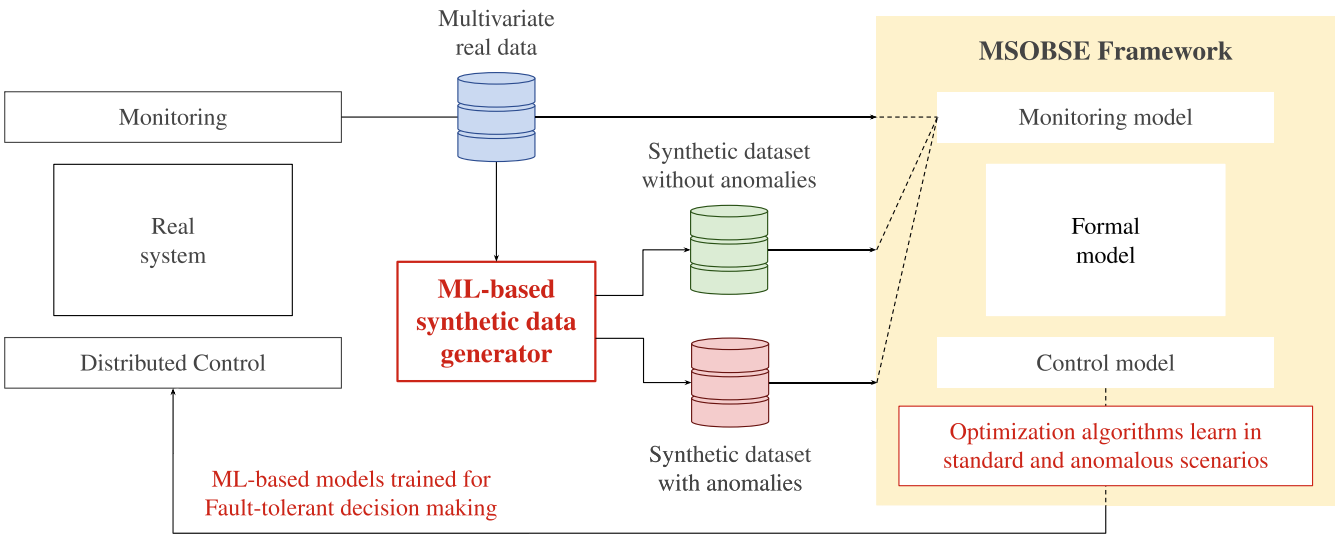

FIGURE 5 Workflow of the integrated approach for fault-tolerant decision making.

anomalies, generating realistic synthetic scenarios that maintain the correlations between variables.[26] Our proposed approach aims to provide trained ML-based models that can assist in energy-efficient fault-tolerant decision-making in real edge infrastructures.

5 | EDGE FEDERATIONS FOR COLLABORATIVE RESOURCE MANAGEMENT

The profile of service requests perceived locally at the network's edge is significantly variable owing to user mobility. Dimensioning EDCs to meet peak demand is impractical due to associated costs and inefficient resource usage. However, ensuring service availability is imperative, given the critical nature of the edge target applications. Consequently, collaboration becomes essential for the edge to effectively accommodate peak demand while guaranteeing QoS. In this paper, we advocate for the federated management of EDCs to enhance energy and cost sustainability while preserving the QoS perceived by users.

Considering the nature of the applications for which the edge is to be deployed, we suggest a model in which computational offloading follows a Function-as-a-Service (FaaS) approach.[27] When end nodes request to open a new service session, the available resources are evaluated, and the resources required by the service are reserved to grant computational offloading while the session is active. When the session is terminated by the end devices, the reserved resources are released and become available for new service session requests.

All EDCs within the same RAN work together and share the current state of their resources. Requests from end devices are usually processed by the closest EDC, especially for latency-critical services. For services that can tolerate higher latency, requests may be forwarded to the EDC with the best energy scenario according to the policies in place (e.g., higher renewable generation, better energy efficiency, lower energy price). However, if one of the EDCs experiences a shortage of resources due to an abnormal peak in demand, incoming requests will be redirected to the next most suitable EDC in the federation.

Furthermore, in the event of congestion at all EDCs in the RAN, EDCs have the capability to forward requests to the public cloud.[27] This ensures that customers whose requests are forwarded to the cloud will not have to wait until an EDC becomes available, even if they experience longer than usual delays. Additionally, computational offloading to the cloud may be necessary if part of the workload is designed to be processed there.

Finally, to take advantage of multi-regional edge computing, computation can be offloaded across edge-cloud DCs to leverage price volatility and intermittent renewable energy. In this case, it is essential to investigate how the virtual network topology and the choice of physical network infrastructure affect offloading costs in terms of delay and energy with an increase in the amount of data transfer.[28]

In this context, heterogeneity poses an inherent challenge in federated management. As shown in Figure 6, the industry recognizes new cooling systems and renewable energies as primary sustainability drivers.[6] The progressive adoption of emerging cooling technologies (such as liquid cooling, single-phase, and two-phase immersion) and green energy sources (including solar photovoltaic, wind power, or green hydrogen) is expected to augment the physical heterogeneity of DCs.

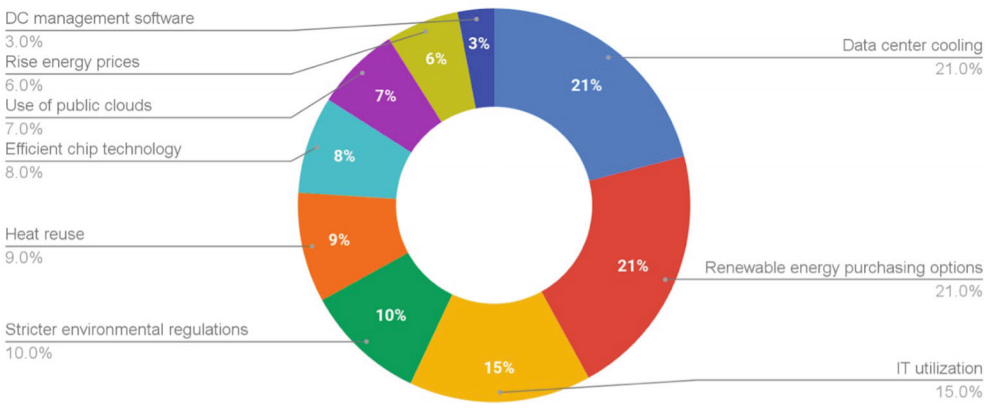

FIGURE 6 Drivers for sustainability gains in the DC industry.

Additionally, the geographical dispersion of EDCs, situated closer to users, introduces another layer of heterogeneity for federated management. Factors such as weather conditions, energy prices, smart grid capabilities, local user demand, and their fluctuation profiles are contingent upon the specific location of each edge infrastructure.

On the other hand, IT utilization emerges as the third major driver perceived by the industry for sustainability improvement, as depicted in Figure 6. Presently, significant efforts are directed towards this goal in cloud DCs[28] given that their average utilization does not surpass 60%.[29] However, this challenge may be even more pronounced for the network's edge infrastructure. Enhancing IT resource utilization presents an inherent challenge in federated edge infrastructures supporting critical applications. Overprovisioning computing equipment to ensure QoS diminishes resource utilization, exerting a negative impact on energy efficiency. Even completely idle servers can consume up to 70% of their maximum power.[30]

5.1 | Significance and innovation of the proposed vision for edge federations

Recent surveys emphasize the crucial role of federated management in edge computing to ensure its optimal functionality. The state-of-the-art research in the field is centered on decentralized and federated AI-based resource management, driven by the extensive distribution and scale of networks that can encompass thousands of devices.

