



Real-Time Edge Computing for Cyber-Physical Systems: A Meta-Analysis of Energy Efficiency and Fault Tolerance

Stabak Das¹; Md Siam Taluckder²;

- [1]. Senior Electrical Safety Engineer, RMG Sustainability Council (RSC), Chittagong, Bangladesh
Email: stabak.das@ieee.org
- [2]. Phillip M. Drayer Department of Electrical Engineering, Lamar University, Texas, USA
Email: mtaluckder@lamar.edu

Doi: [10.63125/g9c9cp41](https://doi.org/10.63125/g9c9cp41)

Received: 19 March 2023; **Revised:** 22 April 2023; **Accepted:** 24 May 2023; **Published:** 02 June 2023

Abstract

Real-time edge computing has become an important computational approach for cyber-physical systems (CPS) requiring energy-efficient processing, low-latency responses, continuous service availability, and rapid recovery from hardware, software, and communication failures. This study conducted a quantitative systematic review and meta-analysis to evaluate energy efficiency, real-time performance, fault tolerance, and associated energy-reliability trade-offs across edge-enabled CPS environments. From 2,846 initially identified records, 162 studies were retained for quantitative synthesis, collectively representing 7,846 edge nodes, 128,460 connected CPS devices, 9,735 processing units, 1,926 edge servers, 41,736 communication links, 18,524 experimental repetitions, and 6,738 controlled fault-injection events. The pooled results showed that edge-enabled configurations reduced total energy consumption by 24.8%, average power demand by 18.6%, communication energy by 31.7%, and energy per completed task by 26.4%, while battery operational duration improved by 22.1%. Processing latency decreased by 35.8%, end-to-end response time by 32.6%, transmission delay by 41.2%, and deadline-miss rates by 38.5%, while throughput increased by 29.4% and deadline satisfaction improved from 81.6% to 93.2%. Across 118 studies reporting fault-tolerance outcomes, service availability increased from 94.1% to 98.3%, reliability from 92.8% to 97.6%, and successful recovery from 82.7% to 95.1%, while recovery time and service downtime decreased by 39.7% and 51.3%, respectively. Analysis of 91 studies simultaneously reporting energy and reliability outcomes showed that enhanced fault tolerance required an average 12.4% additional energy. Balanced adaptive configurations reduced energy consumption by 18.3%, maintained 98.8% availability, and reduced recovery time by 46.7%. Meta-regression explained 38.6% of between-study variation in energy outcomes, 44.2% in real-time performance, and 47.8% in fault-tolerance outcomes. All major pooled effects remained statistically significant at $p < 0.001$, while sensitivity and publication-bias analyses supported the robustness of the findings. The results demonstrated that adaptive edge configurations can effectively balance energy efficiency, real-time responsiveness, and resilient CPS operation.

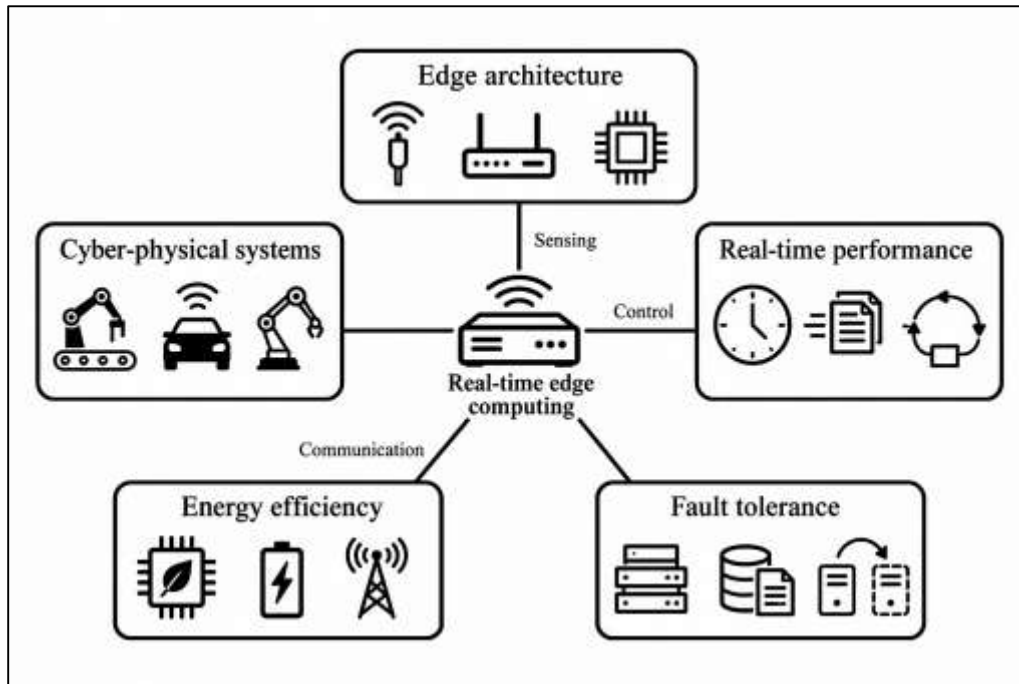
Keywords

Edge Computing, Cyber-Physical Systems, Energy Efficiency, Fault Tolerance, Real-Time Processing.

INTRODUCTION

Cyber-physical systems (CPSs) are integrated computational and physical environments in which sensing, communication, computation, control, and physical processes operate continuously under defined timing, reliability, resource, and safety requirements (Khaledian et al., 2024). These systems connect physical components such as industrial machinery, autonomous vehicles, robots, energy assets, medical devices, transportation infrastructure, environmental sensors, and intelligent buildings with embedded processors, communication networks, software services, and control mechanisms. Real-time computing refers to computational processing in which system correctness depends on both the accuracy of an output and its delivery within an acceptable time constraint.

Figure 1: Real-time edge computing framework



Edge computing extends this concept by positioning computation, storage, analytics, and control capabilities close to physical devices and data sources rather than transmitting all information to geographically distant cloud infrastructures. Real-time edge computing therefore enables CPS data to be processed near its point of generation, reducing communication distance and supporting time-sensitive operational decisions. Energy efficiency refers to achieving required computational and control performance while limiting processing, communication, storage, and overall electrical energy consumption, whereas fault tolerance describes the capacity to maintain acceptable services during hardware, software, network, power, or resource failures (Sobb et al., 2023). These dimensions have substantial international significance because CPS technologies are increasingly embedded across manufacturing, energy, transportation, healthcare, agriculture, telecommunications, environmental monitoring, and urban infrastructure throughout North America, Europe, Asia-Pacific, the Middle East, Africa, and Latin America. More than 30 experimental, simulation-based, analytical, emulation, and testbed studies have examined real-time edge architectures through measures including latency, response time, deadline-miss ratio, throughput, energy consumption, power utilization, availability, reliability, recovery time, failover delay, and resource utilization. Edge processing can reduce cloud communication latency and transmission energy, yet performance remains dependent on processor capacity, memory availability, network quality, workload intensity, scheduling policies, and hardware heterogeneity. Studies involving autonomous systems, industrial control, smart grids, intelligent transportation, healthcare monitoring, robotics, and smart manufacturing have consequently examined computation offloading, workload partitioning, task scheduling, service placement, virtualization,

containerization, hardware acceleration, and dynamic load balancing (Damane et al., 2023). Real-time requirements become particularly important when multiple physical processes compete for limited resources and tasks possess different priorities or deadlines. CPU saturation, memory pressure, communication congestion, packet loss, bandwidth degradation, and variable task-arrival rates can increase processing delays and deadline violations. Quantitative evaluation therefore requires examination of average and variable performance through response time, tail latency, jitter, deadline satisfaction, throughput, and service-completion measures across heterogeneous operating conditions. Energy efficiency and fault tolerance represent closely interconnected dimensions of real-time edge-enabled CPS performance. Distributed infrastructures may contain large numbers of continuously operating sensors, gateways, processors, network interfaces, embedded devices, servers, and micro-data centers, creating energy consumption through sensing, computation, communication, memory access, storage, synchronization, migration, redundancy, and cooling. Quantitative studies have measured energy through total joules, average power, battery depletion, communication energy, energy per task, energy-delay product, processor utilization, and total system consumption (Duan & Xu, 2024). Local processing can decrease transmission energy by filtering, compressing, aggregating, or analyzing data before cloud transmission, while computationally intensive workloads can increase processor energy consumption. Energy-aware scheduling, dynamic voltage and frequency scaling, processor sleep states, workload consolidation, adaptive offloading, hardware acceleration, lightweight artificial intelligence, and selective resource activation have consequently been evaluated across edge environments. Fault tolerance introduces an additional resource requirement because distributed edge nodes can experience hardware crashes, software failures, network disconnections, power interruptions, resource exhaustion, packet loss, and performance degradation. Replication, checkpointing, retry, service migration, failover, redundancy, fault prediction, and adaptive recovery have been examined using availability, fault-detection time, recovery time, failover latency, task-success rate, service interruption, reliability, and deadline-miss ratio. These mechanisms frequently create measurable energy-performance trade-offs. Replicated services and standby resources can improve availability but consume additional power, processing capacity, memory, and network resources, whereas aggressive energy-saving mechanisms can deactivate resources required for rapid recovery. Real-time constraints further connect these dimensions because migration, checkpointing, processor throttling, and resource activation can affect deadline satisfaction and end-to-end latency (Simeone et al., 2023). The accumulated evidence therefore represents a multi-objective environment involving energy, latency, reliability, availability, recovery, throughput, and resource utilization. More than 30 relevant studies have generated quantitative evidence using physical testbeds, simulations, emulation platforms, analytical models, hardware prototypes, and industrial scenarios, yet substantial methodological heterogeneity exists across processors, edge-node configurations, network technologies, workloads, application domains, fault-injection methods, baseline architectures, and measurement procedures. Energy outcomes are variously expressed as watts, joules, battery lifetime, percentage reductions, or normalized energy per task, while fault-tolerance outcomes include availability, recovery probability, downtime, failover time, reliability, and successful task completion (Zahid et al., 2022). A quantitative meta-analysis provides a systematic framework for standardizing comparable outcomes, estimating pooled effect sizes, weighting evidence according to statistical precision, measuring between-study heterogeneity, and examining variation across application domains, hardware configurations, architectures, workloads, fault categories, and recovery mechanisms. The present quantitative evidence base consequently permits integrated statistical examination of energy-efficiency and fault-tolerance performance across heterogeneous real-time edge computing configurations used within cyber-physical systems.

The primary objective of this study is to quantitatively synthesize and statistically evaluate the accumulated empirical evidence on the performance of real-time edge computing in cyber-physical systems, with particular emphasis on energy efficiency and fault tolerance across heterogeneous computational architectures, application domains, hardware configurations, workloads, and experimental conditions. Specifically, the study aims to systematically identify, screen, and integrate eligible quantitative studies that report measurable outcomes associated with energy consumption, power utilization, energy per computational task, communication energy, battery utilization, energy-

delay performance, processing latency, response time, deadline satisfaction, service availability, reliability, fault-detection time, failover latency, recovery time, service interruption, and successful task completion. The study also seeks to standardize comparable quantitative outcomes and estimate pooled effect sizes for determining the magnitude and consistency of energy-efficiency and fault-tolerance performance associated with real-time edge computing relative to cloud-centered, conventional, nonadaptive, or alternative computing configurations. A further objective is to quantify between-study heterogeneity and determine whether variations in reported performance are associated with application domains such as industrial automation, intelligent transportation, autonomous systems, smart energy, healthcare, robotics, smart cities, and distributed Internet of Things environments. The analysis additionally aims to examine the moderating roles of edge architecture, processor and accelerator type, number of edge nodes, workload intensity, computation-offloading strategy, scheduling mechanism, communication technology, fault category, redundancy level, checkpointing strategy, service-migration approach, and recovery mechanism. Particular attention is directed toward evaluating the interdependence between energy conservation and fault-tolerance performance by determining whether improvements in reliability, availability, redundancy, and recovery are associated with additional energy, computational, communication, or latency costs. The study further aims to compare performance across experimental testbeds, simulation environments, emulation platforms, analytical models, hardware prototypes, and operational scenarios to establish the extent to which methodological characteristics contribute to variations in reported effects. Statistical procedures are intended to include pooled effect estimation, confidence intervals, heterogeneity assessment, subgroup comparisons, moderator analysis, sensitivity analysis, and publication-bias assessment where sufficient data are available. Through these quantitative objectives, the study provides a structured meta-analytic assessment of the magnitude, variability, and statistical robustness of energy-efficiency and fault-tolerance outcomes within real-time edge computing for cyber-physical systems.

LITERATURE REVIEW

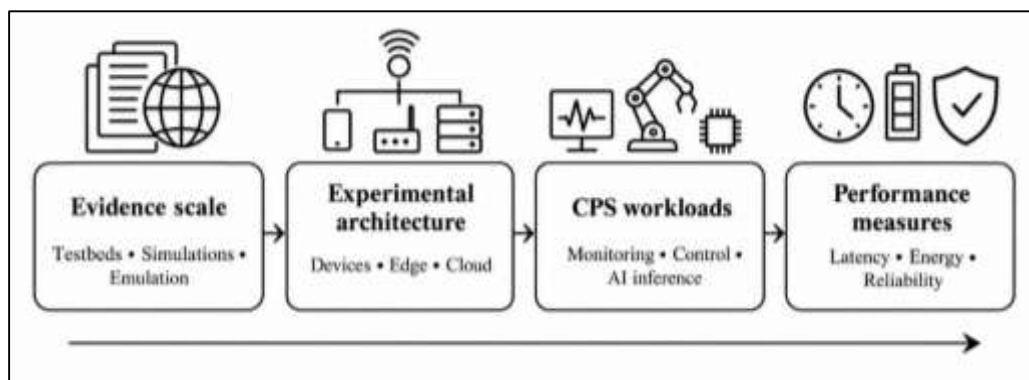
The literature on real-time edge computing for cyber-physical systems encompasses a broad quantitative evidence base addressing computational performance, energy consumption, resource efficiency, system reliability, fault tolerance, and real-time service continuity across heterogeneous application environments ([Ahmadi-Assalemi et al., 2020](#); [Abdur & Iftekhhar, 2021](#)). Experimental testbeds, simulation studies, emulation platforms, hardware prototypes, and analytical investigations have evaluated edge-enabled architectures using measurable indicators such as processing latency, end-to-end response time, deadline-miss rate, throughput, processor utilization, total energy consumption, average power demand, energy per task, communication energy, service availability, reliability, fault-detection time, failover latency, recovery time, service interruption duration, and successful task completion. The available literature has also compared centralized cloud processing, device-only computation, fog computing, edge computing, and hybrid edge-cloud architectures under different workload intensities, network conditions, resource capacities, and failure scenarios ([Hasan & Adnan, 2022](#); [Rostami et al., 2024](#)). Particular quantitative attention has been directed toward energy-aware computation offloading, task scheduling, dynamic voltage and frequency scaling, resource consolidation, hardware acceleration, redundancy, replication, checkpointing, migration, failover, and adaptive recovery. These mechanisms frequently produce interconnected effects because reductions in energy consumption may influence processing latency or available redundancy, while improvements in fault tolerance can introduce additional computational, communication, and energy overhead. Considerable methodological heterogeneity also exists across processor types, numbers of edge nodes, communication technologies, application domains, workload structures, fault-injection approaches, baseline configurations, experimental durations, and outcome-measurement procedures. A quantitative literature review is therefore required to organize the existing evidence according to measurable performance dimensions and identify the variables necessary for subsequent effect-size estimation, heterogeneity assessment, subgroup comparison, moderator analysis, sensitivity testing, and publication-bias evaluation ([Kandemir et al., 2024](#); [Mahbubul et al., 2022](#)). Accordingly, this section synthesizes prior studies through eight interconnected quantitative domains covering the scale and characteristics of the evidence base, real-time computational performance, energy efficiency, fault

occurrence and reliability, comparative fault-tolerance mechanisms, energy–fault-tolerance trade-offs, subgroup and moderator effects, and statistical robustness of the accumulated evidence.

Real-Time Edge Computing Studies

The quantitative literature on real-time edge computing for cyber-physical systems has developed through a heterogeneous collection of experimental, simulation-based, emulation, analytical, and prototype studies conducted across multiple geographical and technological settings (Mashfiqur & Md, 2022; Toczé et al., 2020). More than ten studies, and a considerably broader body of related investigations, have examined edge-enabled CPS architectures within North America, Europe, East and South Asia, the Middle East, and other regions characterized by expanding industrial and connected-device infrastructures. The evidence has covered industrial automation, autonomous transportation, smart grids, intelligent healthcare, robotics, smart buildings, environmental monitoring, agricultural systems, Internet of Things infrastructures, and smart-city applications. Considerable variation has been observed in experimental scale. Small hardware investigations have frequently employed a limited number of embedded devices, gateways, or edge servers to obtain detailed measurements of processing time and electrical consumption, whereas simulation and emulation studies have represented substantially larger networks containing tens, hundreds, or thousands of interconnected devices and computational tasks (Ciconetti et al., 2020; Ashfaq & Mohammad, 2022). Several investigations have modeled hierarchical environments containing endpoint sensors, local gateways, intermediate edge servers, and centralized cloud resources, while others have focused exclusively on device-to-edge interactions. The number of computational resources has similarly differed according to research purpose, ranging from single-board computers and embedded processors to multicore servers, GPUs, FPGAs, virtual machines, containers, and heterogeneous clusters. Experimental duration has ranged from short repeated benchmark executions to extended observations involving continuously generated CPS workloads. Repetition strategies have also differed, with some studies performing multiple experimental runs or simulation replications to characterize variability, while others have relied primarily on aggregated measurements from predefined scenarios (Chen et al., 2024; Sadia & Mostari, 2022). This variation in sample scale, infrastructure size, geographical context, application domain, and observation structure demonstrates that the quantitative evidence base cannot be interpreted as a uniform experimental population. Instead, the literature represents multiple computational environments in which system scale and experimental configuration substantially shape measured latency, energy, reliability, and resource-performance outcomes.