In their survey, Kar et al.[31] identify several criteria for task offloading decisions: (i) Resource constraints, which occur when a task demands more computing resources than the local system can provide. (ii) Addressing latency, as the distance between devices can impact time-sensitive applications. (iii) Load balancing, necessary when the IT infrastructure reaches its maximum capacity, and additional tasks need to be distributed among other entities. (iv) Ensuring privacy, confidentiality, and security, particularly for sensitive data. These findings align with the perspective presented in the survey by Iftikhar et al.[32] The authors highlight that task offloading and resource allocation primarily revolve around response time and energy constraints. Load balancing, on the other hand, is determined based on CPU load, memory usage, latency, and network load, with a focus on meeting user needs.

However, our vision seeks to introduce a new dimension to resource management decision-making in edge federations. This involves considering not only cooling consumption but also energy pricing and smart grid technologies (energy storage and generation models). By incorporating these aspects, we can exploit the heterogeneity of the distributed edge to further optimize energy consumption while ensuring the desired QoS. Furthermore, the edge environment facilitates the execution of specific applications with diverse profiles in terms of load, schedules, and QoS requirements. Our research aims to leverage this feature through inter-agent collaboration to enhance overall resource utilization. Below are additional details regarding our vision to enhance the sustainability of the edge by capitalizing on heterogeneity and secondary market models.

5.1.1 | Leveraging heterogeneity to enhance energy efficiency

Heterogeneity presents a significant opportunity to enhance the sustainability of edge computing through federated management. We propose expanding resource management systems to incorporate the energy impact of scheduling and configuration decisions within the federation. This approach allows the federation to leverage the heterogeneity of each EDC in its specific location at any given time while maintaining awareness of its status as follows.

- **Utilization of IT equipment:** Dynamically driven by the assigned workload.
- **IT power consumption:** Varied based on IT technology (e.g., CPUs, GPUs, etc.), resource utilization, and configuration of their low-power policies.
- **Cooling power consumption:** Subject to fluctuations in the heat generated by IT equipment, cooling technology, and outdoor climatic conditions (i.e., temperature and humidity).
- **Renewable energy production:** Dependent on the deployed energy generation technology and weather factors (i.e., solar radiation intensity, wind intensity).
- **Energy storage:** Determined by storage technology characteristics and charging and discharging policies.
- **Grid energy pricing:** Established by energy companies, contingent on users' daily consumption patterns.
- **QoS requirements:** Specifically agreed upon for the diverse applications and services supported by the federation.

5.1.2 | Leveraging secondary market models to improve resource utilization

We also advocate for the integration of the secondary market model in federated edge management, where resources are leased to other companies, offering two main advantages. The first is to lease tenants' idle resources to enhance resource utilization. The second is to rent or sell outdated infrastructure to extend its service life. When an edge federation is no longer useful for a particular application, it can still be repurposed to serve another application with different resource and QoS requirements. This model reduces the need for companies to establish their edge infrastructure by leveraging the computing resources of existing EDCs.

What we are proposing is a shift from a CAPEX-intensive to an OPEX-intensive approach, where the main benefit is derived from software developments. The OPEX-intensive nature of this approach is significant. On the one hand, it reduces barriers to entry into the market, enabling not only the largest companies but also smaller players to compete. Additionally, it allows for the extension of the infrastructure's useful life and maximizes its utilization. Furthermore, the implementation of a secondary market model has the potential to enhance competition in the DC industry and facilitate deployment in less developed countries.

However, the incorporation of this model introduces challenges, particularly in ensuring that workloads coexist with applications that have critical QoS and security requirements. The federation stands to gain from the secondary market model by considering of the following aspects for each type of workload.

- **Application resource utilization profile:** For example, CPU, memory, I/O, network and so forth, and their fluctuation due to user interaction.
- **User demand profile:** Varied in different patterns such as daily, weekly and so forth.
- **Contracted QoS:** Determined by the nature of the application (best effort vs. critical applications).

Models facilitating detailed predictions of these patterns will be crucial in ensuring QoS to users. Additionally, predicting energy-related variables based on the evolution of resource demand would enable proactive management strategies. Our objective is to enhance the sustainability and efficiency of the edge operation by capitalizing on heterogeneity and secondary market opportunities while upholding the QoS of the applications.

6 | ENVISIONED INTEGRATED APPROACH

The establishment of a sustainable edge requires addressing three key challenges: sustainable deployment, fault-tolerant automated operation, and collaborative resource management. The economic and energy sustainability of a competitive edge infrastructure depends on addressing these open research questions.

6.1 | State-of-the-art vision analysis

At a general level, the state of the art agrees with these lines of action and identifies them as challenges still to be solved and that require further study in energy efficiency.[25,33] First, in terms of sustainable deployments, there is interest in cooling infrastructure consumption. However, reducing cooling requirements while maintaining DC temperature focuses on the application of AI-based energy-efficient techniques in DCs with traditional cooling systems.[34] In addition, there is interest in improving resource efficiency, resource scheduling, and provisioning techniques to reduce IT energy consumption.[25]

Second, with respect to automated operations, there is a strong motivation in the state of the art to explore the combination of AI and edge computing.[25] On the one hand, the challenges of automating the optimization of consumption, scheduling, latency, privacy, and security are prominent. With the proliferation of edge devices, the problems, which are already complex and non-convex, scale rapidly.[35] ML is better suited than traditional methods for solving these types of problems.[36] The other challenge stems from the fact that the nature of the edge makes it a good candidate for performing AI-based tasks closer to the user to improve user experience and reliability.[37]

Third, collaborative resource management through computation offloading at the edge is a frontier for the state of the art.[25] Computation offloading, which has traditionally been applied in the cloud, assumes as a priority to preserve the resources and consumption of the end devices by moving the processing to the DC where access to the power grid

is not a problem. However, in edge computing, EDCs may be also resource and power constrained.[38,39] In this context, current research considers collaborative resource federation between the edge and the cloud as a topic with significant potential.[28,40] Real-time task scheduling, service orchestration considering resource availability, intermittent renewable energy, QoS, and smart grids are marked as hot topics to be explored (not yet approached) in many research efforts.[34,41]

One of the major limitations of the state-of-the-art vision is that it does not consider that energy-aware optimization techniques (i.e., resource efficiency improvement, resource scheduling, resource provisioning, latency reduction, and computation offloading, among others) will be used together to optimize IT and cooling in the face of emerging cooling schemes. In fact, the application of novel two-phase immersion cooling techniques, as discussed in Section3.2, is still limited to some use cases. The size of EDCs is also not considered as an additional constraint in sizing their IT and cooling resources. Additionally, these optimization techniques are not expected to simultaneously account for the high heterogeneity at the edge (e.g., access to intermittent renewables, energy pricing).

Furthermore, the feasibility of these optimization techniques is usually explored in the simulation world based on models of device behavior. The current state of the art neglects two important issues. First, when training optimization algorithms, the bias of the data feeding them is typically overlooked. If this data does not contain possible anomalies that may occur in the infrastructure, these algorithms will not have learned to make reliable decisions in these situations, and the fault tolerance of the system will not be ensured. On the other hand, the leap from the world of simulation to the real world scenario is very large, even more so for a SoS as complex and distributed as the deployment of edge computing.

6.2 | Proposed integrated approach

In this research, we outline three approaches to addressing these challenges, with the goal of collectively delivering the dense, compute-intensive edge infrastructure needed to support the widespread digitization of society.

To improve the environmental and economic sustainability of deployments, we advocate the incorporation of innovative cooling systems, such as two-phase immersion cooling. The technology's ability to reduce power consumption, physical footprint and cost, regardless of environmental conditions such as temperature and humidity, makes it particularly compelling for improving sustainability at the network edge.

The robustness, modularity, and integrability inherent in formal modeling within an MBSE framework can significantly augment automated decision making in complex environments. We propose an M&S&O framework that allows to (i) integrate and validate behavioral models of the different parts of the system, (ii) simulate the system operation to explore the global impact of the proposed policies, (iii) inject synthetic data for ML-based strategies to learn to make decisions in different scenarios, and (iv) reduce the leap to the final real system through HIL, allowing the replacement of computational models with real hardware for real-world scenarios or hybrid environments.