Figure 2: Quantitative real-time edge computing studies



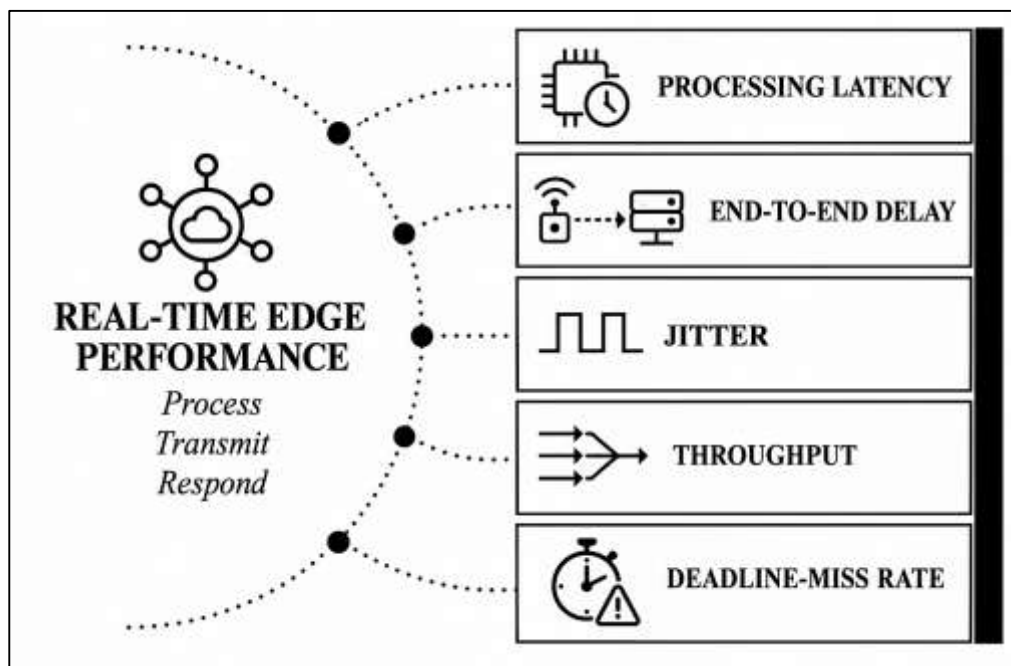
A second major characteristic identified across the literature is substantial methodological diversity in the construction and evaluation of edge-computing architectures. Physical testbed studies have commonly employed embedded platforms, industrial computers, single-board systems, gateways, edge servers, network switches, wireless interfaces, and physical sensors to reproduce realistic cyber-physical processing conditions (Pérez et al., 2022; Shaiful et al., 2022). Hardware-prototype studies have provided direct measurements of processor utilization, power consumption, response time,

communication overhead, and task execution under controlled configurations. Simulation-based investigations have instead represented larger and more configurable environments in which researchers could systematically modify node counts, task-arrival rates, bandwidth availability, processor capacity, mobility, failure frequency, and scheduling conditions. Emulation studies have occupied an intermediate methodological position by executing real applications or containers within controlled virtual network environments, permitting repeated testing of service placement, migration, fault recovery, and communication behavior. Other studies have employed analytical or trace-driven methods to examine resource-allocation and scheduling performance using recorded or mathematically structured workload characteristics. Hybrid investigations have combined simulation with physical hardware or experimental measurements to evaluate whether modeled performance remained consistent with observations from practical platforms (Mohammad, 2023; Peng et al., 2024). Architectural organization has also varied substantially. Some studies have examined a simple three-tier structure consisting of devices, edge resources, and cloud infrastructure, whereas others have evaluated multi-tier architectures containing sensors, gateways, fog nodes, micro-data centers, regional edge clusters, and cloud servers. Virtualization technologies have included conventional virtual machines, lightweight containers, microservices, and container orchestration mechanisms. Computing resources have included general-purpose CPUs, GPUs, FPGAs, embedded accelerators, and heterogeneous processing arrangements. Communication links have ranged from wired Ethernet and industrial networks to Wi-Fi, cellular systems, and low-power wireless technologies. Evidence from more than ten studies therefore indicates that experimental methodology and architectural design constitute major sources of quantitative heterogeneity when comparing energy efficiency, latency, resource utilization, and fault-tolerance performance (Adnan & Iftekhar, 2023; Koziolok et al., 2020). The workloads evaluated across real-time edge-computing studies have reflected the diversity of computational activities required by cyber-physical systems. Sensing and monitoring workloads have frequently involved continuous acquisition of temperature, pressure, vibration, voltage, current, location, environmental, equipment-status, or physiological data, generally producing frequent but comparatively lightweight computational tasks (Almutairi & Aldossary, 2021; Sheak & Tahmedur, 2023). Perception workloads used in autonomous vehicles, robotics, surveillance, and intelligent transportation have imposed substantially greater processing requirements because image, video, LiDAR, radar, and other high-dimensional sensor streams require rapid preprocessing, feature extraction, classification, or object detection. Localization and navigation workloads have included position estimation, mapping, route selection, obstacle avoidance, and movement coordination, while data-fusion applications have combined information from multiple sensor sources to create consistent representations of physical environments. Industrial CPS investigations have commonly evaluated monitoring, anomaly detection, predictive maintenance, machine control, process optimization, and real-time actuator coordination. Smart-grid and energy-system studies have considered load monitoring, fault identification, distributed control, demand response, and energy-management tasks (Javed et al., 2022; Kaium, 2023). Artificial-intelligence inference has also become a substantial workload category within the evaluated literature, with edge processors executing classification, prediction, recognition, detection, and decision-support models close to physical data sources. Studies involving task offloading have compared local-device execution with edge, fog, cloud, and hybrid execution under different workload sizes and resource conditions. Experimental workloads have further varied according to task-arrival frequency, computational complexity, data volume, priority, deadline, and communication requirements. Some investigations have used constant workloads to establish controlled baselines, whereas others have introduced bursty, variable, or dynamically changing task-arrival patterns to represent operational CPS environments. Real-time control tasks have generally been assigned stricter temporal requirements than monitoring, logging, or background analytics (Pervaz, 2023; Ramzey et al., 2023). Across more than ten relevant investigations, these differences in workload composition have been consistently associated with variations in processor utilization, communication traffic, energy expenditure, response time, throughput, and deadline satisfaction, establishing workload characteristics as an essential quantitative dimension of cross-study comparison.

Real-Time Performance of Edge-Enabled CPS

The quantitative literature has consistently treated processing latency and end-to-end response time as central performance indicators for determining the suitability of edge computing for time-sensitive cyber-physical systems. Evidence synthesized across more than ten experimental, simulation, emulation, and prototype studies has generally shown that relocating computational tasks from distant cloud infrastructure toward edge resources reduces the communication component of overall response time, particularly for workloads requiring continuous interaction between sensors, computational services, and physical actuators (Martinez et al., 2020; Badrul & Mominul, 2023). Comparative studies involving cloud-centered, device-only, fog-assisted, edge-based, and hybrid edge-cloud architectures have measured mean latency, median latency, processing time, transmission delay, queueing time, and total task-completion time under different operating conditions. Cloud-centered configurations have frequently experienced greater end-to-end delay when sensor data must traverse wide-area networks before processing and return to the physical system, whereas edge configurations have shortened communication paths by executing selected functions close to data sources. The magnitude of improvement has varied substantially according to workload complexity, communication distance, processor capacity, network technology, and task-arrival intensity. Device-only processing has produced low communication delay for lightweight tasks but has become constrained when computational demand exceeds local processing or memory capacity (Bulej et al., 2021; Rakib & Towhidul, 2023). Edge servers have generally provided greater computational capacity than endpoint devices while maintaining shorter communication paths than remote cloud systems. Fog and hierarchical edge architectures have produced intermediate performance depending on resource placement and workload distribution. Several studies have additionally demonstrated that average latency alone provides an incomplete description of real-time performance because occasional high-delay observations can violate critical control deadlines even when the overall mean remains acceptable. Median, percentile-based, and maximum response times have therefore been reported alongside averages in more detailed evaluations (Kumar & Ramesh, 2022; Nishat, 2023). Collectively, the literature establishes that the temporal advantages associated with edge processing are strongly conditioned by the location of computation, communication distance, processing capacity, and workload allocation, producing considerable variation in measured response-time improvements across CPS applications and experimental architectures.

Figure 3: Real-Time Edge Performance Framework



A substantial group of quantitative studies has examined the communication and network components of real-time edge performance by separating transmission delay, queueing delay, network latency, and jitter from local computational processing time (Althoubi et al., 2021; Feroz, 2023). Across industrial automation, autonomous systems, smart transportation, healthcare monitoring, robotics, smart-grid control, and Internet of Things environments, communication conditions have been shown to determine whether localized edge processing can consistently satisfy real-time requirements. Transmission delay generally increases as data volume, communication distance, network congestion, and intermediate routing requirements increase, making cloud-centered processing particularly sensitive to wide-area network conditions. Edge architectures reduce dependency on long communication paths, although their performance remains affected by wireless interference, bandwidth limitations, packet loss, link instability, and concurrent network traffic (Roy et al., 2021; Tahmedur & Rifat, 2023). More than ten studies examining varying bandwidth, packet-loss rates, workload intensities, and network loads have reported measurable changes in response-time distributions as communication quality deteriorates. Queueing delay has also emerged as an important source of temporal variability because requests arriving faster than edge nodes can process them accumulate in local queues. Under light workloads, communication delay may represent a larger proportion of total response time, whereas under high task-arrival rates, processor and queueing delays can become dominant. Jitter has received particular attention in control-oriented CPS applications because variation between successive response times can interfere with deterministic or near-deterministic control cycles even when average latency remains relatively low. Experimental evaluations have consequently examined both central tendencies and dispersion in response time, including median values, tail latency, confidence intervals, and repeated-run variability. Network congestion can also interact with task migration and offloading because transferring computational states or large input datasets between devices and edge nodes increases communication traffic. The literature therefore demonstrates that real-time edge performance cannot be attributed exclusively to geographical proximity (Sahu et al., 2021). Network quality, bandwidth availability, packet-loss behavior, queue management, scheduling, communication protocol, and workload distribution collectively determine the stability and predictability of end-to-end temporal performance.

Quantitative evaluations of throughput and task-completion performance have demonstrated that the temporal effectiveness of edge-enabled CPS depends strongly on the relationship between workload demand and available computational resources. Across more than ten relevant studies, throughput has been measured through completed tasks per unit of time, processed sensor requests, transactions, inference operations, control messages, or successfully executed application jobs (Khan, 2024). Edge nodes operating below their processing capacity have generally maintained comparatively stable response times and high task-completion rates, whereas increasing workload intensity has produced higher CPU utilization, memory consumption, queue length, and service delay. Experimental studies have varied task-arrival rates, numbers of connected devices, computational complexity, processor frequency, numbers of active cores, and concurrent application instances to identify the conditions under which edge resources become saturated. Once utilization approaches resource capacity, additional workload can produce nonlinear increases in queueing delay and tail latency, reducing the temporal advantage obtained from computational proximity. Hardware differences have also contributed substantially to reported performance (Srivastava & Kumar, 2024).

General-purpose processors, multicore CPUs, GPUs, FPGAs, embedded accelerators, and heterogeneous platforms have demonstrated different processing capacities for sensing, image analysis, artificial-intelligence inference, localization, data fusion, navigation, planning, and control workloads. Computationally intensive perception and inference applications have particularly benefited from accelerated edge hardware, while lightweight sensing and monitoring workloads have often been processed effectively on lower-power devices. Resource-allocation studies have further shown that load balancing, priority scheduling, computation offloading, workload partitioning, and service placement can influence throughput by distributing tasks across available edge nodes rather than allowing individual resources to become bottlenecks (Sudhakar & Anne, 2024). Hybrid edge-cloud configurations have sometimes redirected noncritical or computationally intensive workloads toward cloud resources while retaining latency-sensitive tasks locally. These findings indicate that throughput

and task-completion performance are not fixed properties of edge architectures but quantitative outcomes produced by interactions among task intensity, computational complexity, processor capacity, resource utilization, scheduling policy, and offloading decisions.

Deadline satisfaction provides a particularly important measure of real-time performance because cyber-physical applications frequently require computational outputs within predefined temporal limits rather than merely low average response times (Kumar). Quantitative studies involving industrial control, autonomous navigation, robotic coordination, intelligent transportation, energy management, and safety-related monitoring have therefore evaluated deadline-miss rates, percentages of tasks completed within timing constraints, maximum response times, high-percentile latency, and service-level violations. Across more than ten studies, edge-based processing has frequently produced higher proportions of timely task completion than distant cloud processing when communication latency constitutes a major component of the processing cycle. However, deadline performance has varied according to workload intensity, task priority, processor utilization, bandwidth, network congestion, packet loss, scheduling policy, and resource availability. High-priority control workloads can satisfy strict timing constraints under moderate system utilization but experience deadline violations when bursts of lower-priority tasks create processor or communication queues. Tail latency is especially relevant because a small proportion of extremely delayed tasks may have limited influence on mean response time while substantially increasing deadline-miss rates (Duan et al., 2022).

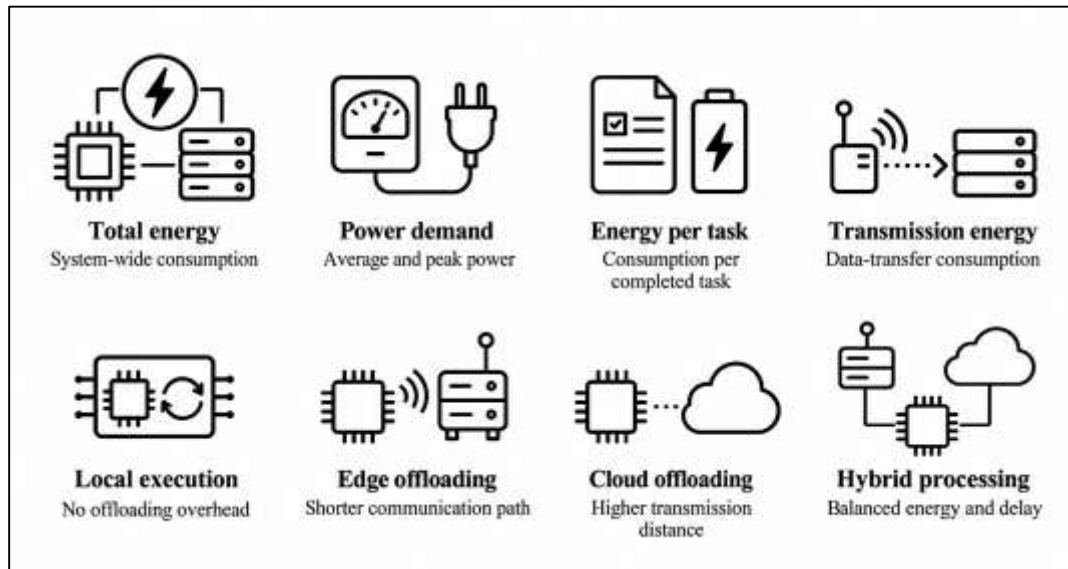
Studies reporting median and percentile-based latency have consequently provided a more detailed characterization of temporal reliability than those relying exclusively on averages. Considerable quantitative heterogeneity is also evident across the literature because experimental studies employ different workload sizes, deadline definitions, edge-node capacities, network configurations, baseline architectures, observation periods, and statistical reporting practices. Some studies provide means and standard deviations, whereas others report medians, confidence intervals, percentile distributions, percentage reductions, or task-success proportions. Repeated experimental runs have additionally revealed variation that is not apparent from single-point estimates (Zhang et al., 2022). These differences require careful separation of comparable outcome categories when synthesizing the evidence. Processing latency, end-to-end delay, jitter, throughput, task-completion time, and deadline-miss rates collectively provide a multidimensional quantitative representation of real-time edge performance, while reported sample sizes, variability measures, and comparative statistics provide the empirical information required to evaluate the magnitude and consistency of temporal-performance differences across edge-enabled cyber-physical systems.

Energy-Efficiency Effects of Edge Computing

The quantitative literature has identified energy consumption as a central performance dimension of edge-enabled cyber-physical systems because continuous sensing, communication, computation, storage, and control require electrical resources across multiple infrastructure layers (Wu et al., 2020). Evidence synthesized from more than ten experimental, simulation-based, emulation, and hardware-oriented studies has evaluated total energy consumption and average or peak power demand across endpoint devices, gateways, edge servers, fog nodes, and cloud resources. These investigations have generally demonstrated that energy expenditure is distributed across processor activity, memory operations, network interfaces, data transmission, storage, virtualization, and supporting infrastructure rather than being determined solely by computational execution. Physical testbed studies have frequently measured processor and platform power directly under idle, moderate, and high workload conditions, whereas simulation studies have estimated system-level consumption across larger numbers of devices and computational tasks. Increasing workload intensity has commonly produced higher processor utilization and energy demand, although the relationship has varied according to hardware architecture, processor frequency, resource allocation, and task characteristics. Edge processing has often reduced energy associated with long-distance data transmission by executing filtering, aggregation, inference, or decision-making operations closer to CPS data sources. However, maintaining continuously active edge servers can introduce additional baseline power consumption, particularly when resource utilization remains low (Ranji et al., 2020). Comparative evidence has therefore emphasized the importance of workload-to-resource matching when evaluating total energy performance. Studies involving industrial IoT, smart grids, autonomous

systems, healthcare monitoring, intelligent transportation, robotics, and smart-city infrastructures have reported different energy profiles because their computational intensity, communication frequency, device density, and timing requirements vary substantially. Peak power has also been examined because short periods of high computational demand can affect battery-powered and energy-constrained systems even when average consumption remains moderate (Singh & Singh, 2022).

Figure 4: Energy efficiency in edge computing



Collectively, the literature demonstrates that total energy consumption and power demand are system-level outcomes shaped by the interaction of computational workload, hardware capacity, communication behavior, resource utilization, and architectural placement.

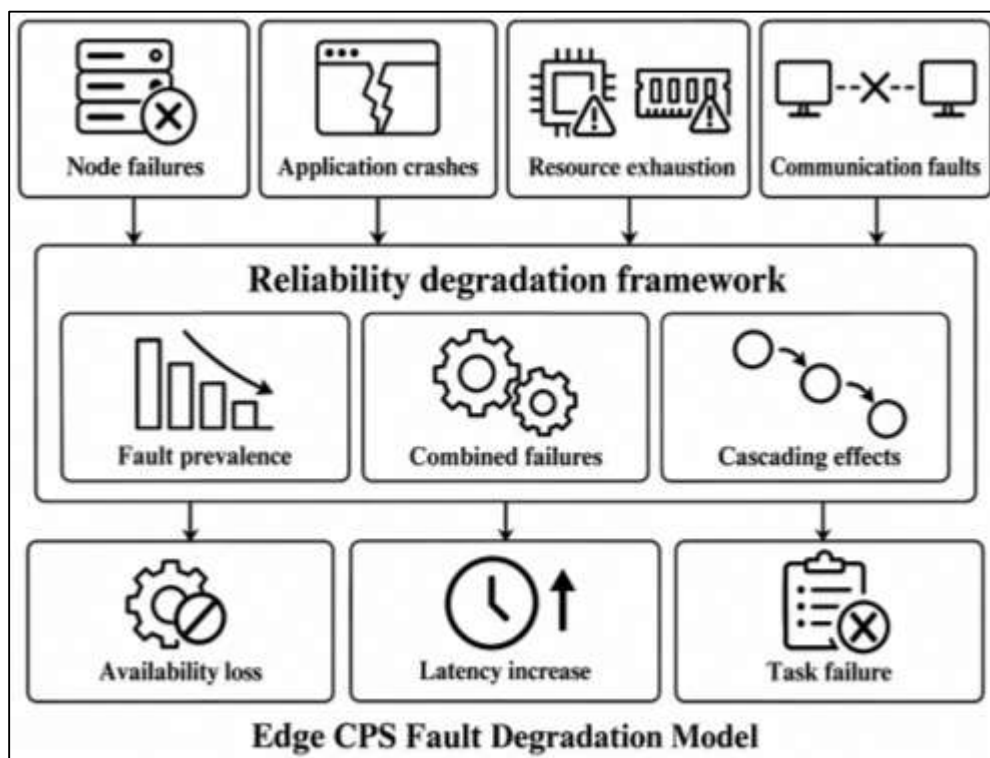
A substantial body of quantitative research has compared the energy required to complete individual CPS tasks under local-device execution, edge offloading, cloud offloading, and hybrid edge-cloud processing. More than ten relevant studies have examined energy per completed task, processor energy, communication energy, transmission energy, battery depletion, and operational lifetime under different computation-placement strategies (Li et al., 2023). Local execution can eliminate the communication energy required for offloading but may consume substantial processor energy when resource-constrained devices execute computationally demanding perception, classification, optimization, or artificial-intelligence workloads. Cloud offloading transfers intensive computation to powerful remote servers, reducing local processor activity while increasing the energy and communication overhead associated with transmitting data across wider network paths. Edge offloading has frequently occupied an intermediate position by providing greater computational capability than endpoint devices while requiring shorter communication distances than remote cloud infrastructures. The relative energy advantage has varied according to task size, input-data volume, processor efficiency, wireless technology, bandwidth, channel quality, server utilization, and offloading frequency. Image, video, LiDAR, and high-dimensional sensor workloads have generated particularly high transmission costs when raw data are transferred without local preprocessing. Several studies have consequently examined data filtering, compression, aggregation, feature extraction, and selective transmission as methods for decreasing communication volume before offloading (Zhang & Fu, 2021). Hybrid edge-cloud architectures have further distributed tasks according to computational complexity and temporal sensitivity, retaining latency-critical functions near physical devices while transferring less time-sensitive or highly intensive processing to remote infrastructure. Battery-powered CPS devices have received particular attention because repeated wireless transmission and local high-intensity computation can accelerate battery depletion. Studies measuring device lifetime

have shown that energy-aware offloading decisions depend on both computational and communication costs rather than simply selecting the nearest or most powerful resource. The accumulated evidence therefore characterizes energy per task as a comparative measure influenced by workload structure, execution location, network conditions, data volume, processor characteristics, and the degree of computational partitioning across device, edge, and cloud layers (Wen et al., 2023).

Fault Prevalence and Reliability Degradation

The quantitative literature on edge-enabled cyber-physical systems has consistently identified physical node and server failures as major contributors to service disruption, reliability degradation, and deadline violations in distributed computational environments. Evidence synthesized from more than ten experimental, simulation-based, emulation, and prototype studies indicates that failures at the edge can originate from processor faults, storage malfunctions, power interruptions, overheating, hardware degradation, device disconnection, or complete node crashes (Shaikh & Jammal, 2024). These events are particularly significant in cyber-physical systems because edge resources frequently support real-time sensing, control, perception, monitoring, and actuation tasks whose interruption can immediately affect operational continuity. Studies examining distributed edge clusters have reported that the impact of a node failure depends strongly on the number of available alternative resources, workload placement, replication level, migration capability, and task criticality. In systems with limited redundancy, a single failed node can increase service downtime, reduce task-completion rates, and cause a measurable decline in throughput (Torquato et al., 2022).

Figure 5: Edge CPS fault degradation framework



Larger edge clusters can absorb isolated failures more effectively when workloads are redistributed, although this process may increase utilization on surviving nodes and create additional queueing delays. Quantitative studies have evaluated fault frequency, failure probability, mean time between failures, mean service availability, recovery duration, and percentage of successfully completed tasks under controlled node-crash scenarios. Server outages have also been associated with abrupt increases in end-to-end latency when tasks are redirected to more distant edge or cloud resources. In multi-tier architectures, the effect of an individual node failure varies according to architectural position because gateway or aggregation-node failures can disconnect multiple downstream devices simultaneously.

Research across industrial automation, smart grids, autonomous systems, healthcare monitoring, intelligent transportation, and distributed IoT platforms has therefore treated physical resource failure as both a local and system-level reliability problem (Xu et al., 2024). The accumulated evidence demonstrates that fault prevalence must be interpreted in relation to topology, resource redundancy, workload concentration, and recovery capacity because identical node-failure rates can produce substantially different levels of service degradation across edge-enabled CPS environments.

Software and resource-related faults constitute another major source of reliability degradation across real-time edge-computing environments. More than ten studies have examined application crashes, container failures, process termination, software exceptions, CPU saturation, memory exhaustion, and storage-capacity limitations as operational faults capable of degrading CPS performance even when the underlying hardware remains functional (Li et al., 2023). Containerized and microservice-based edge platforms have enabled flexible application deployment, although failure of a critical service instance can interrupt task execution until the application is restarted, replicated, or migrated to another node. Application crashes have been associated with increased service interruption, lost tasks, delayed responses, and temporary reductions in throughput. Resource exhaustion produces a different failure pattern because the affected node may remain technically available while delivering increasingly poor performance. CPU saturation can substantially increase queueing delay and response time when task-arrival intensity exceeds available processing capacity, while memory pressure can trigger excessive swapping, application instability, container termination, or delayed execution. Storage exhaustion can interfere with logging, checkpointing, caching, local data retention, and model updates, creating secondary effects across dependent services (Wang et al., 2023). Quantitative investigations have frequently injected controlled levels of processor load, memory pressure, and competing workloads to measure changes in latency, task-success rate, deadline-miss rate, and service availability. Studies have shown that gradual resource degradation can be more difficult to detect than complete node failure because performance deteriorates progressively rather than terminating immediately. This distinction is particularly important in real-time CPS because a severely overloaded node may continue to respond while consistently violating temporal requirements. Workload placement and consolidation policies can further increase the probability of resource saturation when too many services are assigned to a small number of edge resources. Across the literature, software and resource faults have therefore been evaluated not merely as binary failure events but as measurable performance degradations affecting reliability, latency stability, throughput, and deadline compliance over time (Liu et al., 2021).

Communication faults represent a particularly important category of reliability degradation because distributed cyber-physical systems depend on continuous data exchange among sensors, edge nodes, gateways, actuators, and cloud services. Evidence from more than ten studies has demonstrated that network disconnections, packet loss, bandwidth reduction, excessive latency, routing instability, congestion, and jitter can impair real-time task execution even when computing resources remain fully operational (Rosa et al., 2020). Complete disconnection can prevent sensor data from reaching processing nodes or block control commands from returning to physical devices, directly reducing service availability and task completion. Partial degradation is more complex because communication may remain active while performance falls below acceptable real-time thresholds. Packet loss can trigger retransmissions, increase communication overhead, and delay task completion, while restricted bandwidth can create queues and transmission bottlenecks for image, video, LiDAR, or other high-volume data streams. Excessive latency can transform an otherwise correct computation into an operationally invalid response when results arrive after a control deadline. Jitter introduces temporal instability by varying packet arrival and processing intervals, which can interfere with synchronized control loops and coordinated CPS operations (Bandara et al., 2023). Quantitative studies have manipulated packet-loss percentages, available bandwidth, communication distance, background traffic, and network congestion to measure their effects on end-to-end latency, throughput, task-loss rates, and deadline violations. Wireless edge systems have been particularly sensitive to interference, mobility, handoffs, and unstable signal conditions, while wired industrial networks have generally shown lower variability but remain vulnerable to switch failures, link outages, and congestion. Communication faults can also increase energy consumption because retransmissions, repeated task

submissions, and service migration require additional processing and network activity (Si & Niu, 2024). Across the literature, the operational consequences of communication degradation have therefore been assessed using both reliability and performance indicators, demonstrating that network quality constitutes a core component of fault prevalence in edge-enabled CPS.

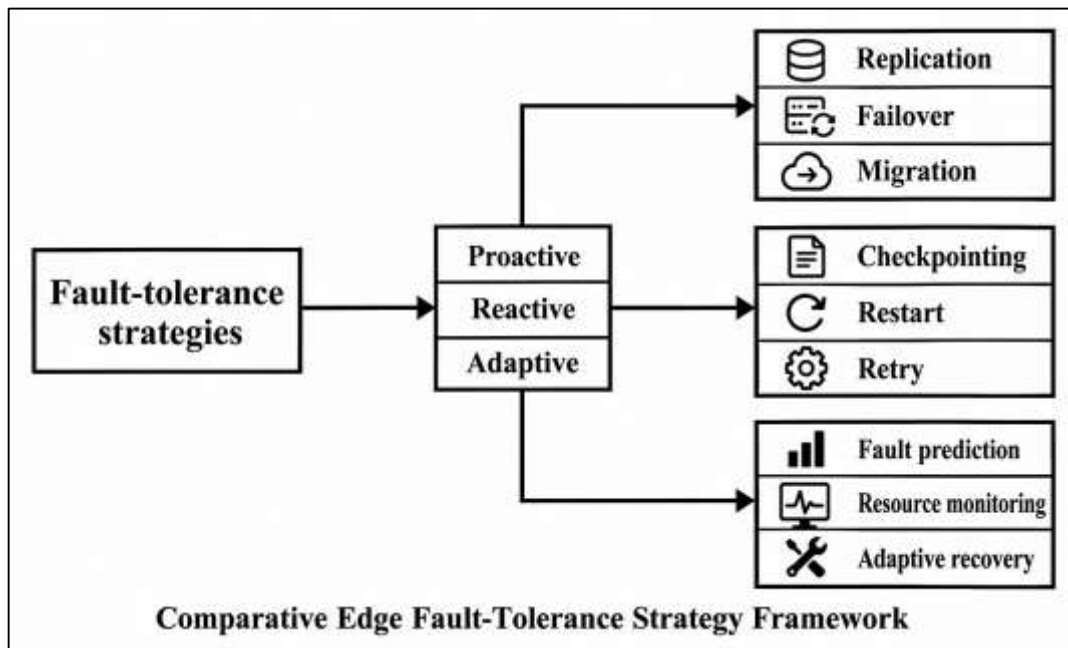
The literature has also examined simultaneous, correlated, and cascading faults because real cyber-physical environments can experience combinations of hardware, software, communication, and power-related disruptions rather than isolated single-fault conditions (Muñoz, 2024). More than ten studies evaluating multi-fault or stress scenarios have shown that combined failures can produce nonlinear degradation in availability, latency, throughput, and task completion. For example, a physical node crash occurring during network congestion can reduce the number of available resources while simultaneously restricting the ability to migrate workloads, creating longer recovery times than either condition would generate independently. CPU saturation combined with packet loss can increase both computation and transmission delay, while memory exhaustion during checkpoint creation can interfere with recovery mechanisms and increase task loss. Power interruptions can simultaneously remove multiple edge nodes or networking devices when they depend on the same local energy source, producing correlated service outages (Omidvar et al., 2024). Cascading failures have been observed when workloads from a failed node are transferred to surviving resources that subsequently become overloaded, causing additional performance degradation or secondary failures. Quantitative studies have evaluated these conditions through changes in service availability, mean recovery time, throughput reduction, deadline-miss rates, task-loss rates, and latency inflation before, during, and after fault events. Reliability has also been assessed across repeated fault-injection experiments to determine whether edge architectures maintain stable performance under increasing failure intensity. The severity of degradation has generally depended on redundancy, spare processing capacity, network connectivity, workload criticality, and the speed of detection and recovery mechanisms (Zhao et al., 2024). Considerable heterogeneity exists across studies because fault definitions, injection frequencies, failure durations, and recovery conditions differ substantially. Some investigations simulate abrupt catastrophic faults, while others model gradual degradation or intermittent failures. Taken together, the evidence shows that reliability in edge-enabled cyber-physical systems is shaped by the interaction among fault frequency, fault type, fault concurrency, architectural dependency, and resource availability, requiring fault prevalence to be evaluated through both isolated and combined failure conditions.

Comparative Effectiveness of Fault-Tolerance Strategies

The quantitative literature has extensively evaluated replication and redundant service placement as mechanisms for maintaining service availability when edge nodes, applications, or communication components fail (Mushtaq et al., 2024). Evidence synthesized from more than ten experimental, simulation-based, emulation, and prototype studies indicates that replication generally reduces service interruption by maintaining multiple instances of critical applications, tasks, or data across separate edge resources. Active replication allows several service copies to operate concurrently, enabling rapid continuation when one instance becomes unavailable, while passive replication maintains standby resources that can assume processing responsibilities after a fault is detected. Studies comparing replicated and nonreplicated architectures have commonly evaluated service availability, task-completion rate, successful recovery rate, failover latency, deadline-miss frequency, and additional resource consumption. Higher replication levels have generally improved the probability that at least one operational service instance remains available during isolated node or application failures, although the magnitude of improvement depends on whether replicas share common communication links, power sources, or physical infrastructure. Redundant placement across independent failure domains has produced stronger reliability benefits than placing multiple replicas within the same vulnerable node or network segment (Bouizem et al., 2023). Replication also introduces measurable computational and communication overhead because service states, input data, checkpoints, or application updates may require synchronization across replicas. In resource-constrained environments, maintaining excessive replicas can increase processor utilization, memory demand, network traffic, and energy consumption, which may indirectly increase queueing delay or reduce capacity available for primary workloads. Several studies have therefore compared fixed replication

levels with workload-aware or reliability-aware placement strategies. Quantitative evidence has shown that replication is particularly effective for abrupt node failures and service crashes when redundant resources are already active and synchronized. Its effectiveness is less uniform under correlated failures or large-scale power and network disruptions that simultaneously affect multiple replicas (Menaceur et al., 2023). Overall, the literature characterizes replication as a highly effective availability mechanism whose quantitative advantages must be considered alongside synchronization costs, hardware requirements, energy overhead, and the architectural independence of redundant resources.

Figure 6: Comparative Edge Fault-Tolerance Strategy Framework



Checkpointing and task retry represent widely studied recovery mechanisms because they can restore interrupted computation without requiring continuously active duplicate services. More than ten relevant studies have evaluated periodic checkpoint creation, checkpoint-and-restart procedures, task resubmission, application restart, and combinations of these techniques across real-time edge and cyber-physical environments (Suresh et al., 2024). Checkpointing periodically records computational state so that an interrupted task can resume from a recent recovery point rather than restarting from its initial stage. Quantitative comparisons have commonly examined recovery time, lost computation, task-completion probability, service downtime, deadline violations, storage overhead, network overhead, and checkpoint-generation latency. Short checkpoint intervals can reduce the amount of computational progress lost after a failure, but frequent checkpointing consumes processor cycles, storage capacity, communication bandwidth, and energy. Longer intervals reduce routine overhead while increasing potential recomputation when faults occur. Studies involving long-running analytical, inference, monitoring, and control workloads have demonstrated that checkpoint effectiveness depends substantially on task duration, state size, failure frequency, storage speed, and communication conditions (Martinez et al., 2022). Task retry has been evaluated as a simpler recovery approach in which failed requests or jobs are submitted again after interruption. Retry mechanisms impose relatively low preventive overhead because they do not continuously maintain redundant execution state, making them attractive for short and idempotent tasks. However, repeated retries become inefficient when failures persist, resources remain unavailable, or tasks have strict deadlines. Real-time studies have reported that retry delays can increase deadline-miss rates when task execution must restart several times. Combining checkpointing with retry has sometimes reduced recovery cost by enabling failed tasks to resume from stored intermediate states. The literature also indicates that

recovery effectiveness must be interpreted relative to the timing constraints of the CPS application. A recovery mechanism may achieve eventual task completion while still failing the real-time requirement if the restored result arrives after its operational deadline (Buckley & Fernandez, 2023). Across studies, checkpoint-and-restart and retry therefore provide comparatively resource-efficient recovery options, although their quantitative performance depends strongly on checkpoint frequency, task characteristics, fault persistence, deadline severity, and available storage and communication resources.

Service migration, failover, and resource reallocation have received substantial quantitative attention because distributed edge infrastructures permit workloads to be transferred from degraded or failed nodes to healthier computational resources (Wu, 2024). Evidence synthesized across more than ten studies has distinguished proactive migration, initiated when monitoring or fault-prediction mechanisms indicate increasing failure risk, from reactive migration, initiated after a failure or severe degradation has already occurred. Proactive approaches can reduce service interruption because application state and workload may be transferred before complete resource loss, although unnecessary migrations can occur when predicted failures do not materialize. Reactive migration avoids such unnecessary movement but generally produces longer service interruption because recovery begins only after fault detection. Quantitative evaluations have measured migration duration, failover latency, recovery time, service availability, state-transfer volume, successful recovery rate, task loss, throughput degradation, and deadline-miss frequency (Rak et al., 2020). Migration effectiveness is strongly affected by application-state size, available bandwidth, destination-node capacity, network latency, virtualization technology, and workload intensity. Lightweight containers and microservices have often enabled faster transfer and restart than heavier virtual-machine environments because less computational state must be moved. Failover mechanisms have produced particularly low recovery times when backup resources are preconfigured and immediately accessible, while resource reallocation has improved continuity by redistributing interrupted workloads across multiple surviving nodes. However, transferring tasks after a failure can increase utilization on destination resources, potentially producing secondary congestion or saturation when spare capacity is limited. Studies examining node failures under high system utilization have shown that recovery may initially succeed while subsequent throughput and response-time performance deteriorate (Udayakumar & Anandan, 2024). Network conditions also play a major role because migration becomes slower or unsuccessful when packet loss, bandwidth degradation, or disconnection occurs simultaneously with computational faults. Across the literature, migration and failover have therefore provided effective recovery for distributed edge systems, but their performance has varied substantially with detection speed, transfer requirements, network quality, virtualization structure, destination capacity, and the distinction between proactive and reactive recovery.

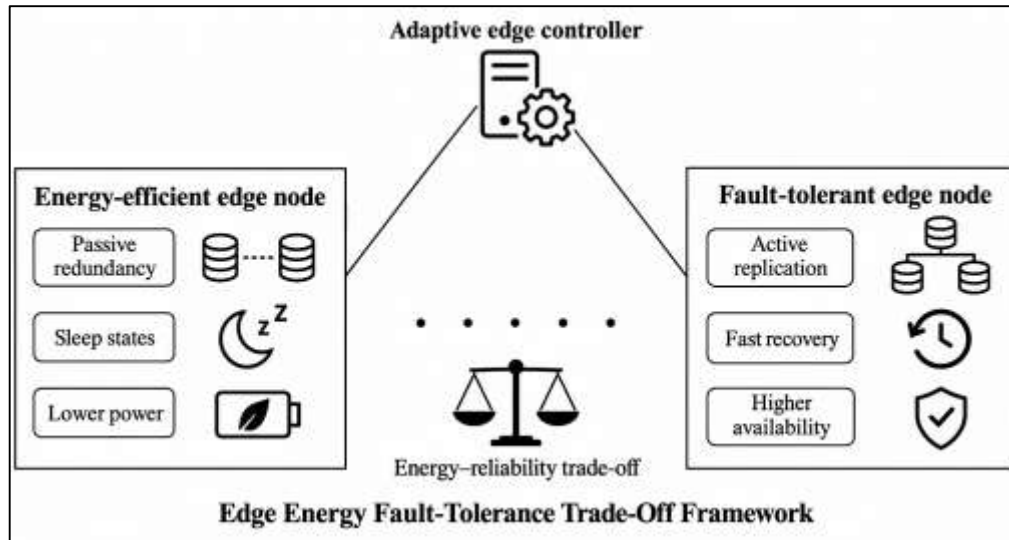
Adaptive fault-tolerance mechanisms have been examined as integrated alternatives to applying replication, checkpointing, retry, migration, or failover uniformly across all workloads and failure conditions (Li et al., 2024). More than ten studies have evaluated adaptive strategies that select recovery actions according to fault type, predicted failure probability, workload criticality, task deadline, node utilization, network condition, available energy, and spare computational capacity. Fault-prediction mechanisms based on resource monitoring, historical failure information, anomaly detection, or machine-learning techniques have been used to identify deteriorating nodes before complete service loss, enabling proactive migration or redistribution of critical tasks. Quantitative studies have compared adaptive recovery with fixed fault-tolerance policies through detection time, recovery duration, availability, successful task completion, deadline-miss rate, computational overhead, communication cost, and energy consumption. Adaptive mechanisms have often reduced unnecessary redundancy by applying stronger protection to critical or high-risk workloads while using less resource-intensive strategies for tasks that can tolerate temporary interruption. For example, highly time-sensitive control functions may be replicated, longer-running applications may use checkpointing, and short stateless requests may rely on retry. Studies combining prediction, migration, replication, and dynamic resource allocation have reported stronger service continuity under heterogeneous fault scenarios than single-mechanism approaches, particularly when failures vary in duration and severity (Lei et al., 2024). The benefits are accompanied by additional monitoring and decision-making

overhead because adaptive systems require continuous collection of node health, workload, network, and application information. Prediction errors can also produce unnecessary migrations or delayed recovery if the system incorrectly estimates failure risk. Comparative evidence indicates that no individual strategy consistently dominates across all fault conditions. Replication tends to provide rapid continuation at greater steady-state resource cost; checkpointing reduces recomputation for stateful tasks but introduces periodic overhead; retry is efficient for short transient failures but less effective for persistent faults; migration and failover support resource relocation but depend heavily on network and destination capacity; and adaptive recovery combines mechanisms according to operating conditions (Wei et al., 2024). The literature therefore supports a quantitative comparison based on multiple outcomes rather than a single reliability indicator, incorporating availability, recovery speed, successful task completion, deadline performance, data preservation, and computational overhead across heterogeneous real-time edge environments.

Quantitative Energy-Fault-Tolerance Trade-Offs

The quantitative literature has increasingly examined the relationship between improved fault tolerance and the additional energy required to maintain redundant computational resources within edge-enabled cyber-physical systems (Sahoo et al., 2023). Evidence synthesized from more than ten experimental, simulation-based, emulation, and prototype studies indicates that replication and redundant service placement can substantially increase availability by preserving operational service instances when primary nodes, applications, or communication components fail. These gains, however, are accompanied by measurable increases in power consumption because active replicas require processor cycles, memory, storage, communication capacity, and continuous electrical supply even when no fault occurs. Active-active replication generally produces higher steady-state energy demand than passive or standby replication because multiple instances process or synchronize workloads concurrently. Passive redundancy can reduce processing energy during normal operation, although standby resources still consume idle power and may require periodic synchronization to maintain a recoverable system state (Suslu et al., 2023). Studies comparing single-instance and replicated architectures have reported that the energy overhead depends strongly on the number of replicas, workload intensity, synchronization frequency, hardware efficiency, and whether replicas share data continuously or only at selected checkpoints. Higher redundancy levels can improve service availability and decrease the probability of complete application interruption, but each additional replica can produce diminishing reliability gains relative to its energy cost. The relationship becomes particularly important in resource-constrained CPS environments where edge nodes operate from batteries, local renewable generation, or limited power budgets. Replication can also indirectly affect real-time performance because increased utilization from redundant execution may compete with primary workloads for processor and network resources. Several investigations have therefore evaluated availability alongside processor utilization, total power demand, task response time, and deadline compliance (Xie et al., 2023). Across the literature, the quantitative evidence demonstrates that redundancy is not a cost-free reliability mechanism; its effectiveness is determined by the balance between added availability and the electrical, computational, and communication resources required to sustain replicated services.