The diversity of EDC configurations (i.e., IT systems, cooling methods, renewable energy generation systems, infrastructure locations) and variable external factors (such as user demand, mobility, energy price, and weather) collectively influence the energy consumption, QoS, and costs associated with edge computing. Therefore, our proposal involves exploiting this heterogeneity through edge federations. Computing can be offloaded across edge-cloud datacenters to enable collaborative management that is resilient to peak user demand, price volatility, and intermittent renewable energy.

Figure 7 illustrates the high-level architectural framework that guides our integrated approach. Our framework encompasses the infrastructure that is essential for the future sustainable delivery of critical IoT services, including:

1. IoT devices. Integrates end nodes with user demand profiles appropriate to the applications for which they are deployed. They enable user mobility and are responsible for requesting computation offload to the federation.
2. Network Connectivity and Core Network. These include the technology solutions for providing network connectivity in the system, such as the radio interface and crosshaul. They enable edge-cloud computation offloading and latency-aware federated resource management.
3. Edge-cloud federation. Integrates EDCs and cloud Data Centers (DCs), including their compute and cooling resources (two-phase immersion cooling or air-based). Enables energy, location, and smart grid-aware policy configuration.
4. Smart Grid. Includes renewables, storage, and management. It allows the configuration of energy management policies with awareness of price profile and renewable energy generation.
5. MBSE Framework: It is an M&S&O framework, so it integrates:
   - The conceptual model of the overall system, which may include the aforementioned subsystems. Each subsystem's formal model, in turn, contains its respective behavioral model, which may have been obtained through analytical

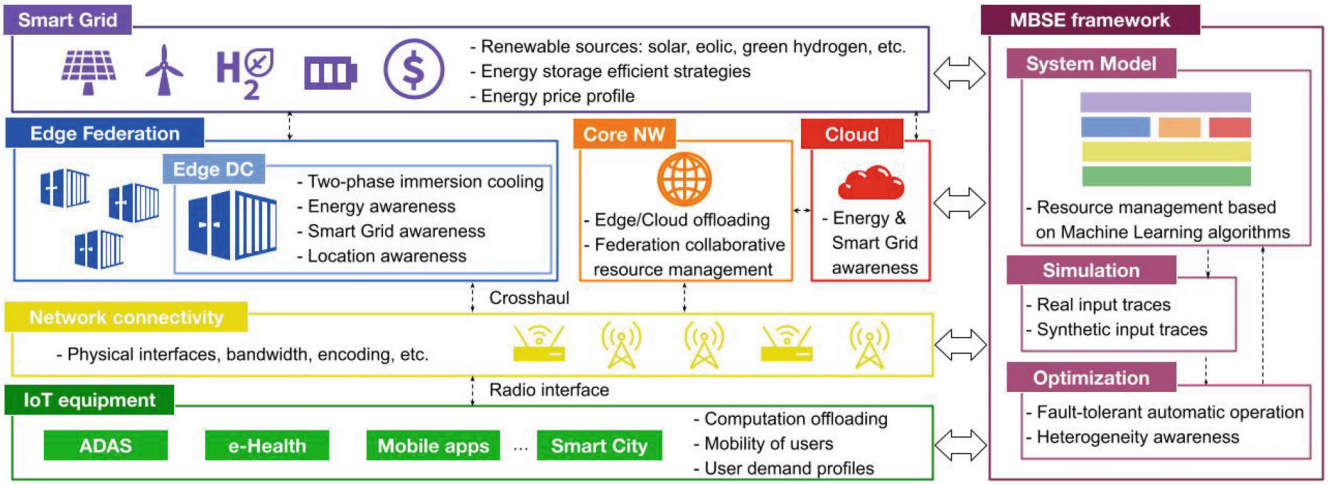

FIGURE 7 A distributed environment for IoT applications integrating two-phase immersion cooling and a modeling, simulation, and optimization framework for edge federations.

or ML-based modeling techniques. The conceptual model is then translated into a computational model (i.e., a computer program). Simulations can be run on the computational model to test the impact of the implemented policies.

- A simulation tool for executing the computational model of the system. The tool allows replacing parts of the computational model with parts of the real system, enabling the framework to operate in pure simulation contexts, real-world scenarios, or hybrid environments. It also allows the injection of traces from the real system and synthetic traces. This makes it possible to include anomalous scenarios to train ML-based algorithms and improve the fault tolerance of the system.
- A scenario optimization framework that allows improving the scenario configuration and its heterogeneity, such as the number of EDCs deployed, their location, their IT sizing, cooling, renewables and batteries. It also allows automatic optimization of the policies deployed in the system to improve its fault tolerance, for example by avoiding outages of critical services.

Our proposed vision synergizes the benefits of simulation and real-world applications to enable holistic evaluation of environmental and economic sustainability policies in distributed environments for future critical IoT applications. This integrated approach is designed to make the next edge, as transformative as society needs it to become.

7 | USE CASE: SUSTAINABLE EDGE COMPUTING FOR ADVANCED DRIVER ASSISTANCE SYSTEMS

Advanced Driver Assistance Systems (ADAS) have emerged from the imperative for a safer driving experience, aiming to reduce accidents and casualties. The primary goal of ADAS is to predict and prevent emergencies by leveraging data collected from vehicles, their surroundings, and their occupants. As per Nvidia,[42] this application generates approximately 2 gigapixels of data per second per vehicle, requiring a processing capacity of 250 trillion operations per second. Incorporating IT systems into vehicles to achieve this computational capacity renders the technology solutions prohibitively expensive, impeding their commercial development. Conversely, transmitting data to the cloud for processing introduces service latency and poses potential risks to the physical integrity of vehicles and their occupants. Therefore, deploying a computing infrastructure at the network's edge becomes indispensable for ADAS to effectively manage its substantial data volumes and address the criticality of processing delays.[43]

Additionally, performing computation offloading in close proximity to the data sources facilitates collaboration between vehicles, Smart Cities, and pedestrians, thereby enhancing the mobility experience. This unlocks the full potential of integrating ADAS into an edge infrastructure for mobility management, allowing the city to be configured based on real-time needs. For instance, in the event of detecting a risk situation, it becomes possible to prevent the escalation

of accidents and enhance the speed of evacuation by coordinating the management of pedestrians, vehicles, and city infrastructure.

Consequently, ADAS, serving as a precursor to autonomous vehicles, stands out as a pivotal application driving the development of edge computing. Hence, exploring the sustainability of its deployment and operation becomes crucial. In this regard, EDCs can benefit from utilizing two-phase immersion cooling to significantly reduce their energy consumption. Additionally, they can improve their resource management strategies and resilience to energy price fluctuations through the use of federated management and smart grids. Smart grid-aware edge computing federations require multi-domain solutions. MBSE facilitates the integration of modeling and simulation (M&S) tools into the system development process.[44] It assists in the exploration and validation of the system of systems under study, providing insights into functionality and potential technical risks of the solution while reducing cost.

This section outlines our initial steps towards ensuring the sustainability of an edge infrastructure for ADAS. These steps encompass: (i) designing a prototype EDC with two-phase immersion cooling, exploring its thermal and power behavior, (ii) modeling the EDC prototype and simulating its federated operation to harness cooling heterogeneity, and (iii) modeling and simulating smart grid-awareness in the federation.

7.1 | Prototype of a two-phase immersion cooling EDC for ADAS

The objective of this use case is to demonstrate that our approaches can contribute to the development of a sustainable edge infrastructure for the digitalization of society. Since 2017, we have collaboratively developed an EDC prototype with the industry to explore the viability and benefits of two-phase immersion cooling technology for edge computing. Our prototype, situated at the Technical School of Telecommunications Engineering at Universidad Politécnica de Madrid, Spain, is a small shipping container adapted as a data room, as illustrated in Figure 8. The immersion tank inside the container accommodates a GPU computing infrastructure tailored for running ML-based applications. The dry cooler positioned on top of the container serves as the heat exchanger with the environment. Our EDC prototype is designed for an IT infrastructure of up to 50 kW.

Figure 9 illustrates the installation of the cooling system outside the container, encompassing the expansion vessel, the drive pump, the dry cooler, and the piping essential for its operation. In this prototype, the sole factor contributing to cooling consumption is the pump responsible for circulating water from the condenser to the dry cooler in a closed circuit. Consequently, the cooling model is contingent solely on the difference in outlet and return temperatures to the immersion tank (T_{out} and T_{in} respectively) and the ambient temperature (T_{amb}).

FIGURE 8 Two-phase immersion cooling EDC prototype.

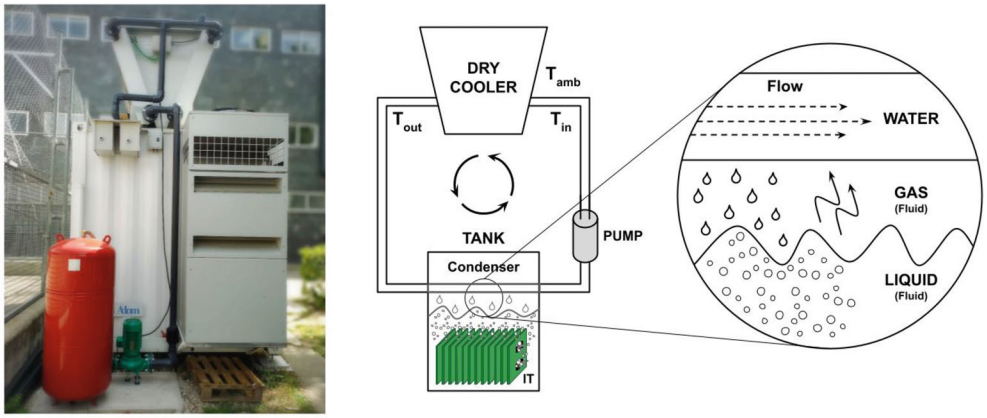

FIGURE 9 Simplified two-phase immersion cooling EDC.

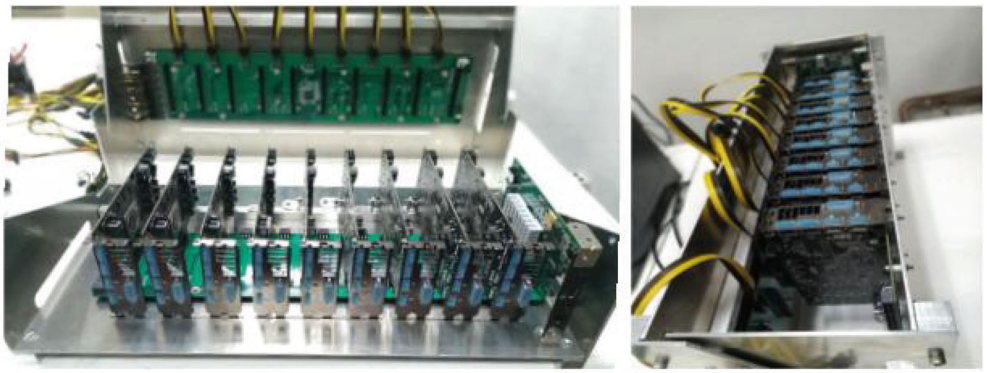

FIGURE 10 Custom rack development for the prototype.

According to the specifications of the Intersam model IDV-130 V-cooler used in the prototype, the specified temperature ranges that interface the computing and cooling systems are $\Delta T_1 \in [0, 20](^{\circ}C)$ and $\Delta T_2 \in [0, 8](^{\circ}C)$, where $\Delta T_1 = T_{in} - T_{amb}$ and $\Delta T_2 = T_{in} - T_{out}$. The Aero-cooler is designed to work in an air-cooled environment where the temperatures are much lower than those found in immersion cooling. Therefore, we will operate at the upper limit of the supported ranges (i.e., if $T_{amb} = 32°C$, then $T_{in} = 52°C$ and $T_{out} = 44°C$). In this scenario, we have used a Wilo IPL 50/115-0.75/2 drive pump. As the drive capacity is oversized, the cooling power will be mainly defined by the minimum flow rate it can provide, which is sufficient for the operating temperature range of the system. Therefore, the cooling power will be around 1300 W in most cases, as discussed in our previous research, which presents the pump's power consumption model (P_{EDC_COOL}) used in this work.[13] This allows us to achieve a theoretical PUE of about 1.03 within the state-of-the-art two-phase immersion technology. Moreover, even for a resource utilization level similar to that found in clouds (60%), the PUE remains around 1.04.

For the computing infrastructure, a custom rack has been developed to submerge the GPUs (Sapphire Pulse Radeon RX 580) in the dielectric fluid. This rack ensures the mechanical security of both the GPUs and their controllers. In our prototype, we eliminated the air-cooling components of the IT, such as fans and heatsinks, which are typically bulky, especially for GPUs. This modification enables the interleaving of two racks, each containing 9 GPUs, thereby enhancing computing density, as illustrated in Figure 10. It is devised to stack multiple units vertically and horizontally in the immersion tank, achieving high computational densities. This configuration allows us to support up to 50 kW in the container. GPUs execute deep learning (DL) algorithms with significantly greater performance than CPUs, making GPU clusters among the top choices for IT infrastructures in real-time video processing applications.[45]

7.1.1 | Description and power profile of the ADAS workload

Our ADAS application[46] is a DL-based service that notifies the user if a distraction is detected while driving. The system captures video footage from inside the vehicle and processes it through a Convolutional Neural Network (CNN) to estimate whether the driver has diverted attention from the road. Due to the computationally intensive nature of DL algorithm training, these processes are offloaded to the edge layer, reducing costs associated with in-vehicle hardware systems. This practice is commonplace in both industry and research, with ongoing exploration of the advantages of Federated Learning to enhance edge computing training efficiency. Models trained at the EDC with user video footage are subsequently transmitted to the vehicle, thereby enhancing prediction quality. Proximity of edge infrastructure to users allows for more frequent model updates with minimal delay. In the event of a risk situation, the edge federation can inform other vehicles, contributing to the improvement of road safety.

The power profile of the workload running on the IT infrastructure of the EDC varies dynamically and depends on the number of users. Additionally, two-phase immersion coolants exhibit turbulent performance when the boiling point is reached. In this study, we utilized Novec 7100, an engineered fluid with a boiling point of 61°C.[13] Therefore, ML serves as an excellent alternative to obtain highly accurate power models to help estimate the workload behavior in our EDC.

We executed the ADAS workload while modifying the main control variables (such as GPU clock frequency settings) and for different numbers of user parallel sessions, in a two-phase immersion setting. Using this data, we modeled the IT power with a FeedForward Neural Network.[13] The result was an accurate power model, $P_{IT}(t)$, as depicted in Figure 11, with a normalized root-mean-square deviation (NRMSD) of about 3% and an R^2 of 98%.

7.2 | Modeling EDCs with two-phase immersion cooling

To facilitate the automatic operation of EDCs within a federation, we propose to apply our M&S&O methodology. Specifically, in this paper, we use Mercury,[47] our MSOBSE framework, to assist optimization algorithms in comprehensively exploring system behavior and automatically enhancing decisions based on formal modeling. Our framework offers detailed models for various system elements, spanning from vehicles to EDCs federations. We have modeled EDCs to resemble the behavior of our two-phase immersion prototype, including the obtained IT power and cooling models.