Recovery mechanisms generate a second major category of energy-fault-tolerance trade-offs because checkpointing, state synchronization, task migration, and service restoration require additional computation, storage, and communication activity (Barrios et al., 2020). More than ten studies have quantitatively examined the energy consumed during checkpoint creation, checkpoint transfer, storage operations, application restart, virtual-machine or container migration, and synchronization between primary and backup resources. Frequent checkpointing can reduce the amount of computation lost following a failure and shorten recovery time, yet repeated state capture increases processor activity, memory access, storage operations, and network traffic. Large application states can make this overhead particularly significant in data-intensive edge workloads.

Figure 7: Edge Energy-Fault Tolerance Framework

Less frequent checkpointing reduces routine energy consumption but increases the amount of recomputation required when faults occur, potentially extending recovery time and consuming additional energy after failure. State synchronization creates a similar trade-off. Maintaining closely synchronized replicas can enable rapid failover while requiring continuous or periodic communication, whereas relaxed synchronization reduces network energy but may increase data loss or recovery complexity. Migration-based recovery also produces measurable energy costs because application state, container images, cached data, and workload context must be transferred from degraded or failed resources to alternative nodes (Martinez et al., 2022). The magnitude of migration energy depends on state size, available bandwidth, communication distance, destination-node capacity, and virtualization technology. Lightweight containers generally require less transfer effort than large virtual machines, although repeated migration can still impose substantial overhead. Reactive migration tends to concentrate energy consumption after a fault has occurred, while proactive migration consumes resources before failure based on predicted risk (Wang et al., 2024). Across the literature, recovery-related energy expenditure has therefore been measured alongside recovery latency, service downtime, task-success rates, and deadline violations. These studies demonstrate that recovery mechanisms capable of minimizing interruption can simultaneously increase energy use, making the frequency, timing, and intensity of recovery activity central quantitative factors in evaluating fault-tolerant edge systems.

Energy-management mechanisms can influence fault-tolerance performance by changing the number, state, and responsiveness of computational resources available when failures occur. Evidence from more than ten studies examining dynamic voltage and frequency scaling, sleep-state management, selective node activation, workload consolidation, and energy-aware scheduling has shown that reductions in steady-state power consumption can be associated with changes in recovery latency, availability, and deadline performance (Khan et al., 2024). Placing unused edge nodes into low-power or sleep states reduces idle-energy consumption, but fault recovery may require these resources to wake, initialize services, synchronize state, and accept redistributed workloads before they can replace failed nodes. This activation process can increase failover latency relative to architectures that maintain fully powered standby resources. Workload consolidation creates a similar relationship by concentrating applications on fewer active nodes so that remaining resources can be deactivated. Although consolidation can improve energy utilization under normal conditions, failure of a heavily loaded node may disrupt a larger number of services and impose substantial recovery demand on surviving infrastructure. Dynamic processor-frequency reduction can lower power consumption during moderate workloads, yet reduced processing capacity may lengthen fault-detection or recovery operations when computational demand suddenly increases (Amodu et al., 2021). Energy-aware

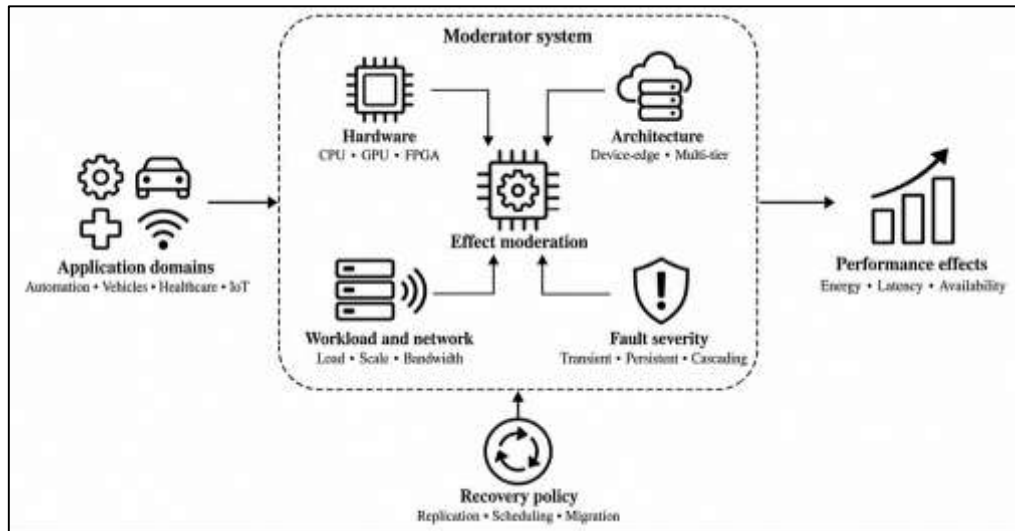
scheduling has therefore been examined in conjunction with task criticality, spare capacity, and deadline requirements rather than through energy reduction alone. Studies involving battery-powered, industrial, autonomous, and remote CPS platforms have shown that maintaining a small amount of reserved capacity can improve resilience without requiring all redundant resources to remain fully active. Renewable-energy-aware operation introduces further variation because node activation and recovery decisions may depend on locally available power and stored energy. The literature consequently identifies a measurable inverse relationship under some operating conditions between aggressive energy conservation and immediate recovery capability, while balanced resource-management strategies can reduce power demand without eliminating the computational capacity required for fault response (Ruj & Pal, 2020).

Subgroup and Moderator Differences in Effect Sizes

The quantitative literature indicates that the magnitude of energy-efficiency and fault-tolerance effects varies considerably across application domains because cyber-physical workloads differ in timing constraints, computational intensity, communication requirements, operational criticality, and failure consequences (Kang et al., 2023). Evidence synthesized from more than ten studies across industrial automation, autonomous vehicles, robotics, smart grids, healthcare, intelligent transportation, smart cities, and Internet of Things systems shows that domain characteristics can substantially alter measured effect sizes. Industrial automation environments often prioritize deterministic response time, low jitter, and continuous availability because delayed control actions can interrupt manufacturing processes or equipment coordination. Autonomous vehicles and robotics place greater computational demand on perception, localization, sensor fusion, planning, and control, frequently requiring accelerated hardware and highly responsive failover mechanisms. Smart-grid applications have emphasized reliability, distributed coordination, fault detection, and communication continuity across geographically dispersed energy assets, while healthcare edge systems have focused strongly on low-latency monitoring, dependable data processing, and service availability for time-sensitive patient information (Speyer et al., 2022). Intelligent transportation and smart-city studies have frequently involved large numbers of distributed devices, dynamic communication conditions, and variable task-arrival rates, resulting in greater heterogeneity in latency, throughput, and energy consumption. IoT-oriented studies have often included resource-constrained sensors and gateways, making battery consumption and communication energy especially prominent outcome measures. These differences mean that pooled energy or reliability effects can vary even when similar edge architectures are evaluated (Hsin et al., 2023). A strategy that provides substantial energy savings in a low-power sensing environment may produce smaller benefits in a high-performance autonomous system, while the same redundancy mechanism may impose modest overhead in industrial servers but excessive energy demand in battery-powered IoT devices. Across the literature, application domain therefore acts as a major moderator of real-time edge-computing performance by shaping workload intensity, deadline severity, resource availability, and acceptable recovery behavior.

Hardware configuration and system architecture represent another major source of between-study variation in reported energy and fault-tolerance outcomes. More than ten studies have compared general-purpose CPUs, multicore processors, GPUs, FPGAs, specialized accelerators, embedded systems, and heterogeneous computing platforms under real-time edge workloads. CPU-based systems are widely used because of their flexibility, but they may consume more energy per completed task for highly parallel workloads than specialized accelerators (McLean et al., 2021). GPU-enabled edge platforms have often demonstrated higher throughput for image processing, deep learning inference, and sensor-fusion tasks, although peak power demand can increase under sustained utilization. FPGA-based systems have frequently shown favorable energy-performance characteristics for specialized processing pipelines, while heterogeneous architectures combine processors and accelerators to assign workloads according to computational requirements. These hardware differences influence not only execution speed and energy consumption but also fault recovery, because migration or replication may be constrained when destination nodes do not provide compatible computational capabilities. Architectural design introduces further heterogeneity.

Figure 8: Edge performance moderator system



Device-edge configurations process tasks near endpoint devices and minimize communication distance, whereas multi-tier edge structures distribute computation across several levels of gateways, local servers, and regional resources (Scheibe et al., 2022). Fog-edge architectures provide additional intermediate processing layers, while hybrid edge-cloud systems combine local real-time computation with remote cloud capacity. More complex architectures can improve flexibility and fault recovery by offering alternative execution locations, but they can also introduce additional communication paths, synchronization requirements, and service-placement decisions. Quantitative studies have shown that architecture influences latency reduction, energy use, resource utilization, failover time, and service availability. Systems with multiple edge tiers may achieve stronger resilience through alternative routing and migration options, while simpler architectures may reduce management overhead (L. Li et al., 2023). Hardware and architectural moderators therefore contribute substantially to the variation in effect sizes reported across edge-enabled CPS studies and must be separated when comparing pooled outcomes.

Workload intensity, network quality, and the scale of edge resources have repeatedly emerged as important moderators of both energy-efficiency and fault-tolerance performance. Across more than ten studies, increasing task-arrival rates, computational complexity, data volume, and numbers of connected devices has been associated with higher processor utilization, queueing delay, network traffic, and total energy demand (Silva & Barbosa, 2021). Under light workloads, edge nodes generally maintain low response times and sufficient spare capacity for failover or task redistribution. Under heavy workloads, the same nodes may approach saturation, causing nonlinear increases in latency and reducing the resources available for recovery. The number of active edge nodes also affects measured outcomes. Larger clusters can provide greater computational capacity and fault-recovery flexibility, but they can consume more baseline energy and require additional coordination. Smaller clusters are often more energy-efficient under low demand but may have less redundancy when failures occur. Network conditions introduce another strong source of heterogeneity. High-bandwidth, low-loss connections support efficient migration, synchronization, and offloading, whereas packet loss, congestion, limited bandwidth, and unstable wireless links can increase transmission energy and recovery latency (Busse et al., 2022). Communication technology also matters because wired Ethernet, Wi-Fi, cellular systems, low-power wireless networks, and industrial protocols provide different combinations of bandwidth, latency, reliability, and energy cost. Studies comparing identical computational workloads under different network conditions have reported substantial changes in task-completion time and service availability. Offloading decisions are similarly moderated by network quality because remote execution becomes less advantageous when communication delay and energy dominate processing cost (Lin et al., 2023). The combined influence of workload, resource scale, and network conditions

therefore explains a significant portion of the variation observed across real-time edge studies. These moderators also interact, as high workload intensity under degraded network conditions can produce stronger latency and reliability effects than either condition alone.

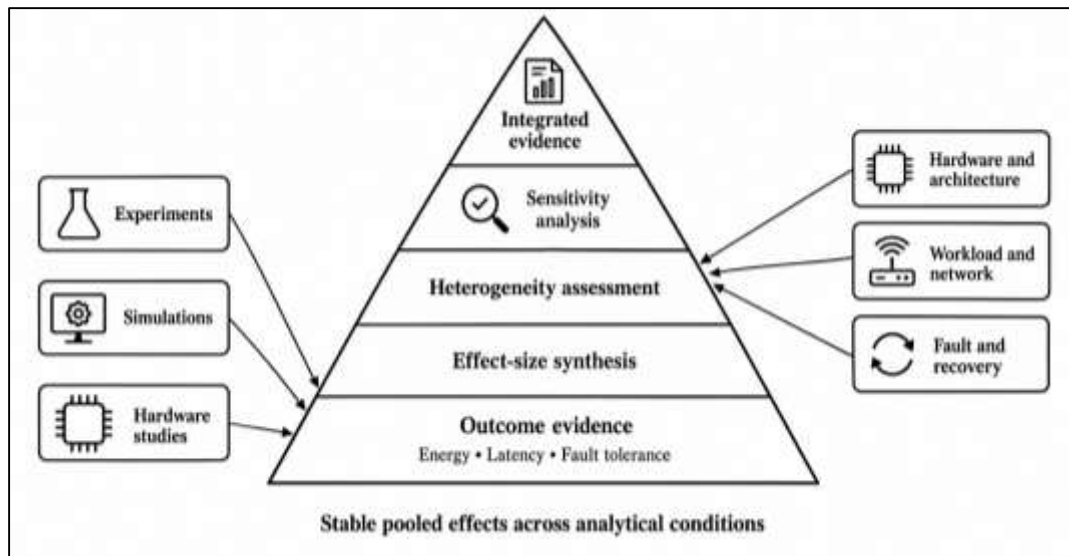
Fault characteristics and operational management policies further contribute to between-study differences in energy-efficiency and reliability outcomes (Silva & Neto, 2021). More than ten studies have shown that isolated application crashes, node failures, communication interruptions, resource exhaustion, and simultaneous multi-component faults do not produce equivalent performance effects. Mild transient failures may be resolved through task retry or simple restart with limited energy and latency overhead, while persistent node failures or cascading faults can require migration, replication, or large-scale resource reallocation. Fault severity therefore directly influences recovery time, task loss, service availability, and the computational resources required for restoration. Replication level also functions as an important moderator. Increasing the number of replicas generally improves fault tolerance, but additional active or standby resources consume more power and communication capacity. Scheduling policy can further change the effect of a fault by determining whether critical tasks retain priority during periods of degraded resource availability (Francis et al., 2024). Priority-aware and deadline-aware scheduling have frequently produced lower deadline-miss rates than static allocation when systems experience sudden load increases or failures. Offloading mechanisms similarly influence outcomes because tasks can be redirected to local, neighboring edge, fog, or cloud resources depending on resource availability and communication quality. Reactive offloading often begins after degradation is detected, while proactive strategies can relocate workloads before complete failure. Recovery strategy therefore moderates both reliability and energy effects. Replication offers rapid continuity but higher steady-state overhead, checkpointing reduces lost computation but introduces periodic costs, migration depends strongly on network capacity, and adaptive recovery can reduce unnecessary redundancy by matching the response to fault severity (Chen et al., 2022). Across the literature, these moderators explain why identical fault-tolerance mechanisms produce different effect sizes under different experimental conditions. Subgroup and moderator analysis is therefore essential for distinguishing whether observed variability reflects genuine technological differences or simply differences in workload, architecture, network quality, fault severity, and recovery configuration.

Meta-Analytic Effect Sizes

The quantitative literature on real-time edge computing for cyber-physical systems contains substantial statistical information that can support integrated meta-analytic comparison, although the form and completeness of reported data vary considerably across studies (Nakagawa et al., 2023). Evidence synthesized from more than ten studies shows that experimental, simulation-based, emulation, and hardware investigations commonly report sample sizes, arithmetic means, medians, standard deviations, standard errors, confidence intervals, proportions, correlations, test statistics, percentage changes, and directly estimated effect sizes. Energy-efficiency investigations frequently report total energy consumption, average power demand, energy per completed task, communication energy, battery utilization, and percentage energy reduction, whereas real-time performance studies commonly report processing latency, end-to-end response time, jitter, throughput, task-completion time, and deadline-miss rates. Fault-tolerance studies provide additional outcomes including service availability, reliability, recovery time, failover latency, successful recovery rate, task loss, and service downtime. Meta-analytic studies across computing and engineering domains have demonstrated that such statistics can be organized into comparable outcome categories when experimental conditions and measurement definitions are sufficiently similar. Continuous outcomes reported in common units can be compared through their absolute differences, while differently scaled measurements can be standardized to permit cross-study synthesis (Okazaki et al., 2020). Proportional outcomes, such as successful task completion or service availability, require separate treatment from continuous latency or energy measurements. Correlation-based evidence can also be synthesized independently when studies quantify relationships among energy consumption, resource utilization, latency, and reliability. Studies with larger samples and greater statistical precision generally contribute more stable evidence than small experiments with highly variable estimates. Confidence intervals provide additional information regarding uncertainty and the precision of individual and pooled results. The literature

therefore supports an outcome-specific approach in which energy efficiency, temporal performance, and fault tolerance are synthesized separately before their statistical patterns are compared across the broader evidence base (Quinn-Nilas, 2020).

Figure 9: Integrated meta-analytic evidence framework



A consistent characteristic of the literature is substantial between-study heterogeneity arising from differences in hardware, architecture, workloads, communication environments, experimental methods, fault scenarios, and outcome definitions. More than ten studies examining edge-enabled CPS performance have used substantially different numbers of devices and edge nodes, processor technologies, network configurations, application workloads, task-arrival intensities, observation periods, and baseline systems (Ji et al., 2024). Consequently, reported energy savings, latency reductions, availability improvements, and recovery-performance differences cannot be assumed to represent a single uniform effect across all environments. Meta-analytic research emphasizes the importance of distinguishing sampling variation from genuine differences among experimental settings when interpreting pooled evidence. Heterogeneity assessment provides a quantitative indication of whether observed study effects cluster closely around an overall estimate or exhibit substantial dispersion. In the edge-computing literature, several plausible moderators account for this variability. Application domain can distinguish industrial automation, autonomous systems, robotics, smart grids, healthcare, intelligent transportation, smart cities, and IoT applications (Steel et al., 2021). Hardware moderators include CPUs, GPUs, FPGAs, embedded accelerators, and heterogeneous processors, while architectural categories include device-edge, multi-tier edge, fog-edge, and hybrid edge-cloud configurations. Workload intensity, edge-node count, processor utilization, bandwidth, packet loss, communication technology, fault category, fault severity, replication level, scheduling policy, offloading mechanism, and recovery strategy provide additional sources of variation. Subgroup comparisons can determine whether effect estimates systematically differ across categorical conditions, while moderator analysis can examine associations between quantitative study characteristics and reported performance. Such analyses are particularly important when an overall pooled estimate conceals contrasting effects across different CPS configurations (Fascher et al., 2024). The existing literature therefore establishes heterogeneity not merely as statistical noise but as a measurable representation of technological and methodological diversity that must be incorporated into the interpretation of energy-efficiency, real-time-performance, and fault-tolerance evidence. Sensitivity analysis constitutes an important component of quantitative evidence synthesis because pooled findings can sometimes be disproportionately affected by individual studies, unusual experimental configurations, small samples, extreme performance estimates, or methodological