Figure 12 illustrates the models that define the system, delineating the various levels of granularity requisite for the EDC federation. $P_{EDC\_IT}(t)$ represents the dynamic aggregated consumption of all GPUs in the immersion tank. $P_{EDC\_COOL}(t)$ denotes the dynamic cooling consumption based on the pump's flow rate ($\phi_{EDC}(t)$), dependent on the temperature difference provided by the dry cooler (T_{in}-T_{out}) and $P_{EDC\_IT}(t)$.[13]

7.3 | Modeling the smart grid for edge federations

We have incorporated a smart grid formal model into our MSOBSE framework, as illustrated in Figure 13. The energy provider (PROVR) supplies electricity to the smart grid at a variable price ($E_{price}(t)$). Solar panels generate free green power ($P_{solar}(t)$) with a variable profile based on solar irradiation. If this generated power exceeds the total power demand of the

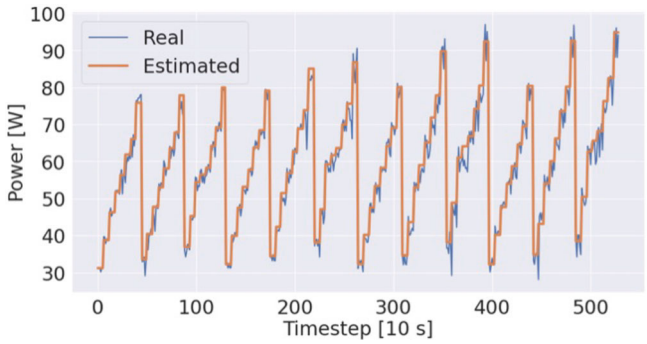

FIGURE 11 Power model estimations for the IT equipment in the immersion-cooled EDC.

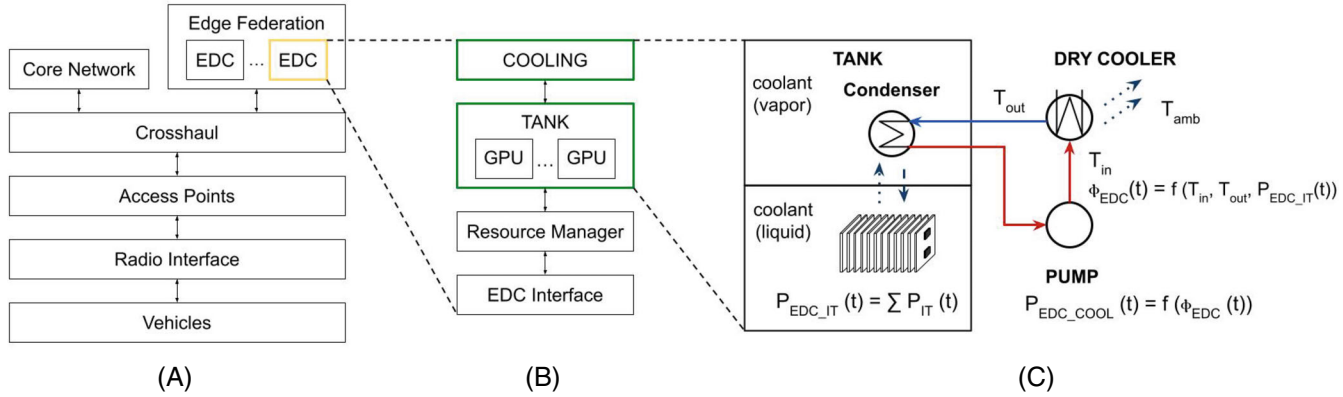

FIGURE 12 Conceptual model of the system. (A) Edge federation model, (B) edge data center model, (C) two-phase immersion cooling model.

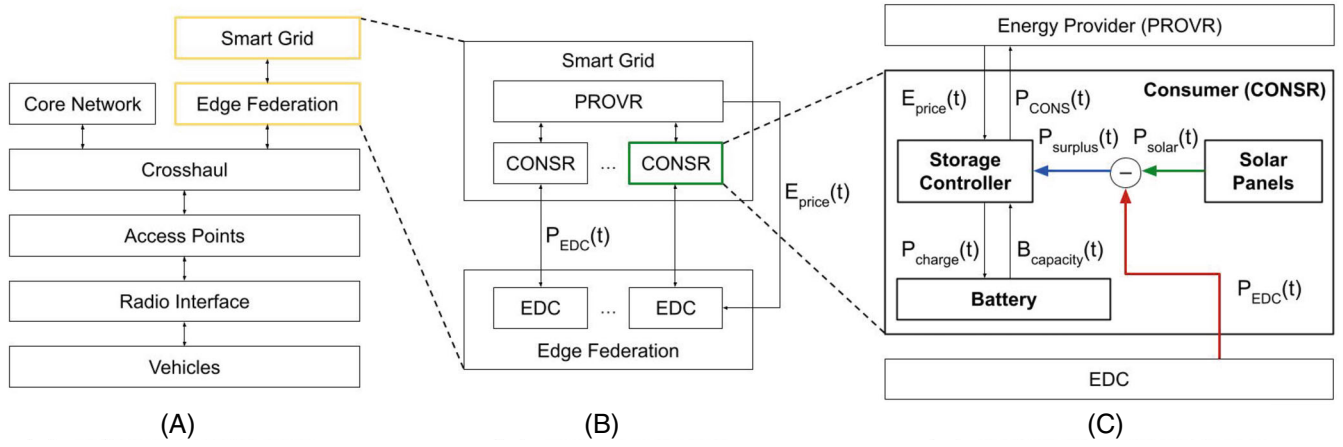

FIGURE 13 Conceptual model of the smart grid. (A) Edge federation model, (B) smart grid model, (C) smart grid consumer model.

EDC ($P_{EDC}(t) = P_{EDC_IT}(t) + P_{EDC_{COOL}}(t)$), a surplus of energy ($P_{surplus}(t)$) is anticipated. Each EDC in the federation has a smart grid consumer model (CONSR) that incorporates its solar panels, a battery, and a controller. The storage controller makes charging and discharging decisions for the battery ($P_{charge}(t)$) based on its capacity ($B_{capacity}(t)$), the energy price, and the surplus. The controller also manages how much electricity is consumed from the grid ($P_{CONS}(t)$).

Figure 14 displays the flowchart of the storage controller's behaviour. The controller makes decisions based on two thresholds. On the one hand, $E_{THRES\_CHARGE\_PRICE}$ is the energy price below which the consumer decides to charge the battery (grid energy price is so cheap than they can benefit from storing energy in the battery). On the other hand, $E_{THRES\_DISCHARGE\_PRICE}$ is the energy price above which the consumer decides to discharge the battery (since the price is so expensive that it is better to use the stored energy than to pay for it to the provider). The battery also has a maximum charge rate of $P_{MAX\_CHARGE\_RATE}$ and a maximum capacity of $B_{MAX\_CAPACITY}$.

When provided energy price is lower than $E_{THRES\_CHARGE\_PRICE}$, as long as the battery is not at full capacity, the battery is charged at the maximum charge rate. On the other hand, when the price exceeds this threshold, the controller will act based on the surplus calculation. If the solar panel generates more energy than the EDC demand, providing that the battery is not fully charged, then the surplus energy will be used to charge the battery at the minimum of $P_{MAX\_CHARGE\_RATE}$ and the surplus. Otherwise, when the surplus is negative and the grid energy price is higher than $E_{THRES\_DISCHARGE\_PRICE}$, the battery will be discharged to the maximum rate between $-P_{MAX\_CHARGE\_RATE}$ and the surplus. For further details on the controller management workflow and EDCs resource manager policy, refer to our previous work.[27]

This model allows us to explore the coordinated management of power generation, storage, cooling, and computing capabilities for ADAS. Our goal is to minimize the energy consumption of the federation providing the ADAS service by leveraging the heterogeneity of EDCs' locations, their environmental conditions, and the energy price.