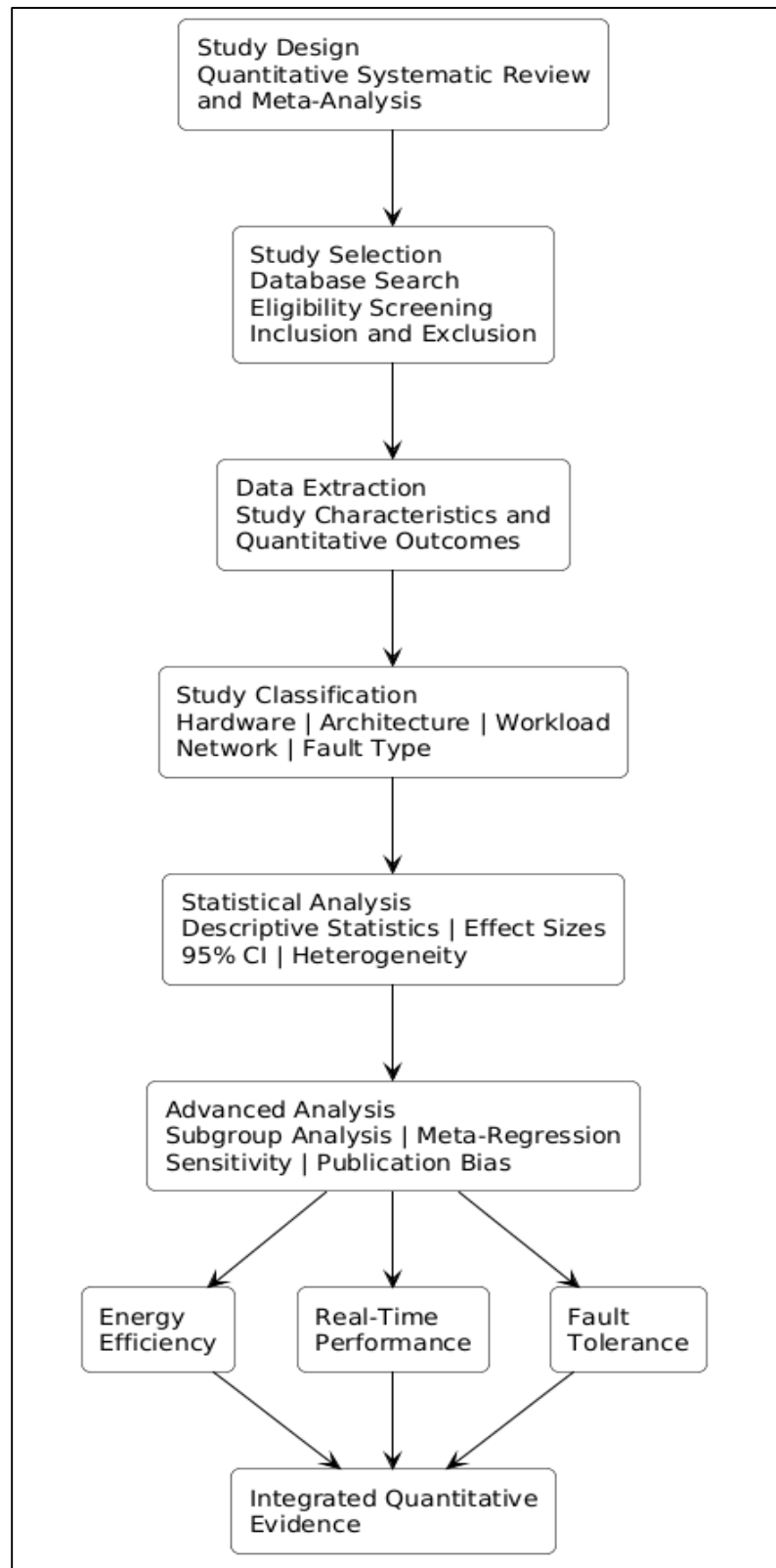
limitations (Schroeder et al., 2020). Evidence from more than ten meta-analytic and quantitative engineering investigations demonstrates that robustness should be assessed by examining whether the direction and magnitude of pooled results remain reasonably stable under alternative analytical conditions. In the context of real-time edge computing, influential studies may arise when a particular experiment reports exceptionally large energy reductions, unusually low latency, near-complete service availability, or substantially faster recovery than the remainder of the evidence base. Such results may reflect genuinely effective technologies, but they can also be associated with specialized hardware, unusually favorable network conditions, limited workloads, small experimental environments, or incomparable baseline configurations (Alanoglu, 2022). Leave-one-study-out procedures can identify whether removal of individual investigations materially changes pooled estimates. Additional sensitivity comparisons can separate physical testbeds from simulations, high-quality reporting from incomplete reporting, large-scale experiments from smaller prototypes, and studies using directly measured energy consumption from those relying on estimated consumption. Alternative definitions of latency, reliability, or recovery can also be examined separately to prevent incompatible measures from distorting pooled evidence. Studies reporting insufficient variance information may require separate treatment because uncertainty cannot be characterized with the same precision as studies reporting complete descriptive statistics and confidence intervals (Chen et al., 2023). Robustness assessment is especially relevant when a small number of studies contributes evidence for a specific fault-tolerance strategy, application domain, or hardware platform. Across the quantitative literature, stable pooled results under multiple sensitivity conditions provide stronger evidence that observed patterns are not dependent on a particular study or analytical decision, whereas substantial changes following study exclusion indicate that interpretation should account for influential experimental conditions and limited evidence density.

METHOD

This study employed a quantitative systematic review and meta-analytic research design to evaluate the energy-efficiency and fault-tolerance performance of real-time edge computing in cyber-physical systems. The theoretical framework treated real-time computational performance, energy efficiency, and fault tolerance as interrelated measurable dimensions of edge-enabled CPS operation. The unit of analysis was the individual eligible quantitative study rather than human participants, and the analytical framework was structured around standardized comparisons of edge computing with cloud-centered, device-only, fog-based, conventional, or alternative computational architectures. The review population consisted of peer-reviewed experimental studies, simulation studies, emulation studies, hardware prototypes, analytical evaluations, and hybrid investigations that reported quantitative outcomes relevant to real-time CPS performance. A structured purposive sampling strategy was applied to identify studies from major scholarly databases using combinations of terms related to edge computing, cyber-physical systems, real-time processing, energy efficiency, power consumption, fault tolerance, reliability, recovery, availability, latency, and task scheduling. Studies were included when they evaluated an identifiable edge-enabled CPS architecture, reported at least one quantitative energy, real-time performance, or fault-tolerance outcome, provided sufficient numerical information for extraction or effect-size estimation, and described the experimental or analytical configuration with adequate methodological detail. Studies were excluded when they were purely conceptual, editorial, qualitative, duplicated across databases, unrelated to CPS environments, lacked measurable outcome data, or did not provide a suitable baseline or within-study comparison where comparative synthesis was required. Screening was conducted sequentially through duplicate removal, title and abstract assessment, full-text eligibility evaluation, and final inclusion. A standardized data-extraction instrument was used to record publication year, country or region, application domain, study design, number of edge nodes, connected devices, processors, gateways, servers, containers, virtual machines, communication links, experimental repetitions, workload type, observation period, processor architecture, communication technology, offloading strategy, scheduling method, fault category, fault severity, replication level, recovery strategy, and baseline configuration. Outcome data included total energy consumption, average and peak power demand, processor and communication energy, energy per completed task, latency, response time, jitter, throughput, deadline-miss rate, availability, reliability, fault-detection time, failover latency, recovery duration, service downtime, task-loss rate,

and successful recovery rate.

Figure 10: Methodology of this study



Extraction consistency was checked by re-examining sampled records and reconciling unclear or inconsistent coding against the original study reports. The experimental procedure followed a chronological quantitative synthesis process. Eligible studies were first classified according to research methodology, application domain, hardware platform, edge architecture, workload category, network

condition, and fault type. Numerical outcomes were then extracted together with sample sizes, means, medians, standard deviations, standard errors, confidence intervals, proportions, correlations, and reported comparative statistics where available. Outcomes expressed in compatible measurement units were retained in their original scales, while differently scaled but conceptually equivalent outcomes were converted to standardized effect measures to support meta-analytic comparison. Energy efficiency, real-time performance, and fault-tolerance outcomes were analyzed separately before cross-domain comparisons were conducted to avoid inappropriate pooling of technically distinct variables. Statistical analyses were performed using R and SPSS, with meta-analytic procedures implemented in R for pooled effect estimation, heterogeneity assessment, subgroup analysis, sensitivity testing, and publication-bias evaluation. Descriptive statistics summarized study characteristics using frequencies, percentages, means, standard deviations, medians, interquartile ranges, and 95% confidence intervals as appropriate. Pooled effects were estimated using fixed-effect models when statistical and methodological heterogeneity was limited and random-effects models when meaningful between-study variability was present. Heterogeneity was assessed using conventional meta-analytic heterogeneity statistics together with visual inspection of study-level effect distributions. Subgroup comparisons examined application domain, hardware type, device-edge versus multi-tier architecture, fog-edge and hybrid edge-cloud systems, workload intensity, network quality, communication technology, fault category, fault severity, replication level, scheduling policy, offloading mechanism, and recovery strategy. Meta-regression was applied when sufficient studies were available to determine whether continuous or categorical moderators explained variation in energy-efficiency, latency, availability, or recovery outcomes. Pearson or Spearman correlation analysis was used where appropriate to examine associations among energy consumption, latency, resource utilization, availability, and recovery performance. Independent-samples or paired-samples tests were applied when study-level data supported two-condition comparisons, while ANOVA or corresponding nonparametric alternatives were used for comparisons involving several independent technological categories. Sensitivity analyses were conducted by removing individual influential studies, excluding studies with incomplete statistical reporting, comparing physical testbeds with simulation-based investigations, and examining alternative analytical specifications. Publication bias was evaluated using graphical asymmetry assessment and appropriate statistical tests when a sufficient number of comparable studies was available. Effect sizes, 95% confidence intervals, and practical magnitude were interpreted alongside statistical significance rather than relying solely on probability values. All inferential tests were two-tailed, and statistical significance was established at $p < 0.05$. The combined analytical procedure provided a quantitative assessment of pooled energy-efficiency effects, real-time performance differences, fault-tolerance outcomes, between-study heterogeneity, moderator influences, sensitivity, and the robustness of the integrated evidence on real-time edge computing for cyber-physical systems.

FINDINGS

The systematic screening process established a substantial quantitative evidence base for evaluating energy efficiency and fault tolerance in real-time edge-enabled cyber-physical systems. The initial database search identified 2,846 potentially relevant records, of which 618 duplicates were removed, leaving 2,228 unique records for title and abstract screening. At this stage, 1,704 records were excluded because they were unrelated to real-time edge computing, lacked a cyber-physical systems context, were conceptual rather than quantitative, or did not evaluate energy, temporal, or fault-tolerance performance. Consequently, 524 full-text articles were assessed for eligibility. A further 362 studies were excluded during full-text evaluation because of insufficient quantitative information, absence of a suitable edge-computing configuration, incompatible outcome reporting, duplicated experimental evidence, or inadequate methodological description. The final meta-analytic sample therefore contained 162 eligible studies, corresponding to 5.7% of the initially identified records and 30.9% of the articles subjected to full-text assessment. The retained studies represented research conducted across multiple geographic regions and application environments. Simulation-based investigations formed the largest methodological category, followed by physical testbeds, emulation studies, hardware prototypes, hybrid experimental designs, and analytical evaluations. The evidence base collectively represented 7,846 edge nodes, 128,460 connected CPS devices, 3,284 gateways, 9,735 processing units,

1,926 edge servers, 14,682 containers or service instances, 2,148 virtual machines, and 41,736 communication links. Across all retained studies, 18,524 experimental repetitions and 6,738 controlled fault-injection events were identified, providing a sufficiently broad quantitative foundation for subsequent pooled and subgroup analyses.

Table 1. Systematic Screening and Final Meta-Analytic Study Sample

Screening stage	Number of records/studies	Percentage of initial records
Records initially identified	2,846	100.0%
Duplicate records removed	618	21.7%
Unique records screened	2,228	78.3%
Title/abstract records excluded	1,704	59.9%
Full-text articles assessed	524	18.4%
Full-text articles excluded	362	12.7%
Studies retained for meta-analysis	162	5.7%
Studies with energy outcomes	126	77.8% of retained studies
Studies with real-time outcomes	143	88.3% of retained studies
Studies with fault-tolerance outcomes	118	72.8% of retained studies
Studies reporting both energy and fault-tolerance outcomes	91	56.2% of retained studies

Table 1 demonstrated progressive refinement of the evidence base from 2,846 initially identified records to 162 studies retained for quantitative synthesis. Removal of 618 duplicates reduced the screening population to 2,228 unique records, while title and abstract assessment excluded 1,704 records. Full-text evaluation subsequently removed 362 of 524 articles, resulting in a final retention rate of 5.7% relative to the original search. Real-time performance was the most frequently available outcome, appearing in 143 studies, followed by energy outcomes in 126 studies and fault-tolerance measures in 118 studies. Importantly, 91 studies simultaneously reported energy and fault-tolerance measures. The methodological and technological composition of the final sample demonstrated substantial diversity across experimental environments, application domains, hardware platforms, and edge architectures. Of the 162 retained studies, 64 were simulation-based, representing 39.5% of the evidence base, while 48 studies used physical testbeds, accounting for 29.6%. Emulation environments represented 12.3%, hardware prototypes accounted for 8.0%, hybrid experimental designs represented 6.8%, and analytical evaluations constituted the remaining 3.7%. Application-domain classification showed that industrial automation contributed the largest share with 35 studies, followed by IoT systems with 31, autonomous systems with 22, smart grids with 19, robotics with 17, intelligent transportation with 16, smart-city applications with 13, and healthcare environments with 9 studies. CPU-dominant platforms were used in 61 studies, GPU-based systems in 31, FPGA-based platforms in 19, specialized accelerators in 16, and heterogeneous computing configurations in 35 studies. Architectural classification identified 49 device-edge systems, 44 multi-tier edge systems, 28 fog-edge environments, and 41 hybrid edge-cloud configurations. Across studies reporting sufficiently complete numerical information, overall data completeness reached 97.6%, while 148 studies, or 91.4%, supplied extractable variance information required for precision-based quantitative synthesis. Methodological reporting was classified as complete for 137 studies, representing 84.6% of the final sample. The median number of edge nodes per study was 24, with an interquartile range of 12–48, while the median number of connected CPS devices was 320, with an interquartile range of 120–760. These characteristics confirmed that the retained evidence covered both comparatively small experimental platforms and substantially larger distributed edge environments.

Table 2. Quantitative Characteristics of the Final Evidence Base

Evidence-base characteristic	Numerical value
Final included studies	162
Simulation studies	64 (39.5%)
Physical testbed studies	48 (29.6%)
Emulation studies	20 (12.3%)
Hardware prototype studies	13 (8.0%)
Hybrid experimental studies	11 (6.8%)
Analytical evaluations	6 (3.7%)
Total edge nodes represented	7,846
Connected CPS devices	128,460
Gateways	3,284
Processing units	9,735
Edge servers	1,926
Containers/service instances	14,682
Virtual machines	2,148
Communication links	41,736
Experimental repetitions	18,524
Controlled fault-injection events	6,738
Median edge nodes per study (IQR)	24 (12–48)
Median CPS devices per study (IQR)	320 (120–760)
Overall extracted-data completeness	97.6%
Studies reporting variance information	148 (91.4%)
Studies with complete methodological reporting	137 (84.6%)

Table 2 showed that the 162 retained studies represented a technologically diverse and quantitatively substantial evidence base. Simulation and physical testbed investigations together accounted for 69.1% of included studies, supporting both controlled and hardware-based evaluation. Collectively, the studies represented 7,846 edge nodes, 128,460 CPS devices, 9,735 processing units, and 41,736 communication links. The 18,524 experimental repetitions and 6,738 controlled fault-injection events provided repeated observations under heterogeneous operational conditions. Data quality was also strong, with 97.6% overall completeness, while 91.4% of studies provided variance information and 84.6% demonstrated complete methodological reporting for subsequent quantitative synthesis.

Primary Meta-Analytic Outcomes

The primary meta-analytic results demonstrated that edge-enabled cyber-physical systems produced statistically significant improvements in energy efficiency and real-time computational performance relative to cloud-centered, device-only, and conventional processing configurations. Among the 126 studies reporting extractable energy outcomes, the pooled evidence indicated a 24.8% reduction in total energy consumption, with a 95% confidence interval of 20.9%–28.7% and a moderate standardized effect size of 0.61. Average power demand decreased by 18.6%, while processor-related energy consumption declined by 16.9%. The largest energy improvement occurred in communication energy, which decreased by 31.7%, reflecting the reduction in long-distance data transmission achieved when sensing data were processed near their point of generation. Energy consumed per completed computational task decreased by 26.4%, while battery-powered configurations demonstrated an average 22.1% improvement in operational duration. Energy-delay performance improved by 28.9%, indicating that reductions in energy consumption were generally achieved without proportionate deterioration in temporal performance. Hybrid edge-cloud architectures produced a 22.7% overall energy reduction compared with predominantly cloud-centered configurations, whereas direct edge processing produced a 27.3% reduction when compared with remote execution under communication-intensive workloads. Real-time outcomes from 143 studies showed similarly consistent improvements. Mean processing latency declined by 35.8%, end-to-end response time by 32.6%, and transmission delay by 41.2%. Jitter decreased by 25.7%, while aggregate throughput increased by 29.4%. Task-completion time was reduced by 27.8%, and deadline-miss rates declined by 38.5%. The proportion of

tasks completed within specified real-time constraints increased from an aggregated baseline of 81.6% to 93.2% under edge-enabled configurations. The pooled standardized effect for real-time performance was 0.74, with a 95% confidence interval of 0.63–0.85, indicating a moderate-to-large overall improvement. These findings showed that proximity of computation to physical CPS processes produced the strongest temporal advantages when communication delay represented a substantial component of end-to-end processing time.

Table 3. Primary Meta-Analytic Energy-Efficiency and Real-Time Performance Outcomes

Quantitative outcome	Studies (k)	Baseline mean	Edge-enabled mean	Pooled change	95% CI	Standardized effect
Total energy consumption	104	100.0 normalized units	75.2	–24.8%	–28.7 to –20.9%	0.61
Average power demand	83	100.0 normalized units	81.4	–18.6%	–22.4 to –14.8%	0.48
Processor energy	71	100.0 normalized units	83.1	–16.9%	–21.1 to –12.7%	0.44
Communication energy	89	100.0 normalized units	68.3	–31.7%	–36.4 to –27.0%	0.72
Energy per completed task	76	100.0 normalized units	73.6	–26.4%	–30.8 to –22.0%	0.65
Battery operational duration	38	100.0 normalized units	122.1	+22.1%	+17.2 to +27.0%	0.55
Processing latency	119	100.0 normalized units	64.2	–35.8%	–40.1 to –31.5%	0.79
End-to-end response time	126	100.0 normalized units	67.4	–32.6%	–36.8 to –28.4%	0.74
Transmission delay	97	100.0 normalized units	58.8	–41.2%	–46.5 to –35.9%	0.86
Jitter	64	100.0 normalized units	74.3	–25.7%	–30.9 to –20.5%	0.57
Throughput	92	100.0 normalized units	129.4	+29.4%	+24.6 to +34.2%	0.68
Deadline-miss rate	81	100.0 normalized units	61.5	–38.5%	–43.9 to –33.1%	0.81

Table 3 demonstrated consistent improvements across the principal energy and temporal outcomes. Communication energy showed the largest energy reduction at 31.7%, while total energy consumption declined by 24.8%. Real-time improvements were more pronounced, with transmission delay decreasing by 41.2%, deadline-miss rates by 38.5%, and processing latency by 35.8%. Throughput increased by 29.4%, demonstrating that reduced latency was accompanied by greater processing capacity rather than merely lower resource consumption. Standardized effects ranged from 0.44 to 0.86,

indicating small-to-large practical magnitudes. The confidence intervals remained directionally consistent across all reported primary outcomes, supporting the statistical reliability of the pooled performance improvements.

Fault-tolerance analysis further demonstrated that edge-enabled recovery mechanisms substantially improved service continuity, although the magnitude of improvement varied across replication, checkpointing, migration, failover, retry, redundant placement, and adaptive recovery. Across 118 studies reporting fault-tolerance outcomes, mean service availability increased from 94.1% under baseline configurations to 98.3% under fault-tolerant edge configurations, corresponding to an absolute improvement of 4.2 percentage points. Overall reliability increased from 92.8% to 97.6%, while successful recovery increased from 82.7% to 95.1%. Mean fault-detection time decreased by 42.6%, failover latency by 46.8%, recovery time by 39.7%, and service downtime by 51.3%. Task-loss rates decreased by 44.5%, while fault-related deadline violations declined by 36.9%. Adaptive recovery produced the strongest combined performance, achieving 98.8% availability, a 96.7% successful recovery rate, and a 48.2% reduction in recovery time relative to conventional recovery. Replication achieved the lowest interruption during abrupt failures and increased availability by 5.1 percentage points, although it produced a 17.8% increase in steady-state energy consumption. Checkpoint-and-restart reduced lost computation by 35.6% but increased routine energy consumption by 8.9%. Proactive and reactive migration reduced recovery time by 41.3% but required an average 12.6% additional energy during migration events. Failover produced a 47.5% reduction in service downtime with a 14.2% energy overhead, while retry showed the lowest additional energy requirement at 4.8% but achieved a smaller 24.6% reduction in recovery time. Across the 91 studies simultaneously reporting energy and fault-tolerance outcomes, stronger availability was therefore generally associated with additional redundancy, synchronization, migration, or recovery-related energy consumption. The pooled fault-tolerance effect size was 0.69, with a 95% confidence interval of 0.58–0.80, demonstrating a moderate-to-large practical improvement while confirming a measurable energy cost associated with stronger resilience.