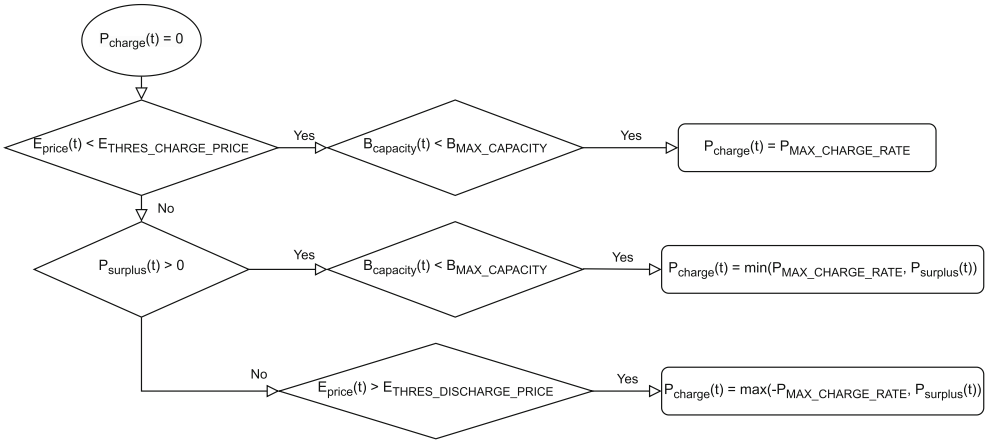

FIGURE 14 Flowchart for the battery model.

7.3.1 | Performance evaluation

We present the evaluation of our proposed approach, which combines two-phase immersion cooling, MSOBSE, ML, and federated management with the objective of reducing the impact of energy price fluctuations on the overall ADAS service cost. To achieve a controllable environment, we have chosen the Mercury MSOBSE framework[47] to simulate a federation of EDCs. In this work, we have extended the simulator to include the two-phase immersion cooling infrastructure and the smart grid models explained previously. We leverage the DEVS formalism[48] to delineate the conceptual model, from which computational models are derived.[49] Our framework employs the xDEVS simulation engine[50] for simulating the computational models.

To evaluate the performance of our approach, we utilize a computation offloading service demand that corresponds to real mobility traces of 535 taxis in the San Francisco Bay Area, USA.[51] The workload involves running the ADAS application described in Section 7.1.1. Our experimental setup includes a federation of 3 EDCs, each containing 20 AMD Sapphire Pulse Radeon RX 580 GPUs, served by 19 access points. To determine the location of access points and EDCs, we used Mercury's allocation manager tool. In this way, the sessions are divided into three different geographic areas where each EDC can provide the ADAS service to up to 100 vehicles. The electricity price profile corresponds to the average hourly day-ahead wholesale energy price in California in 2017.[27] The power generation profile for the solar panels was obtained from the PVGIS-NSRDB database based on the location of these EDCs. The storage controller is configured with charge and discharge thresholds for the battery of $E_{THRES\_CHARGE\_PRICE} = 20\$/Wh$ and $E_{THRES\_DISCHARGE\_PRICE} = 35\$/Wh$ respectively. The battery is a PylonTech US3000C model. It has a maximum charge rate of $P_{MAX\_CHARGE\_RATE} = 1.78$ kW and a maximum capacity of $B_{MAX\_CAPACITY} = 3.37$ kWh. Further details on scenario configuration can be consulted in our previous work.[27]

We conducted a 24-h scenario simulation in which the ADAS application is served from the edge infrastructure with a variable demand profile that depends on the vehicles in the coverage range of the system's access points. In this experiment, the ADAS service is delivered to each vehicle from its nearest EDC to isolate the benefits of the smart grid. Figure 15A shows the charging and discharging profile of the batteries of the three EDCs of the federation. The storage controller takes advantage of the off-peak hour to charge the batteries when the energy price is minimal. During peak hours, when the price is higher, the controller prioritizes using stored and solar energy. This increases the demand for grid power during off-peak hours, minimizing its use during peak hours, as shown in Figure 15B for the federation. Thus, the cost profile of the federated EDCs, presented in Figure 15C, only shows steep rises when the computation required by the vehicles demands power consumption that exceeds the solar generation and battery storage capacity.

The baseline used for comparing the results of our work involves the same scenario but excludes the integration of any aspect of the presented smart grid model (i.e., energy price awareness, energy storage systems, and renewable energy production). Our results demonstrate that leveraging smart grid heterogeneity for ADAS offers two main benefits. First, it can reduce federation energy consumption by 20% due to the solar photovoltaic system installed in the EDCs. Second, operating costs can also be reduced by 30% due to the energy storage management, which effectively harnesses renewable

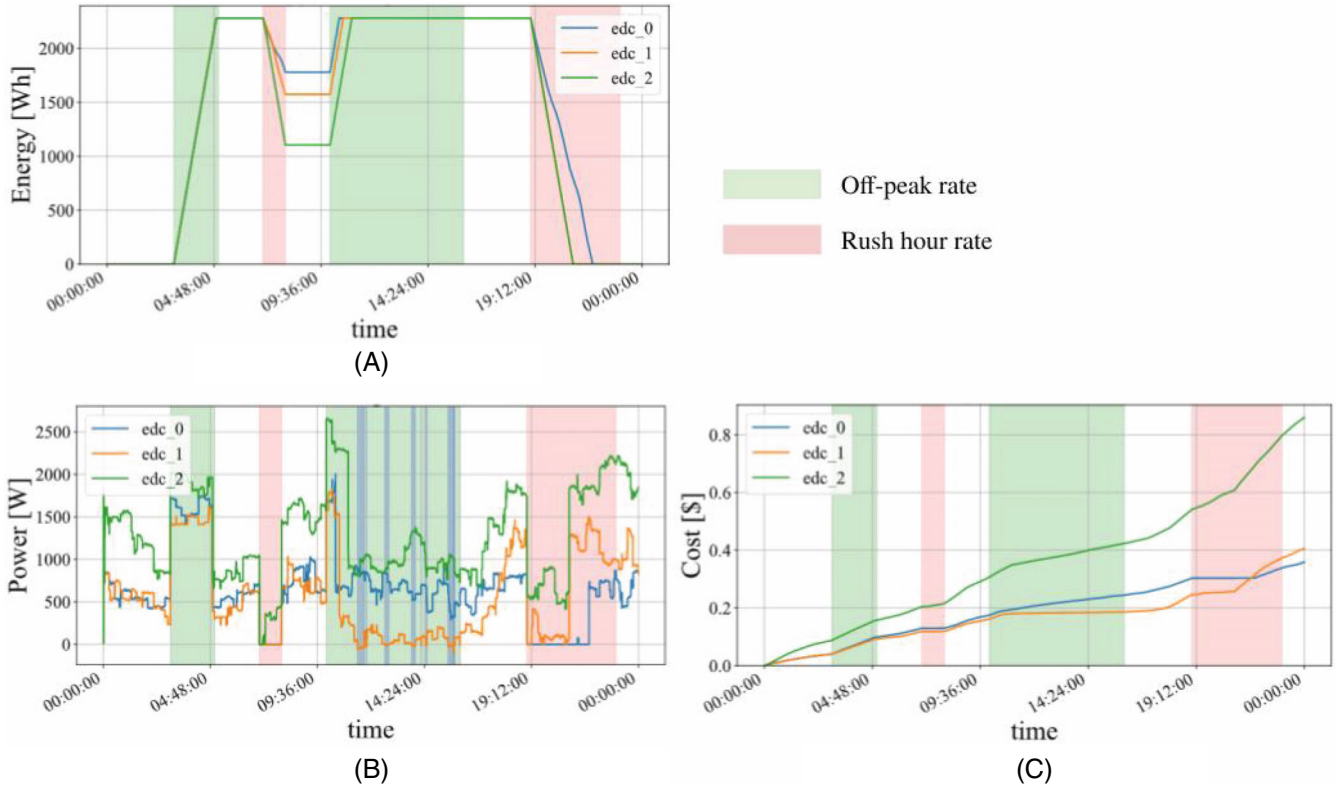

FIGURE 15 Battery energy storage, grid power consumption and energy cost. (A) Energy stored per EDC, (B) power consumption drawn from the power grid per EDC, (C) energy cost profile per EDC.

energy and battery power when the electricity price is high. This improvement significantly enhances the energy and economic sustainability of the edge infrastructure.

In this research, we optimized the energy consumption of the federation through workload allocation, considering two-phase immersion cooling, user mobility, smart grid with solar generation and batteries, and energy prices. The presented scenario opens the door to in-depth investigations on how federated edge computing management can self-organize to absorb demand peaks. Additionally, it can be exploited to delve into the management of secondary markets, where workloads with diverse resource and QoS requirements, as well as demand profiles, coexist, as expected in personalized medicine. The objective of our approach is to harness the M&S&O framework to deploy and operate real edge federations that are aware of the heterogeneity of their resources and environment, thereby enhancing their energy and economic sustainability.

8 | CONCLUSIONS AND FUTURE DIRECTIONS

Edge computing has the potential to facilitate the digitization of society, driving the deployment of critical applications that enhance people's quality of life. However, in practice, implementing the edge paradigm proves to be more disruptive than anticipated, primarily due to the limitations of applying current cloud-based strategies at the network's edge. In line with sustainability commitments shaping industry regulations, edge computing is currently perceived as a threat due to the energy inefficiency of colocation-based solutions and the rising density of computing.

Our primary objective is to transform this perceived threat into an opportunity, fostering the sustainable development of future edge infrastructures and making them environmentally and economically competitive to expedite their adoption. In this article, we articulate our vision for tackling the fundamental challenges associated with energy-aware edge deployment and operation, including efficient cooling, fault-tolerant automation, and collaborative orchestration. We put forward an innovative approach that integrates two-phase immersion cooling, formal modeling, simulation and optimization techniques, ML, and federated management to harness heterogeneity and propel the sustainability of edge computing.

We present a high-level architectural framework encompassing the essential infrastructure for the sustainable deployment of critical IoT services within an edge federation. The framework includes IoT equipment, cooling systems, network connectivity, smart grid, and cloud resources. Our architecture features a Model-Based System Engineering (MBSE) framework designed for fault-tolerant and heterogeneity-aware deployment and operation. Our integrated approach is presented, demonstrating how it can aid in exploring the global impact of energy-related dimensions in edge-based solutions.

To showcase the advantages of our methodology, we outline our early steps toward establishing the sustainability of an edge infrastructure for an ADAS application. We conceived a prototype EDC with an state-of-the-art PUE using two-phase immersion cooling, investigating its thermal and power dynamics. Subsequently, we formulated a model for our DC prototype to simulate and optimize its federated operation. Lastly, we evaluate the performance of our prototype by simulating an edge federation that incorporates a smart grid model to leverage environmental heterogeneity. Our findings demonstrate that our approach can reduce edge federation consumption by 20% and operating costs by 30%, thereby significantly enhancing the energy and economic sustainability of the edge infrastructure. These preliminary findings underscore the potential advantages of our vision for a more sustainable future in edge computing. In pursuit of this goal, our future work will focus on the following research areas:

- Investigate the thresholds of two-phase immersion cooling for accommodating higher IT power densities approaching the coolant's critical heat flux.
- Evaluate innovative methods for fluid condensation to develop two-phase immersion systems that incorporate free-cooling mechanisms.
- Examine secure workload allocation and consolidation strategies that account for the heat exchange capacity limitations of two-phase immersion coolants.
- Enhance the fault-tolerant operation of edge federations to bolster security against anomalies and vulnerabilities.
- Devise sustainable approaches for dimensioning and deploying edge federations tailored to the requirements and dynamic resource demands of emerging critical applications.
- Introduce novel energy optimization techniques for fostering collaboration among geographically distributed edge federations and clouds, capitalizing on their inherent heterogeneity.
- Formulate collaborative strategies for competitive multi-tenant environments, particularly for critical applications.

AUTHOR CONTRIBUTIONS

Patricia Arroba: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing - original draft, visualization, supervision. **Rajkumar Buyya:** Conceptualization, writing - review & editing, supervision. **Román Cárdenas:** Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing - original draft, visualization. **José L. Risco-Martín:** Conceptualization, methodology, formal analysis, writing - review & editing. **José M. Moya:** Conceptualization, writing - review & editing, supervision, resources, funding acquisition, project administration.

ACKNOWLEDGMENTS

We would like to thank 3M, Adam Data Centers, ImesAPI, and Tychetools for their support of this research. This work is partially supported by the HiPEAC6 Network, financed by the European Union's Horizon2020 research and innovation programme, and by the University of Melbourne. This research has also been supported by the Centre for the Development of Industrial Technology (CDTI) and State R&D Program Oriented to the Challenges of the Society (Retos Colaboración 2017) under contracts IDI-20171194 and RTC-2017-6090-3 from the Spanish Ministry of Science and Innovation, and by the Spanish State Research Agency under grant PID2019-110866RB-I00/AEI/10.13039/501100011033.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Rajkumar Buyya https://orcid.org/0000-0001-9754-6496
José L. Risco-Martín https://orcid.org/0000-0002-3127-6507