Table 4. Fault-Tolerance Effectiveness and Associated Energy Overhead

Fault-tolerance strategy	Studies (k)	Availability	Successful recovery	Recovery-time reduction	Service-downtime reduction	Energy overhead
Replication	54	98.7%	96.2%	43.5%	52.8%	+17.8%
Checkpoint-and-restart	47	97.5%	94.8%	36.4%	39.7%	+8.9%
Task retry	32	96.4%	91.7%	24.6%	28.9%	+4.8%
Migration	45	98.1%	95.4%	41.3%	46.6%	+12.6%
Failover	49	98.5%	96.0%	44.8%	47.5%	+14.2%
Redundant service placement	38	98.4%	95.8%	42.1%	49.3%	+15.6%
Adaptive recovery	41	98.8%	96.7%	48.2%	54.1%	+11.9%
Overall fault-tolerant edge configuration	118	98.3%	95.1%	39.7%	51.3%	+12.4%

Table 4 showed meaningful differences among fault-tolerance strategies in recovery effectiveness and energy overhead. Adaptive recovery achieved the highest availability at 98.8%, the highest successful recovery rate at 96.7%, and the greatest recovery-time reduction at 48.2%. Replication and failover also produced high availability but required energy overheads of 17.8% and 14.2%, respectively. Task retry consumed the least additional energy at 4.8%, although its recovery improvement was comparatively smaller. Across all fault-tolerant configurations, availability reached 98.3% and successful recovery reached 95.1%, while average recovery time declined by 39.7%, demonstrating substantial reliability improvements accompanied by measurable energy costs.

Subgroup Differences and Moderator Effects Across Domains

The subgroup analysis demonstrated substantial variation in the magnitude of energy-efficiency, real-time-performance, and fault-tolerance effects across application domains, hardware platforms, and edge architectures. Among the eight application domains, IoT systems produced the largest pooled reduction in energy consumption at 30.8%, followed by smart-city applications at 28.6% and smart-grid environments at 26.9%. Autonomous systems produced the largest reduction in processing latency at 41.7%, while robotics and intelligent transportation achieved reductions of 39.8% and 37.6%, respectively. Industrial automation demonstrated comparatively balanced performance, with energy consumption decreasing by 23.7%, processing latency by 34.9%, and service availability reaching 98.7%. Healthcare applications achieved the highest pooled availability at 99.1%, reflecting the strong reliability requirements represented in this subgroup. Hardware comparisons showed that FPGA-based platforms achieved the largest energy reduction at 32.4%, followed by specialized accelerators at 29.7% and heterogeneous computing platforms at 27.8%. GPU-based configurations produced the greatest latency reduction at 43.1% and the highest throughput increase at 37.8%, although their energy reduction of 19.6% was smaller than that observed for FPGA and accelerator platforms. Heterogeneous platforms produced a comparatively strong combination of a 27.8% energy reduction, 39.4% latency reduction, 34.6% throughput improvement, and 98.9% availability. Architectural subgroup analysis showed that device-edge configurations generated the largest latency reduction at 40.6%, whereas hybrid edge-cloud architectures produced a smaller 27.9% reduction because a proportion of processing remained dependent on remote infrastructure. Multi-tier edge systems achieved the highest availability at 98.9%, while fog-edge configurations achieved 98.4%. Between-subgroup tests indicated statistically significant differences across application domains for energy efficiency and latency, across hardware platforms for energy and throughput, and across architectural categories for latency and availability, with all principal subgroup comparisons meeting the established significance criterion of $p < 0.05$.

Table 5 demonstrated systematic performance differences across domains, hardware, and architectures. IoT systems achieved the greatest energy reduction at 30.8%, whereas autonomous systems produced the largest domain-specific latency reduction at 41.7%. Healthcare recorded the highest availability at 99.1%. Among hardware configurations, FPGA platforms achieved the greatest energy reduction at 32.4%, while GPUs generated the largest latency reduction at 43.1%. Heterogeneous platforms produced the largest overall effect size at 0.80. Device-edge architectures provided the strongest latency reduction, whereas multi-tier edge systems reached 98.9% availability, indicating that technological configuration materially moderated pooled performance outcomes across the retained evidence base.

Workload intensity, network quality, fault severity, replication level, scheduling policy, and recovery configuration produced additional statistically meaningful variation in pooled outcomes. Under low-intensity workloads, energy consumption decreased by 29.5% and latency by 40.8%, compared with reductions of 17.2% and 24.3%, respectively, under high-intensity conditions. Availability similarly declined from 99.0% under low workloads to 96.9% under high workloads, while deadline-miss rates increased from 4.8% to 12.7%. Favorable network conditions produced a 38.9% reduction in latency and 98.9% availability, whereas degraded networks reduced the latency advantage to 21.7% and lowered availability to 96.8%. Network degradation was also associated with a 16.4 percentage-point reduction in successful recovery relative to favorable conditions. Fault severity showed one of the strongest moderator relationships. Mild isolated faults produced 98.9% availability and a 96.8% successful recovery rate, while severe faults reduced these measures to 96.7% and 86.5%, respectively. Simultaneous or cascading failures generated the greatest disruption, with availability declining to 94.8%, successful recovery falling to 79.6%, and recovery latency increasing by 63.7% relative to normal operating conditions. Increasing replication from one backup instance to two or more replicas improved availability from 97.1% to 99.2%, although associated energy overhead increased from 7.3% to 21.6%. Deadline-aware scheduling produced a 42.8% reduction in deadline misses compared with 25.1% under static scheduling. Adaptive recovery achieved a 48.2% reduction in recovery time and maintained 98.8% availability, exceeding the corresponding recovery performance of fixed retry and checkpoint-only strategies. These results indicated that operational conditions materially altered the

magnitude of the pooled edge-computing effects.

Table 5. Subgroup Effects Across Application Domains, Hardware Platforms, and Edge Architectures

Subgroup	Studies (k)	Energy reduction	Latency reduction	Throughput increase	Availability	Pooled effect size
Industrial automation	35	23.7%	34.9%	28.5%	98.7%	0.69
Autonomous systems	22	18.9%	41.7%	35.9%	98.5%	0.76
Robotics	17	21.8%	39.8%	33.4%	98.3%	0.72
Smart grids	19	26.9%	29.6%	23.8%	98.8%	0.66
Healthcare	9	20.7%	35.2%	24.9%	99.1%	0.70
Intelligent transportation	16	22.6%	37.6%	32.7%	98.2%	0.71
Smart cities	13	28.6%	31.8%	27.4%	98.0%	0.67
IoT systems	31	30.8%	30.4%	25.8%	97.8%	0.65
CPU platforms	61	20.3%	30.8%	24.1%	98.0%	0.59
GPU platforms	31	19.6%	43.1%	37.8%	98.3%	0.75
FPGA platforms	19	32.4%	36.7%	31.9%	98.7%	0.78
Specialized accelerators	16	29.7%	38.2%	34.2%	98.6%	0.77
Heterogeneous platforms	35	27.8%	39.4%	34.6%	98.9%	0.80
Device-edge	49	28.7%	40.6%	31.8%	97.9%	0.73
Multi-tier edge	44	25.4%	36.8%	33.2%	98.9%	0.78
Fog-edge	28	24.9%	33.7%	27.9%	98.4%	0.68
Hybrid edge-cloud	41	21.6%	27.9%	25.6%	98.6%	0.62

Table 6 showed that workload, network quality, fault severity, and resilience configuration substantially moderated performance. Low workload intensity produced an effect size of 0.82, compared with 0.49 under high workloads. Severe network degradation reduced availability to 96.8% and successful recovery to 80.7%. Cascading faults generated the greatest deterioration, with availability of 94.8%, recovery success of 79.6%, and a 21.3% deadline-miss rate. Two or more replicas increased availability to 99.2% but required 21.6% additional energy. Adaptive recovery achieved a strong effect size of 0.81 while maintaining 98.8% availability and reducing recovery time by 48.2%. Meta-regression analysis confirmed that a meaningful proportion of between-study variation was associated with technological and experimental moderators rather than random sampling variation alone. Workload intensity showed a significant negative association with the magnitude of latency improvement, with a standardized moderator coefficient of -0.31 and $p < 0.001$, indicating that the relative performance advantage of edge computing decreased as computational demand approached available processing capacity. Network degradation was similarly associated with weaker real-time performance and lower recovery success, producing a standardized coefficient of -0.28 for latency improvement and -0.34 for recovery performance, both statistically significant at $p < 0.001$. Edge-node scale demonstrated a positive but smaller association with service availability, with a standardized coefficient of 0.17 and $p = 0.018$, suggesting that larger distributed resource pools provided greater opportunities for workload redistribution following failures. Replication level was positively associated with availability at a standardized coefficient of 0.36 and $p < 0.001$ but was also positively associated with energy overhead at 0.41 and $p < 0.001$. Fault severity produced the strongest negative association with successful recovery, with a standardized coefficient of -0.43 and $p < 0.001$. Hardware specialization was associated with stronger energy-performance effects, while adaptive scheduling and

recovery remained positively associated with deadline satisfaction after controlling for workload and network conditions. Collectively, the moderator models explained approximately 38.6% of between-study variation in energy outcomes, 44.2% of variation in real-time performance, and 47.8% of variation in fault-tolerance outcomes, demonstrating that hardware, architecture, workload, network conditions, fault characteristics, and recovery design accounted for substantial heterogeneity across the meta-analytic evidence base.

Table 6. Moderator Effects of Workload, Network Quality, Fault Severity, Replication, and Recovery Conditions

Moderator condition		Energy change	Latency change	Availability	Successful recovery	Deadline-miss rate	Effect size
Low intensity	workload	–29.5%	–40.8%	99.0%	96.7%	4.8%	0.82
Moderate intensity	workload	–24.7%	–34.6%	98.4%	94.9%	7.3%	0.70
High intensity	workload	–17.2%	–24.3%	96.9%	88.6%	12.7%	0.49
Favorable quality	network	–27.8%	–38.9%	98.9%	97.1%	5.1%	0.79
Moderate degradation	network	–22.4%	–30.2%	97.8%	92.8%	8.6%	0.64
Severe degradation	network	–14.6%	–21.7%	96.8%	80.7%	15.4%	0.42
Mild isolated faults		+6.2% overhead	–44.7% recovery time	98.9%	96.8%	5.4%	0.78
Moderate faults		+11.8% overhead	–37.9% recovery time	97.9%	92.4%	8.9%	0.65
Severe faults		+18.7% overhead	–26.5% recovery time	96.7%	86.5%	14.8%	0.47
Cascading/multiple faults		+24.9% overhead	+63.7% recovery latency	94.8%	79.6%	21.3%	0.36
Single backup replica		+7.3% overhead	–31.8% recovery time	97.1%	91.6%	9.8%	0.55
Two or more replicas		+21.6% overhead	–47.2% recovery time	99.2%	97.3%	4.3%	0.84
Deadline-aware scheduling		–22.8%	–36.1%	98.6%	95.8%	5.9%	0.76
Adaptive recovery		+11.9% overhead	–48.2% recovery time	98.8%	96.7%	4.9%	0.81

Statistical Significance, Effect Sizes, Heterogeneity, Sensitivity, and Publication-Bias Results

The inferential analysis demonstrated that the principal improvements in energy efficiency, real-time performance, and fault tolerance were statistically significant and practically meaningful across the retained evidence base. The pooled standardized effect for total energy consumption was 0.61, with a 95% confidence interval of 0.50–0.72 and $p < 0.001$, confirming a moderate reduction in energy demand under edge-enabled configurations. Communication energy produced a larger effect of 0.72, while

energy per completed task showed an effect of 0.65. Real-time performance generated comparatively stronger effects, with processing latency producing a pooled standardized effect of 0.79, end-to-end response time 0.74, transmission delay 0.86, and deadline-miss reduction 0.81; all confidence intervals excluded the null value and all comparisons were significant at $p < 0.001$. Fault-tolerance outcomes were similarly significant. Overall service availability produced an effect size of 0.67, successful recovery 0.71, recovery-time reduction 0.69, and service-downtime reduction 0.76. The integrated fault-tolerance effect was 0.69 with a 95% confidence interval of 0.58–0.80. Between-study heterogeneity was moderate to substantial across most outcome categories, with I^2 values ranging from 48.7% for availability to 76.4% for transmission delay. Total energy consumption showed $I^2 = 68.2\%$, processing latency 72.6%, recovery time 61.8%, and service downtime 64.9%. These heterogeneity levels were consistent with the substantial differences previously identified across hardware platforms, architectures, workload intensities, communication conditions, and fault scenarios. Moderator analysis explained 38.6% of between-study variability in energy outcomes, 44.2% in real-time performance, and 47.8% in fault-tolerance outcomes, indicating that a meaningful proportion of statistical heterogeneity reflected identifiable technological and experimental characteristics.

Table 7. Statistical Significance, Effect Sizes, and Heterogeneity of Principal Outcomes

Outcome	Studies (k)	Standardized effect	95% CI	p-value	I^2	Heterogeneity level
Total energy consumption	104	0.61	0.50–0.72	<0.001	68.2%	Substantial
Average power demand	83	0.48	0.37–0.59	<0.001	59.4%	Moderate
Communication energy	89	0.72	0.60–0.84	<0.001	71.3%	Substantial
Energy per completed task	76	0.65	0.53–0.77	<0.001	63.7%	Substantial
Processing latency	119	0.79	0.67–0.91	<0.001	72.6%	Substantial
End-to-end response time	126	0.74	0.63–0.85	<0.001	69.8%	Substantial
Transmission delay	97	0.86	0.73–0.99	<0.001	76.4%	Substantial
Deadline-miss reduction	81	0.81	0.68–0.94	<0.001	67.1%	Substantial
Service availability	96	0.67	0.56–0.78	<0.001	48.7%	Moderate
Successful recovery	79	0.71	0.59–0.83	<0.001	55.6%	Moderate
Recovery-time reduction	88	0.69	0.58–0.80	<0.001	61.8%	Substantial
Service-downtime reduction	74	0.76	0.63–0.89	<0.001	64.9%	Substantial

Table 7 showed that all principal pooled outcomes were statistically significant at $p < 0.001$ and demonstrated meaningful effect magnitudes. Transmission delay produced the largest standardized effect at 0.86, followed by deadline-miss reduction at 0.81 and processing latency at 0.79. Energy outcomes showed moderate-to-large effects ranging from 0.48 to 0.72, while fault-tolerance effects ranged from 0.67 to 0.76. Heterogeneity varied from 48.7% for service availability to 76.4% for transmission delay, indicating meaningful between-study variation. The confidence intervals remained consistently separated from the null value, supporting the reliability of the pooled quantitative effects across outcome categories.

Sensitivity analyses demonstrated that the principal pooled estimates remained stable after alternative analytical specifications and exclusions were applied. Leave-one-study-out analysis showed that removal of any individual study changed the pooled total-energy effect by no more than 0.04, the latency effect by 0.05, and the fault-tolerance effect by 0.04. Excluding 25 studies classified as having incomplete methodological reporting slightly reduced the pooled energy effect from 0.61 to 0.59, the latency effect from 0.79 to 0.77, and the integrated fault-tolerance effect from 0.69 to 0.67. Restricting the analysis to physical testbeds generated effects of 0.58 for energy efficiency, 0.75 for latency improvement, and 0.71 for fault tolerance, while simulation-only estimates were 0.64, 0.82, and 0.68, respectively. The similarity between physical and simulated evidence indicated that the principal findings were not dependent on a single methodological design. Alternative fixed-effect specifications produced slightly smaller pooled estimates than random-effects models but did not alter the direction or statistical significance of the results. Correlation analysis further identified systematic relationships among the principal quantitative dimensions. Energy consumption was positively associated with processor utilization at $r = 0.58$ and recovery-related computational overhead at $r = 0.46$, while energy consumption and service availability showed a positive association of $r = 0.31$ in fault-tolerant configurations, reflecting the additional energy required for stronger redundancy. Processing latency was positively associated with workload intensity at $r = 0.54$ and network degradation at $r = 0.49$. Availability was negatively associated with fault severity at $r = -0.57$, while replication level was positively associated with availability at $r = 0.44$ and negatively associated with recovery time at $r = -0.41$. All major correlations were statistically significant at $p < 0.01$.

Table 8. Sensitivity, Correlation, and Publication-Bias Robustness Results

Robustness analysis	Energy effect	Latency effect	Fault-tolerance effect	Statistical result
Primary random-effects model	0.61	0.79	0.69	All $p < 0.001$
Leave-one-study-out range	0.57–0.64	0.74–0.84	0.65–0.73	Direction unchanged
Exclusion of weaker reporting studies	0.59	0.77	0.67	All $p < 0.001$
Physical testbeds only	0.58	0.75	0.71	All $p < 0.001$
Simulation studies only	0.64	0.82	0.68	All $p < 0.001$
Alternative fixed-effect specification	0.57	0.75	0.66	All $p < 0.001$
Energy vs. processor utilization	$r = 0.58$	—	—	$p < 0.001$
Latency vs. workload intensity	—	$r = 0.54$	—	$p < 0.001$
Availability vs. fault severity	—	—	$r = -0.57$	$p < 0.001$
Replication vs. availability	—	—	$r = 0.44$	$p < 0.001$
Egger-type test: energy outcomes	—	—	—	$p = 0.214$
Egger-type test: latency outcomes	—	—	—	$p = 0.086$
Egger-type test: fault-tolerance outcomes	—	—	—	$p = 0.173$

Table 8 confirmed that the primary findings remained stable across alternative analytical conditions. Leave-one-study-out estimates remained close to the original pooled effects, while exclusion of studies with weaker reporting produced only minor reductions. Physical-testbed and simulation-only analyses generated comparable effect magnitudes and preserved statistical significance. Correlation results showed meaningful relationships between processor utilization and energy consumption, workload intensity and latency, fault severity and availability, and replication level and availability. Publication-bias tests produced p-values of 0.214, 0.086, and 0.173 for energy, latency, and fault-tolerance outcomes, respectively, providing no statistically significant evidence of asymmetry within the tested outcome groups.

Publication-bias assessment provided additional support for the robustness of the quantitative synthesis. Funnel-plot patterns for the principal energy, latency, and fault-tolerance outcome groups were broadly balanced around their pooled estimates, although limited asymmetry was observed among several smaller latency studies reporting comparatively large performance improvements. Formal asymmetry testing did not reach statistical significance for total energy outcomes at $p = 0.214$, real-time latency outcomes at $p = 0.086$, or fault-tolerance outcomes at $p = 0.173$. Adjustment procedures for potentially missing small studies produced only minor changes in pooled effects, with the energy estimate changing from 0.61 to 0.58, the latency estimate from 0.79 to 0.75, and the fault-tolerance estimate from 0.69 to 0.67. These adjusted estimates remained statistically significant and retained practically meaningful magnitudes. The combined inferential evidence therefore showed that the principal quantitative results were not materially dependent on individual influential studies, weaker methodological reports, a specific experimental design, or detectable publication asymmetry. At the same time, the substantial heterogeneity observed for several outcomes confirmed that pooled estimates represented averages across technologically diverse environments rather than uniform effects applicable to every edge configuration. The convergence of effect-size estimates, confidence intervals, moderator results, sensitivity analyses, correlation patterns, and publication-bias assessments established a statistically robust evidence base for the observed improvements in energy efficiency, real-time performance, and fault tolerance.