REFERENCES

1. Cisco. *Establishing the Edge. A New Infrastructure Model for Service Providers*. White paper. tech. rep. Cisco; 2022.
2. Bittman T, Iams T, Unni S, Goodness E, Gill B. *Predicts 2024: Edge Computing Technologies Are Gaining Traction and Maturity*. tech. rep. Gartner Research; 2023.
3. Lin L, Liao X, Jin H, Li P. Computation offloading toward edge computing. *Proc IEEE*. 2019;107(8):1584-1607.
4. Bala R, Gill B, Smith D, Wright D, Ji K. Magic quadrant for cloud infrastructure and platform services. *Gartner Reprints*. https://www.gartner.com/doc/reprints. 2020.
5. Mytton D. Hiding greenhouse gas emissions in the cloud. *Nat Clim Chang*. 2020;10(8):701.
6. Davis J, Bizo D, Lawrence A, et al. *Uptime Institute Global Data Center Survey 2022*. tech. rep. Uptime Institute; 2022.
7. Dodge J, Prewitt T, Combes T, et al. Measuring the carbon intensity of AI in cloud instances. *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM; 2022:1877-1894.
8. Vos S, Lago P, Verdecchia R, Heitlager I. Architectural tactics to optimize software for energy efficiency in the public cloud. *2022 International Conference on ICT for Sustainability (ICT4S)*. IEEE; 2022:77-87.
9. Li H, Dai H, Liu Z, Fu H, Zou Y. Dynamic energy-efficient scheduling for streaming applications in storm. *Computing*. 2022;104(2):413-432.
10. Synergy Research Group. *Colocation Market Benefitting Greatly from Hyperscale Operator Clients*. tech. rep. Synergy Research Group; 2020.
11. Bittman T, Gill B, Zimmerman T, Friedman T, MacDonald N, Brown K. *Gartner Predicts 2022: The Distributed Enterprise Drives Computing to the Edge*. tech. rep. Gartner Research; 2021.
12. Singh R, Sukapuram R, Chakraborty S. A survey of mobility-aware multi-access edge computing: Challenges, use cases and future directions. *Ad Hoc Netw*. 2023;140:103044. doi:10.1016/j.adhoc.2022.103044
13. Pérez S, Arroba P, Moya JM. Energy-conscious optimization of Edge Computing through Deep Reinforcement Learning and two-phase immersion cooling. *Futur Gener Comput Syst*. 2021;125:891-907.
14. Haghshenas K, Setz B, Blosch Y, Aiello M. Enough hot air: the role of immersion cooling. *Energy Inform*. 2023;6(1):14.
15. Capes J, Tuma P. *Liquid Cooling: The Key to Data Center Sustainability*. tech. rep. LiquidStack; 2022.
16. Yu L, Yang J, Shia D, Zhang M. A study on compatibility of thermal interface materials with coolants for data center immersion cooling. *2022 21st IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm)*. IEEE; 2022:1-5.
17. Jalili M, Manousakis I, Goiri I, et al. Cost-efficient overclocking in immersion-cooled datacenters. *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE; 2021:623-636.
18. Raniwala A. Bringing 2-phase immersion cooling to hyperscale cloud. *2022 Optical Fiber Communications Conference and Exhibition (OFC)*. IEEE; 2022:1-3.
19. Chen P, Harmand S, Ouenzerfi S. Immersion cooling effect of dielectric liquid and self-rewetting fluid on smooth and porous surface. *Appl Therm Eng*. 2020;180:115862. doi:10.1016/j.applthermaleng.2020.115862
20. Ledin JA. Hardware-in-the-loop simulation. *Embed Syst Program*. 1999;12:42-62.
21. Wang C, Gill C, Lu C. FRAME: fault tolerant and real-time messaging for edge computing. *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE; 2019:976-985.
22. Tuli S, Casale G, Jennings NR. PreGAN: preemptive migration prediction network for proactive fault-tolerant edge computing. *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*. IEEE; 2022:670-679.
23. Mudassar M, Zhai Y, Lejian L. Adaptive fault-tolerant strategy for latency-aware IoT application executing in edge computing environment. *IEEE Internet Things J*. 2022;9(15):13250-13262. doi:10.1109/JIOT.2022.3144026
24. Feng C, Wang Y, Chen Q, Ding Y, Strbac G, Kang C. Smart grid encounters edge computing: opportunities and applications. *Adv Appl Energy*. 2021;1:100006. doi:10.1016/j.adapen.2020.100006
25. Hua H, Li Y, Wang T, Dong N, Li W, Cao J. Edge computing with artificial intelligence: a machine learning perspective. *ACM Comput Surv*. 2023;55(9):1-35. doi:10.1145/3555802
26. Perez J, Arroba P, Moya JM. Data augmentation through multivariate scenario forecasting in data centers using generative adversarial networks. *Appl Intell*. 2023;53(2):1469-1486.
27. Cárdenas R, Arroba P, Risco-Martín JL, Moya JM. Modeling and simulation of smart grid-aware edge computing federations. *Clust Comput*. 2022;26:719-743.
28. Buyya R, Ilager S, Arroba P. Energy-efficiency and sustainability in new generation cloud computing: a vision and directions for integrated management of data centre resources and workloads. *Softw Pract Experience*. 2024;54(1):24-38. doi:10.1002/spe.3248
29. Khan T, Tian W, Ilager S, Buyya R. Workload forecasting and energy state estimation in cloud data centres: ML-centric approach. *Futur Gener Comput Syst*. 2022;128:320-332.
30. Yadav R, Zhang W, Li K, Liu C, Laghari AA. Managing overloaded hosts for energy-efficiency in cloud data centers. *Clust Comput*. 2021;24:1-15.
31. Kar B, Yahya W, Lin YD, Ali A. Offloading using traditional optimization and machine learning in federated cloud–edge–fog systems: a survey. *IEEE Commun Surv Tutor*. 2023;25(2):1199-1226. doi:10.1109/COMST.2023.3239579
32. Iftikhar S, Gill SS, Song C, et al. AI-based fog and edge computing: a systematic review, taxonomy and future directions. *Internet Things*. 2023;21:100674. doi:10.1016/j.iot.2022.100674
33. Chakraborty A, Kumar M, Chaurasia N, Gill SS. Journey from cloud of things to fog of things: Survey, new trends, and research directions. *Softw Pract Experience*. 2023;53(2):496-551. doi:10.1002/spe.3157
34. Gill SS. Quantum and blockchain based Serverless edge computing: a vision, model, new trends and future directions. *Internet Technol Lett*. 2024;7(1):e275. doi:10.1002/itl2.275

35. Yang Z, Pan C, Wang K, Shikh-Bahaei M. Energy efficient resource allocation in UAV-enabled mobile edge computing networks. *IEEE Trans Wirel Commun*. 2019;18(9):4576-4589.
36. Kiran N, Pan C, Wang S, Yin C. Joint resource allocation and computation offloading in mobile edge computing for SDN based wireless networks. *J Commun Netw*. 2019;22(1):1-11.
37. Hossain MS, Muhammad G, Amin SU. Improving consumer satisfaction in smart cities using edge computing and caching: a case study of date fruits classification. *Futur Gener Comput Syst*. 2018;88:333-341.
38. Ali Z, Jiao L, Baker T, Abbas G, Abbas ZH, Khaf S. A deep learning approach for energy efficient computational offloading in mobile edge computing. *IEEE Access*. 2019;7:149623-149633.
39. Hewage TB, Ilager S, Rodriguez MA, Arroba P, Buyya R. DEMOTS: a decentralized task scheduling algorithm for micro-clouds with dynamic power-budgets. *2023 IEEE 16th International Conference on Cloud Computing (CLOUD)*. IEEE; 2023:418-427.
40. Sharma M, Tomar A, Hazra A. Edge computing for industry 5.0: fundamental, applications and research challenges. *IEEE Internet Things J*. 2024;1. doi:10.1109/JIOT.2024.3359297
41. Das R, Inuwa MM. A review on fog computing: Issues, characteristics, challenges, and potential applications. *Telematics Informat Reports*. 2023;10:100049. doi:10.1016/j.teler.2023.100049
42. NVIDIA. *Self-driving safety report*. tech. rep. NVIDIA Corp; 2018.
43. Chekired DA, Togou MA, Khoukhi L, Ksentini A. 5G-slicing-enabled scalable SDN core network: toward an ultra-low latency of autonomous driving service. *IEEE J Sel Areas Commun*. 2019;37(8):1769-1782.
44. Mittal S, Tolk A. *Complexity Challenges in Cyber Physical Systems: Using Modeling and Simulation (M&S) to Support Intelligence*. Wiley; 2019.
45. Ananthanarayanan G, Bahl P, Bodík P, et al. Real-time video analytics: The killer app for edge computing. *Computer*. 2017;50(10):58-67.
46. Pérez S, Pérez J, Arroba P, Blanco R, Ayala JL, Moya JM. Predictive GPU-based ADAS management in energy-conscious smart cities. *2019 IEEE International Smart Cities Conference (ISC2)*. IEEE; 2019:349-354.
47. Cárdenas R, Arroba P, Blanco R, Malagón P, Risco-Martín JL, Moya JM. Mercury: a modeling, simulation, and optimization framework for data stream-oriented IoT applications. *Simul Model Pract Theory*. 2020;101:102037.
48. Zeigler BP, Moon Y, Kim D, Ball G. The DEVS environment for high-performance modeling and simulation. *IEEE Comput Sci Eng*. 1997;4(3):61-71.
49. Cárdenas R, Arroba P, Risco Martin JL. Bringing AI to the edge: a formal M&S specification to deploy effective IoT architectures. *J Simulat*. 2022;16(5):494-511.
50. Risco-Martín JL, Mittal S, Henares K, Cardenas R, Arroba P. xDEVS: a toolkit for interoperable modeling and simulation of formal discrete event systems. *Softw Pract Experience*. 2023;53(3):748-789.
51. Piorkowski M, Sarafijanovoc-Djukic N, Grossglauser M. A parsimonious model of mobile partitioned networks with clustering. *The First International Conference on COMmunication Systems and NETworkS (COMSNETS)*. IEEE; 2009.

How to cite this article: Arroba P, Buyya R, Cárdenas R, Risco-Martín JL, Moya JM. Sustainable edge computing: Challenges and future directions. *Softw: Pract Exper*. 2024;54(11):2272-2296. doi: 10.1002/spe.3340