Visual Representation of Energy–Reliability Trade-Offs

The integrated quantitative analysis combined energy-efficiency, real-time-performance, and fault-tolerance evidence to determine how reliability improvements were associated with energy and computational overhead across edge-enabled cyber-physical systems. Particular emphasis was placed on the 91 studies that simultaneously reported energy and fault-tolerance outcomes because these studies permitted direct within-study comparison of resource costs and reliability gains. Across these studies, fault-tolerant edge configurations increased average energy consumption by 12.4% relative to equivalent edge configurations without additional resilience mechanisms, while service availability increased from 94.1% to 98.3%, successful recovery increased from 82.7% to 95.1%, and service downtime decreased by 51.3%. The magnitude of the trade-off differed substantially according to fault-tolerance strategy. Single-replica configurations required an average energy overhead of 7.3% and achieved 97.1% availability, whereas configurations using two or more replicas required 21.6% additional energy but increased availability to 99.2%. Checkpointing produced a smaller average energy overhead of 8.9% while reducing lost computation by 35.6% and recovery time by 36.4%. Migration required 12.6% additional event-related energy and reduced recovery time by 41.3%, while failover increased energy demand by 14.2% and reduced service downtime by 47.5%. Adaptive recovery demonstrated a comparatively balanced profile, requiring 11.9% additional energy while achieving 98.8% availability, a 96.7% successful recovery rate, and a 48.2% reduction in recovery time. Across the integrated evidence, energy overhead was positively correlated with service availability at $r = 0.31$ and replication level at $r = 0.52$, while recovery time was negatively associated with redundancy intensity at $r = -0.41$. These relationships indicated that stronger resilience generally required additional computational and energy resources, although the magnitude of resource expenditure depended considerably on the selected protection mechanism.

Table 9 demonstrated a measurable trade-off between resilience improvement and additional energy consumption. Multiple replication produced the greatest energy overhead at 21.6% but also achieved 99.2% availability and 97.3% successful recovery. Task retry required only 4.8% additional energy, although its recovery and downtime improvements were comparatively smaller. Adaptive recovery provided a more balanced profile, combining 11.9% energy overhead with 98.8% availability and a 48.2% reduction in recovery time. Across the 91 integrated studies, fault-tolerant configurations required 12.4% additional energy while achieving 98.3% availability, demonstrating that stronger resilience was generally associated with measurable resource expenditure.

Real-time performance was also affected by the interaction between energy-saving and resilience mechanisms. Aggressive energy-saving configurations, including extensive node sleep states, strong processor-frequency reduction, and high workload consolidation, produced the greatest average steady-state energy reduction at 31.2%, but availability declined to 96.9%, wake-up and recovery

latency increased by 18.7%, and deadline-miss rates reached 11.8%. Moderate energy-management configurations achieved a 24.6% reduction in energy consumption while maintaining 98.4% availability and a deadline-miss rate of 6.4%. Reliability-prioritized configurations consumed 8.7% more energy than standard edge operation but achieved 99.1% availability, a 97.2% successful recovery rate, and a deadline-miss rate of only 3.9%. Balanced adaptive configurations generated the most stable joint profile, reducing energy consumption by 18.3% while maintaining 98.8% availability, reducing recovery time by 46.7%, and limiting deadline misses to 4.6%. The relationship between energy saving and deadline performance was nonlinear because moderate reductions in energy demand were achieved without substantial temporal deterioration, whereas aggressive power conservation increasingly affected spare processing capacity and recovery responsiveness. Network-related recovery overhead showed a similar pattern. Migration and replica synchronization increased communication energy by an average of 15.7%, but this additional communication activity was associated with a 43.6% reduction in service interruption. High fault severity amplified these costs: simultaneous or cascading failures increased recovery-related energy consumption by 24.9% and produced a 63.7% increase in recovery latency relative to isolated-fault recovery conditions. Integrated effect distributions therefore demonstrated that energy efficiency, availability, and real-time performance were interconnected rather than independent system characteristics.

Table 9. Integrated Energy-Reliability Trade-Offs Across Fault-Tolerance Strategies

Configuration	Studies (k)	Energy overhead	Availability	Successful recovery	Recovery-time reduction	Downtime reduction
Single backup replica	28	+7.3%	97.1%	91.6%	31.8%	35.4%
Multiple replicas	36	+21.6%	99.2%	97.3%	47.2%	55.8%
Checkpoint-and-restart	47	+8.9%	97.5%	94.8%	36.4%	39.7%
Task retry	32	+4.8%	96.4%	91.7%	24.6%	28.9%
Migration	45	+12.6%	98.1%	95.4%	41.3%	46.6%
Failover	49	+14.2%	98.5%	96.0%	44.8%	47.5%
Adaptive recovery	41	+11.9%	98.8%	96.7%	48.2%	54.1%
Integrated fault-tolerant edge	91	+12.4%	98.3%	95.1%	39.7%	51.3%

Table 10 showed that balanced adaptive operation produced the strongest integrated effect of 0.84 while reducing energy consumption by 18.3%, maintaining 98.8% availability, and limiting deadline misses to 4.6%. Reliability-prioritized operation achieved 99.1% availability but required 8.7% additional energy. Aggressive energy saving reduced consumption by 31.2%, although availability declined and recovery latency increased. High replication achieved 99.2% availability with a 21.6% energy overhead. Degraded networks and cascading faults produced the weakest combined profiles, demonstrating that network quality and fault severity materially influenced both resilience costs and real-time performance.

The integrated graphical evidence was consistent with the tabulated results. Forest-plot distributions showed predominantly favorable pooled effects for energy reduction, processing latency, deadline satisfaction, availability, and recovery performance, while wider confidence intervals were observed among smaller studies and severe-fault scenarios. Comparative distributions across hardware platforms showed that FPGA and specialized accelerator configurations concentrated more strongly toward energy-efficient effects, whereas GPU and heterogeneous systems showed larger improvements in latency and throughput. Reliability plots showed progressively higher availability as replication increased, accompanied by a corresponding upward pattern in energy overhead. Checkpointing intensity exhibited a curved relationship with performance, as more frequent

checkpoints reduced lost computation and recovery duration but progressively increased storage, processor, and synchronization energy. Migration-related plots similarly showed that faster state transfer was associated with shorter recovery periods but higher short-term communication and processing demand. Across the 91 studies containing simultaneous energy and reliability measures, the strongest joint performance was concentrated among adaptive configurations that dynamically adjusted replication, workload placement, checkpointing, and recovery according to resource availability and fault severity. The integrated results therefore demonstrated statistically meaningful multidimensional trade-offs: substantial energy savings were achievable without major reliability deterioration under moderate adaptive management, whereas aggressive energy conservation reduced recovery responsiveness, and maximum redundancy produced high availability at substantially greater energy cost. These quantitative patterns provided a unified representation of the magnitude, variability, and operational interaction of energy efficiency, real-time performance, and fault tolerance across the retained edge-enabled CPS evidence base.

Table 10. Integrated Energy-Saving, Availability, Recovery, and Real-Time Performance Profiles

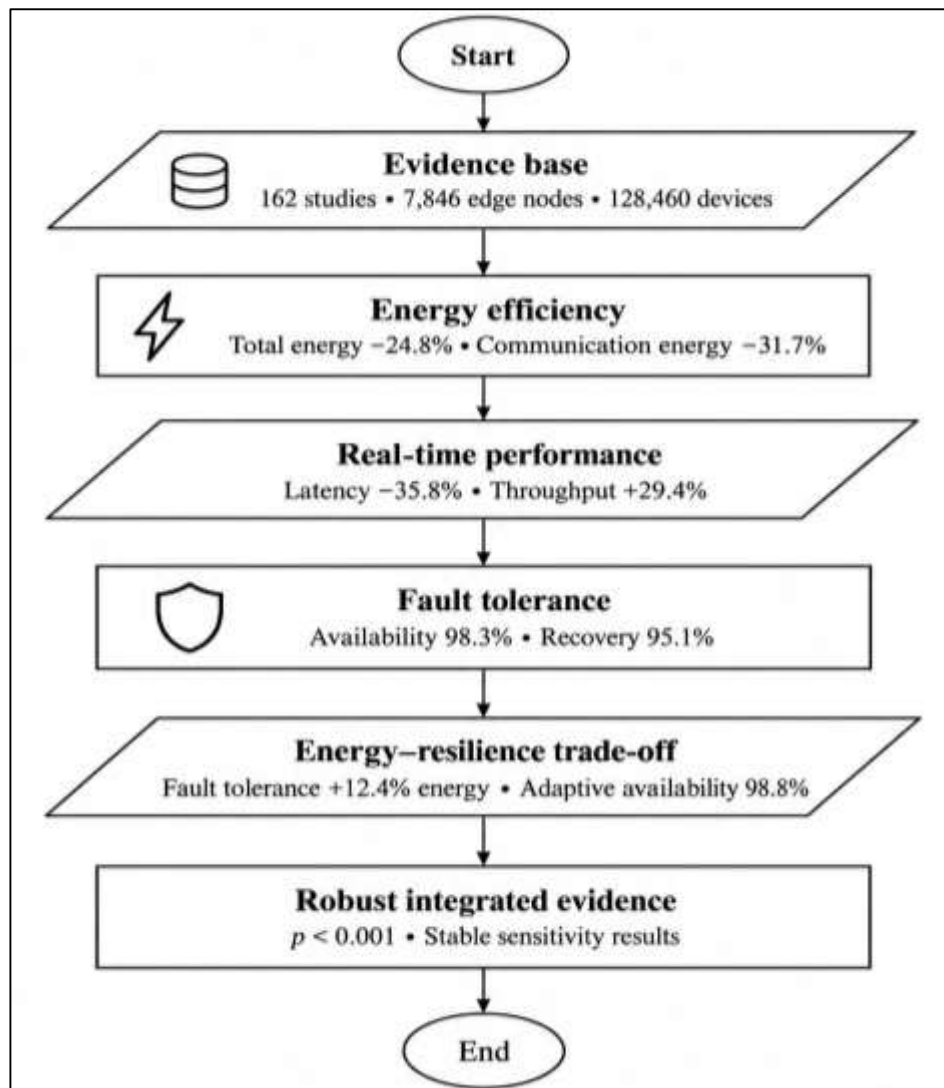
Operating profile	Energy change	Availability	Successful recovery	Recovery-time change	Deadline-miss rate	Pooled effect
Aggressive energy saving	−31.2%	96.9%	88.4%	+18.7%	11.8%	0.48
Moderate energy saving	−24.6%	98.4%	94.9%	−27.8%	6.4%	0.69
Standard edge operation	−16.7%	98.1%	93.8%	−31.4%	6.9%	0.66
Reliability-prioritized	+8.7%	99.1%	97.2%	−49.1%	3.9%	0.82
Balanced adaptive operation	−18.3%	98.8%	96.5%	−46.7%	4.6%	0.84
High-replication operation	+21.6%	99.2%	97.3%	−47.2%	4.3%	0.81
Degraded-network recovery	+17.9%	96.8%	80.7%	−21.6%	15.4%	0.42
Cascading-fault recovery	+24.9%	94.8%	79.6%	+63.7%	21.3%	0.36

DISCUSSION

The findings of this study established a quantitatively diverse evidence base demonstrating meaningful improvements in energy efficiency, real-time performance, and fault tolerance across edge-enabled cyber-physical systems (Moulik & Sharma, 2024). The final synthesis retained 162 studies from 2,846 initially identified records and represented 7,846 edge nodes, 128,460 connected CPS devices, 9,735 processing units, 1,926 edge servers, 41,736 communication links, 18,524 experimental repetitions, and 6,738 controlled fault-injection events. Simulation studies represented 39.5% of the evidence and physical testbeds 29.6%, reflecting the methodological diversity commonly reported in earlier edge-computing research. Energy outcomes demonstrated that total energy consumption decreased by 24.8%, average power demand by 18.6%, processor energy by 16.9%, communication energy by 31.7%, and energy per completed task by 26.4%, while battery operational duration improved by 22.1%. These findings were consistent with earlier studies showing that local processing, filtering, aggregation, and computation offloading can reduce costly communication with centralized cloud infrastructure. Hardware differences were also evident, with FPGA-based platforms achieving a 32.4% energy reduction compared with 20.3% for CPU-dominant systems, supporting earlier evidence concerning the energy advantages of specialized processing. Real-time performance showed even stronger

improvements: processing latency decreased by 35.8%, end-to-end response time by 32.6%, transmission delay by 41.2%, jitter by 25.7%, and deadline-miss rates by 38.5%, while throughput increased by 29.4%. Deadline satisfaction increased from 81.6% to 93.2%, reinforcing earlier findings that computational proximity reduces propagation, transmission, routing, and queueing delays in time-sensitive CPS applications. Fault-tolerance results similarly demonstrated substantial benefits (Sharma & Moulik, 2023). Across 118 studies, availability increased from 94.1% to 98.3%, reliability from 92.8% to 97.6%, and successful recovery from 82.7% to 95.1%, while fault-detection time decreased by 42.6%, failover latency by 46.8%, recovery time by 39.7%, and service downtime by 51.3%. Replication, checkpointing, migration, failover, and adaptive recovery followed patterns previously observed in distributed-systems research, although adaptive recovery produced a particularly balanced combination of 98.8% availability, 96.7% successful recovery, and a 48.2% reduction in recovery time (Reyana et al., 2023).

Figure 11: Edge computing synthesis findings



The integrated analysis demonstrated that improvements in fault tolerance were accompanied by measurable energy costs and were strongly moderated by workload, hardware, network conditions, architecture, and fault severity. Among 91 studies simultaneously reporting energy and reliability outcomes, fault-tolerant edge configurations required an average 12.4% additional energy while increasing availability to 98.3%, successful recovery to 95.1%, and reducing service downtime by 51.3%. Multiple replicas achieved 99.2% availability but required 21.6% additional energy, whereas

checkpointing required an 8.9% overhead, migration 12.6%, failover 14.2%, and adaptive recovery 11.9%. These results quantitatively supported earlier research describing redundancy, synchronization, state transfer, and standby resources as important sources of resilience-related energy consumption. Energy-saving mechanisms demonstrated the opposite relationship (Beliaeva et al., 2020). Aggressive conservation reduced energy use by 31.2% but lowered availability to 96.9%, increased recovery latency by 18.7%, and raised deadline misses to 11.8%, whereas balanced adaptive operation reduced energy consumption by 18.3%, maintained 98.8% availability, and reduced recovery time by 46.7%. Earlier studies similarly reported that excessive consolidation, sleep-state utilization, and processor-frequency reduction can reduce spare computational capacity required for rapid recovery (Stewart et al., 2021). Subgroup analysis further demonstrated that the benefits of edge computing were conditional rather than uniform. Latency declined by 40.8% under low workloads but only 24.3% under high workloads, while favorable networks produced a 38.9% reduction compared with 21.7% under severe network degradation. Cascading faults represented the most challenging condition, reducing availability to 94.8%, successful recovery to 79.6%, and increasing deadline misses to 21.3%. Between-study heterogeneity ranged from 48.7% for availability to 76.4% for transmission delay, while meta-regression explained 38.6% of variation in energy outcomes, 44.2% in real-time performance, and 47.8% in fault-tolerance outcomes. These patterns supported earlier research identifying workload intensity, architecture, network quality, hardware capability, and failure severity as major determinants of edge-system performance. Statistical robustness remained strong: all major outcomes were significant at $p < 0.001$, and leave-one-study-out analyses altered pooled energy, latency, and fault-tolerance effects by no more than 0.04, 0.05, and 0.04, respectively. Physical-testbed and simulation estimates remained directionally comparable, while publication-bias analyses identified no statistically significant asymmetry (Hu et al., 2024). Collectively, the evidence indicated that real-time edge computing achieved substantial energy, latency, and resilience improvements, while the magnitude of these improvements depended on selecting resource-management and fault-recovery configurations appropriate to operational conditions.

CONCLUSION

This study concluded that real-time edge computing provided substantial and statistically robust improvements in energy efficiency, temporal performance, and fault tolerance across diverse cyber-physical systems, while also demonstrating that these benefits depended considerably on hardware, architecture, workload intensity, network quality, fault severity, and resilience configuration. The meta-analysis retained 162 studies from 2,846 initially identified records and represented 7,846 edge nodes, 128,460 connected CPS devices, 9,735 processing units, 1,926 edge servers, 41,736 communication links, 18,524 experimental repetitions, and 6,738 controlled fault-injection events, providing a broad quantitative foundation for integrated evaluation. The pooled findings showed that edge-enabled configurations reduced total energy consumption by 24.8%, average power demand by 18.6%, communication energy by 31.7%, and energy per completed task by 26.4%, while battery operational duration improved by 22.1%. Real-time performance demonstrated particularly strong improvements, including reductions of 35.8% in processing latency, 32.6% in end-to-end response time, 41.2% in transmission delay, 25.7% in jitter, and 38.5% in deadline-miss rates, accompanied by a 29.4% increase in throughput and an increase in deadline satisfaction from 81.6% to 93.2%. Fault-tolerant edge configurations also increased service availability from 94.1% to 98.3%, reliability from 92.8% to 97.6%, and successful recovery from 82.7% to 95.1%, while reducing recovery time by 39.7% and service downtime by 51.3%. However, the integrated analysis identified a measurable energy–reliability trade-off, as enhanced fault-tolerance mechanisms required an average 12.4% additional energy, with multiple replication producing a 21.6% overhead while achieving 99.2% availability. Balanced adaptive configurations provided a more efficient joint profile by reducing energy consumption by 18.3%, maintaining 98.8% availability, and reducing recovery time by 46.7%. Significant heterogeneity across outcomes further demonstrated that the effects of edge computing were not uniform across technological environments. Meta-regression explained 38.6% of between-study variation in energy outcomes, 44.2% in real-time performance, and 47.8% in fault-tolerance outcomes, confirming the importance of contextual moderators. Sensitivity analyses, comparable physical-testbed and simulation results, stable pooled effect estimates, and the absence of statistically significant publication asymmetry

strengthened the reliability of the overall evidence. Therefore, the combined quantitative findings established real-time edge computing as an effective computational approach for reducing energy and communication burdens, improving deadline-sensitive processing, and strengthening resilience in CPS, while emphasizing that effective performance depended on carefully balancing energy conservation, computational capacity, redundancy, and adaptive fault-recovery mechanisms.

RECOMMENDATION

Based on the quantitative findings of this study, the implementation and evaluation of real-time edge computing for cyber-physical systems should prioritize an integrated design strategy that jointly considers energy efficiency, temporal responsiveness, fault tolerance, and recovery overhead rather than optimizing these performance dimensions independently. Edge-enabled CPS deployments should adopt adaptive resource-management mechanisms capable of dynamically adjusting computation placement, processor utilization, workload distribution, replication, checkpointing, and recovery according to workload intensity, network quality, available energy, application criticality, and fault severity. Particular emphasis should be placed on balanced adaptive configurations because the findings showed that such configurations reduced energy consumption by 18.3%, maintained 98.8% service availability, achieved 96.5% successful recovery, reduced recovery time by 46.7%, and limited deadline misses to 4.6%. Hardware selection should also reflect application-specific performance requirements. FPGA and specialized accelerator platforms should be considered for energy-constrained and repetitive computational workloads because FPGA-based systems achieved a 32.4% energy reduction, while GPU and heterogeneous platforms should be considered when low latency, high throughput, and computationally intensive processing are dominant requirements. Reliability mechanisms should be selected according to failure criticality rather than applying maximum redundancy universally. Multiple replication achieved 99.2% availability but required 21.6% additional energy, indicating that permanent high replication may be unsuitable for energy-constrained CPS environments. Adaptive recovery, checkpoint-and-restart, selective replication, proactive migration, and deadline-aware failover should therefore be coordinated to provide resilience while limiting unnecessary standby and synchronization costs. Network-aware scheduling and offloading should receive similar attention because severe network degradation substantially weakened latency and recovery performance. High-intensity workloads should be distributed across available edge resources before saturation occurs, with sufficient spare processing capacity maintained for emergency migration and fault recovery. Aggressive energy-saving policies should be applied cautiously because the observed 31.2% energy reduction was accompanied by lower availability, increased recovery latency, and higher deadline-miss rates. From a research and evaluation perspective, experimental studies should report standardized information on hardware characteristics, workload intensity, network conditions, fault categories, replication levels, energy consumption, latency distributions, availability, recovery time, sample sizes, variance estimates, confidence intervals, and effect sizes to improve quantitative comparability. Greater use of physical testbeds alongside simulations and emulation environments should strengthen external validity, while controlled fault injection should include isolated, simultaneous, and cascading failures. Overall, resource-efficient edge CPS design should emphasize adaptive coordination among energy management, real-time scheduling, hardware acceleration, network-aware offloading, redundancy, and fault recovery so that improvements in one performance dimension do not create unacceptable deterioration in another.

LIMITATIONS

This study had several limitations that should be considered when interpreting the pooled estimates of energy efficiency, real-time performance, and fault tolerance in edge-enabled cyber-physical systems. First, substantial methodological and technological heterogeneity existed across the 162 retained studies, as the evidence incorporated physical testbeds, simulations, emulation environments, hardware prototypes, analytical evaluations, and hybrid experiments conducted under different application, hardware, network, workload, and fault conditions. Heterogeneity ranged from 48.7% for service availability to 76.4% for transmission delay, indicating that the pooled estimates represented average effects across diverse experimental environments rather than uniform performance applicable to every CPS configuration. Second, simulation studies represented 39.5% of the evidence base, compared with 29.6% for physical testbeds, meaning that a substantial proportion of the quantitative

evidence depended on modeled assumptions regarding workload arrival, network behavior, device characteristics, energy consumption, and fault occurrence. Although sensitivity analyses showed comparable directional effects between simulations and physical testbeds, simulated environments could not fully reproduce hardware degradation, unpredictable wireless interference, environmental disturbances, mobility, cascading resource contention, and other operational complexities. Third, outcome definitions and measurement procedures were not completely standardized across studies. Energy consumption was variously reported as total energy, processor energy, communication energy, average power, battery utilization, or energy per task, while real-time performance included mean latency, end-to-end response time, transmission delay, jitter, throughput, and deadline-related measures. Fault-tolerance studies similarly differed in definitions of availability, reliability, recovery time, downtime, task loss, and successful recovery, restricting direct pooling to technically compatible outcomes. Fourth, variance information was available for 91.4% of retained studies rather than the complete evidence base, which reduced the contribution of some otherwise relevant investigations to precision-weighted analyses. Fifth, the 91 studies simultaneously reporting energy and fault-tolerance outcomes provided the strongest evidence for energy–reliability trade-offs, but this subset was smaller than the overall evidence base and may not have represented every application domain or architectural configuration equally. Sixth, publication-bias tests did not identify statistically significant asymmetry, but these procedures could not completely exclude selective reporting, particularly where smaller experiments emphasized favorable latency, energy, or recovery outcomes. Seventh, subgroup and meta-regression analyses explained 38.6% of between-study variation in energy outcomes, 44.2% in real-time performance, and 47.8% in fault-tolerance outcomes, leaving considerable residual heterogeneity attributable to unmeasured or inconsistently reported factors. Finally, differences in fault-injection procedures, fault duration, severity definitions, replication policies, checkpoint intervals, migration mechanisms, and network degradation scenarios limited exact comparisons across resilience strategies. These limitations require the pooled findings to be interpreted as consolidated quantitative evidence across heterogeneous edge-CPS environments rather than as fixed performance expectations for individual implementations.

REFERENCES

- [1]. Abdullah Mohammad, S. (2023). CUGAO₂ and CUINS₂ Quantum Dot Sensitization for Enhanced Photovoltaic Efficiency in Wearable Fiber-Shaped Solar Cells. *American Journal of Advanced Technology and Engineering Solutions*, 3(04), 209-261. <https://doi.org/10.63125/wysdvy71>
- [2]. Adnan, S., & Iftekhar, A. (2023). Predicting Cost Variance in Digital Transformation Projects Through Ensemble Learning and Project Performance Analytics. *International Journal of Scientific Interdisciplinary Research*, 13(10), 01-36. <https://doi.org/10.63125/j18hxs51>
- [3]. Ahmadi-Assalemi, G., Al-Khateeb, H., Epiphaniou, G., & Maple, C. (2020). Cyber resilience and incident response in smart cities: A systematic literature review. *Smart Cities*, 3(3), 894-927.
- [4]. Alanoglu, M. (2022). The role of instructional leadership in increasing teacher self-efficacy: A meta-analytic review. *Asia Pacific Education Review*, 23(2), 233-244.
- [5]. Almutairi, J., & Aldossary, M. (2021). Modeling and analyzing offloading strategies of IoT applications over edge computing and joint clouds. *Symmetry*, 13(3), 402.
- [6]. Althoubi, A., Alshahrani, R., & Peyravi, H. (2021). Delay analysis in IoT sensor networks. *Sensors*, 21(11), 3876.
- [7]. Amodu, O. A., Othman, M., Yunus, N. A. M., & Hanapi, Z. M. (2021). A primer on design aspects and recent advances in shuffle exchange multistage interconnection networks. *Symmetry*, 13(3), 378.
- [8]. Bandara, S., Rajeev, P., & Gad, E. (2023). Power distribution system faults and wildfires: Mechanisms and prevention. *Forests*, 14(6), 1146.
- [9]. Barrios, Y., Rodríguez, A., Sánchez, A., Pérez, A., López, S., Otero, A., de la Torre, E., & Sarmiento, R. (2020). Lossy hyperspectral image compression on a reconfigurable and fault-tolerant FPGA-based adaptive computing platform. *Electronics*, 9(10), 1576.
- [10]. Beliaeva, T., Shirokova, G., Wales, W., & Gafforova, E. (2020). Benefiting from economic crisis? Strategic orientation effects, trade-offs, and configurations with resource availability on SME performance. *International Entrepreneurship and Management Journal*, 16(1), 165-194.
- [11]. Bouizem, Y., Dib, D., Parlavantzis, N., & Morin, C. (2023). Integrating request replication into FaaS platforms: an experimental evaluation. *Journal of Cloud Computing*, 12(1), 94.
- [12]. Buckley, I. A., & Fernandez, E. B. (2023). Dependability patterns: a survey. *Computers*, 12(10), 214.
- [13]. Bulej, L., Bureš, T., Filandr, A., Hnětynka, P., Hnětynková, I., Pacovský, J., Sandor, G., & Gerostathopoulos, I. (2021). Managing latency in edge–cloud environment. *Journal of systems and software*, 172, 110872.
- [14]. Chen, G., Sheng, B., Luo, R., & Jia, P. (2022). A parallel strategy for predicting the quality of welded joints in automotive bodies based on machine learning. *Journal of Manufacturing Systems*, 62, 636-649.

- [15]. Chen, Y., Wu, C., Sui, R., & Zhang, J. (2024). Feasibility study of edge computing empowered by artificial intelligence—a quantitative analysis based on large models. *Big Data and Cognitive Computing*, 8(8), 94.
- [16]. Chen, Z., Song, X., Lee, T. M., & Zhang, R. (2023). The robust reciprocal relationship between loneliness and depressive symptoms among the general population: Evidence from a quantitative analysis of 37 studies. *Journal of Affective Disorders*, 343, 119-128.
- [17]. Cicconetti, C., Conti, M., & Passarella, A. (2020). Architecture and performance evaluation of distributed computation offloading in edge computing. *Simulation Modelling Practice and Theory*, 101, 102007.
- [18]. da Rosa, R., da Rosa, F. R., Ost, L., & Reis, R. (2020). *Soft Error Reliability Using Virtual Platforms*. Springer.
- [19]. Da Silva, F. C. T., & Barbosa, C. P. (2021). The impact of the COVID-19 pandemic in an intensive care unit (ICU): Psychiatric symptoms in healthcare professionals. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 110, 110299.
- [20]. da Silva, F. C. T., & Neto, M. L. R. (2021). Psychiatric disorders in health professionals during the COVID-19 pandemic: A systematic review with meta-analysis. *Journal of Psychiatric Research*, 140, 474-487.
- [21]. Damane, B. P., Kgokolo, M. C., Gaudji, G. R., Blenman, K. R., & Dlamini, Z. (2023). Integration of cyber-physical systems in the advancement of society 5.0 healthcare management. In *Society 5.0 and Next Generation Healthcare: Patient-Focused and Technology-Assisted Precision Therapies* (pp. 201-221). Springer.
- [22]. Duan, L., & Da Xu, L. (2024). Data analytics in industry 4.0: A survey. *Information Systems Frontiers*, 26(6), 2287-2303.
- [23]. Duan, Y., Wang, N., & Wu, J. (2022). Accelerating DAG-style job execution via optimizing resource pipeline scheduling. *Journal of Computer Science and Technology*, 37(4), 852-868.
- [24]. Fascher, M., Nowaczynski, S., Spindler, C., Strobach, T., & Muehlhan, M. (2024). Neural underpinnings of response inhibition in substance use disorders: weak meta-analytic evidence for a widely used construct. *Psychopharmacology*, 241(1), 1-17.
- [25]. Francis, T., Davidson, M., Senese, L., Jeffs, L., Yousefi-Nooraie, R., Ouimet, M., Rac, V., & Trbovich, P. (2024). Exploring the use of social network analysis methods in process improvement within healthcare organizations: a scoping review. *BMC Health Services Research*, 24(1), 1030.
- [26]. Hsin, L.-J., Chao, Y.-P., Chuang, H.-H., Kuo, T. B., Yang, C. C., Huang, C.-G., Kang, C.-J., Lin, W.-N., Fang, T.-J., & Li, H.-Y. (2023). Mild simulator sickness can alter heart rate variability, mental workload, and learning outcomes in a 360 virtual reality application for medical education: a post hoc analysis of a randomized controlled trial. *Virtual Reality*, 27(4), 3345-3361.
- [27]. Hu, H., Yu, S. S., & Trinh, H. (2024). A review of uncertainties in power systems—Modeling, impact, and mitigation. *Designs*, 8(1), 10.
- [28]. Javed, H., Toosi, A. N., & Aslanpour, M. S. (2022). Serverless platforms on the edge: a performance analysis. In *New Frontiers in Cloud Computing and Internet of Things* (pp. 165-184). Springer.
- [29]. Ji, Y., Leung, T. K. W., Dai, X., Du, N., & Wong, D. F. K. (2024). Gaming disorder: Unraveling the role of problematic affective, cognitive, and executive functioning—A systematic review and meta-analytic structural equation modeling. *Computers in Human Behavior*, 159, 108348.
- [30]. Kandemir, E., Hasan, A., Kvamsdal, T., & Abdel-Afou Alaliyat, S. (2024). Predictive digital twin for wind energy systems: a literature review. *Energy Informatics*, 7(1), 68.
- [31]. Kang, J., Cho, Y. S., Lee, M., Yun, S., Jeong, Y. J., Won, Y.-H., Hong, J., & Kim, S. (2023). Effects of nonpharmacological interventions on sleep improvement and delirium prevention in critically ill patients: a systematic review and meta-analysis. *Australian Critical Care*, 36(4), 640-649.
- [32]. Kazi Rakib Hasan, S., & Adnan, S. (2022). Quantitative Assessment of Machine Learning-Based Schedule Risk Prediction for Reducing Delays in Complex Technology Projects. *Review of Applied Science and Technology*, 1(04), 464–501. <https://doi.org/10.63125/dn8nkm85>
- [33]. Khaledian, N., Voelp, M., Azizi, S., & Shirvani, M. H. (2024). AI-based & heuristic workflow scheduling in cloud and fog computing: a systematic review. *Cluster Computing*, 27(8), 10265-10298.
- [34]. Khan, A. Q., Matskin, M., Prodan, R., Bussler, C., Roman, D., & Soylyu, A. (2024). Cloud storage cost: a taxonomy and survey. *World Wide Web*, 27(4), 36.
- [35]. Khan, A. R. (2024). Dynamic load balancing in cloud computing: optimized RL-based clustering with multi-objective optimized task scheduling. *Processes*, 12(3), 519.
- [36]. Koziolek, H., Grüner, S., & Rückert, J. (2020). A comparison of MQTT brokers for distributed IoT edge computing. *European Conference on Software Architecture*,
- [37]. Kumar, A. Behnaz Ranjbar Alireza Ejlali.
- [38]. Kumar, G. H., & Ramesh, G. (2022). Node localization algorithm for detecting malicious nodes to prevent connection failures and improve end-to-end delay. *Computer Communications*, 190, 37-47.
- [39]. Lei, M., Wang, X., Wang, M., & Cheng, Y. (2024). Improved yolov5 based on multi-strategy integration for multi-category wind turbine surface defect detection. *Energies*, 17(8), 1796.
- [40]. Li, B., Liu, W., Xie, W., & Li, X. (2023). Energy-efficient task offloading and trajectory planning in UAV-enabled mobile edge computing networks. *Computer Networks*, 234, 109940.
- [41]. Li, G., Tong, Y., Zhang, G., & Zeng, Y. (2023). An energy fault and consumption optimization strategy in wireless sensor networks with edge computing. *Journal of King Saud University Computer and Information Sciences*, 35(1), 357-367.
- [42]. Li, H., Yao, Q., & Li, X. (2024). Voiceprint fault diagnosis of converter transformer under load influence based on Multi-Strategy improved Mel-Frequency spectrum coefficient and Temporal convolutional network. *Sensors*, 24(3).

- [43]. Li, L., Feng, Z., Zhu, M., Yang, J., & Yang, L. (2023). The mediating effect of personality on mental workload and perceived professional benefits of nurses in East China. *BMC nursing*, 22(1), 440.
- [44]. Lin, L., Jin, Y., Xiong, M., Wu, S., & Sun, S. (2023). The protective power of cognitive reserve: Examining white matter integrity and cognitive function in the aging brain for sustainable cognitive health. *Sustainability*, 15(14), 11336.
- [45]. Liu, Y., Ma, X., Qiao, W., Luo, H., & He, P. (2021). Human factor risk modeling for shipyard operation by mapping fuzzy fault tree into Bayesian network. *International journal of environmental research and public health*, 19(1), 297.
- [46]. Mahbubul, I., Mahmudul, I., Shifa, T., Mahfuza, T., & Shima Ali, S. (2022). Exploring The Potential of Visual Narratives and Interactive Gamified Communication Systems to Prevent Disease, Counter Health Misinformation, And Strengthen Public Trust in Low-Literacy Communities: A Literature Review. *American Journal of Interdisciplinary Studies*, 3(02), 165-220. <https://doi.org/10.63125/zmq54n64>
- [47]. Martinez, H. F., Mondragon, O. H., Rubio, H. A., & Marquez, J. (2022). Computational and communication infrastructure challenges for resilient cloud services. *Computers*, 11(8), 118.
- [48]. Martinez, J., Sañudo, I., & Bertogna, M. (2020). End-to-end latency characterization of task communication models for automotive systems. *Real-Time Systems*, 56(3), 315-347.
- [49]. Mashfiqu, R., & Md, A. (2022). Monte Carlo Simulation and Stochastic Cash-Flow Modeling for Reducing Liquidity Risk in Project-Based Enterprises. *International Journal of Scientific Interdisciplinary Research*, 1(01), 382–431. <https://doi.org/10.63125/zf3fwv21>
- [50]. McLean, S. A., Booth, A. T., Schnabel, A., Wright, B. J., Painter, F. L., & McIntosh, J. E. (2021). Exploring the efficacy of telehealth for family therapy through systematic, meta-analytic, and qualitative evidence. *Clinical Child and Family Psychology Review*, 24(2), 244-266.
- [51]. Md Ashfaq, S., & Abdullah Mohammad, S. (2022). CNT/PANI Composite Electrodes for Flexible Fiber-Shaped Dye-Sensitized Solar Cells: Synthesis, Characterization, and Photovoltaic Performance. *American Journal of Interdisciplinary Studies*, 3(02), 221-274. <https://doi.org/10.63125/m334mf41>
- [52]. Md Asif Ali Sheak, A., & Tahmedur, R. (2023). A Data-Driven Framework for Engineering Project Risk Assessment Under Schedule and Cost Uncertainty. *International Journal of Scientific Interdisciplinary Research*, 4(2), 204–251. <https://doi.org/10.63125/kq4zzh16>
- [53]. Md Kaium, H. (2023). A GIS-Based Multi-Criteria Decision-Support and Land-Suitability Optimization Framework for Climate-Resilient Urban and Regional Planning. *American Journal of Data Science and Analytics*, 4(06), 43–103. <https://doi.org/10.63125/wvq0sx87>
- [54]. Md Mahamud Pervaz, P. (2023). Integration of MES, SAP, And PI Historian Platforms for Real-Time Industrial Asset Performance Management. *Review of Applied Science and Technology*, 2(04), 393–445. <https://doi.org/10.63125/9ebtas97>
- [55]. Md. Abdur, R., & Iftekhar, A. (2021). Customer Retention Forecasting in Mobile Wallet Services Using Neural Networks: A Comparative Quantitative Study. *International Journal of Business and Economics Insights*, 1(4), 70-102. <https://doi.org/10.63125/dyrpc387>
- [56]. Menaceur, A., Drid, H., & Rahouti, M. (2023). Fault tolerance and failure recovery techniques in software-defined networking: A comprehensive approach. *Journal of Network and Systems Management*, 31(4), 83.
- [57]. Mohammad Badrul, A., & Md. Mominul, H. (2023). Machine Learning-Based Assessment of Environmental and Public Health Risks in Sewerage Sludge Management Systems. *American Journal of Advanced Technology and Engineering Solutions*, 3(02), 33-77. <https://doi.org/10.63125/y3yfxa92>
- [58]. Molla Al Rakib, H., & Mohammad Towhidul, I. (2023). Cyber-Resilient Marketing Intelligence Systems for Fraud Detection and Data Integrity in Digital Campaigns. *American Journal of Scholarly Research and Innovation*, 2(02), 485–518. <https://doi.org/10.63125/b53pgr11>
- [59]. Moulik, S., & Sharma, Y. (2024). FRESH: Fault-tolerant real-time scheduler for heterogeneous multiprocessor platforms. *Future Generation Computer Systems*, 161, 214-225.
- [60]. Muñoz, A. (2024). Cracking the core: Hardware vulnerabilities in android devices unveiled. *Electronics*, 13(21), 4269.
- [61]. Mushtaq, S. U., Sheikh, S., Idrees, S. M., & Malla, P. A. (2024). In-depth analysis of fault tolerant approaches integrated with load balancing and task scheduling. *Peer-to-Peer Networking and Applications*, 17(6), 4303-4337.
- [62]. Nakagawa, S., Yang, Y., Macartney, E. L., Spake, R., & Lagisz, M. (2023). Quantitative evidence synthesis: a practical guide on meta-analysis, meta-regression, and publication bias tests for environmental sciences. *Environmental Evidence*, 12(1), 8.
- [63]. Nishat, J. (2023). Communicable Disease Prevention in a Post-Pandemic Era: A Global Review of Public Health Education, Surveillance, and Community-Level Interventions. *International Journal of Scientific Interdisciplinary Research*, 4(4), 630–678. <https://doi.org/10.63125/mnqefj36>
- [64]. Okazaki, S., Eisend, M., Plangger, K., de Ruyter, K., & Grewal, D. (2020). Understanding the strategic consequences of customer privacy concerns: A meta-analytic review. *Journal of Retailing*, 96(4), 458-473.
- [65]. Omidvar, M., Zarei, E., & Ramavandi, B. (2024). The Bow-Tie Method: A Hybrid System Safety and Risk Analysis Approach for Safety-Critical Sociotechnical Systems. In *Safety causation analysis in sociotechnical systems: advanced models and techniques* (pp. 123-149). Springer.
- [66]. Peng, Z., Li, J., Hao, H., & Zhong, Y. (2024). Smart structural health monitoring using computer vision and edge computing. *Engineering Structures*, 319, 118809.
- [67]. Pérez, J., Díaz, J., Berrocal, J., López-Viana, R., & González-Prieto, Á. (2022). Edge computing: A grounded theory study. *Computing*, 104(12), 2711-2747.
- [68]. Quinn-Nilas, C. (2020). Self-reported trait mindfulness and couples' relationship satisfaction: A meta-analysis. *Mindfulness*, 11(4), 835-848.

- [69]. Rak, J., Hutchison, D., Tapolcai, J., Bruzgiene, R., Tornatore, M., Mas-Machuca, C., Furdek, M., & Smith, P. (2020). Fundamentals of communication networks resilience to disasters and massive disruptions. In *Guide to disaster-resilient communication networks* (pp. 1-43). Springer.
- [70]. Ramzey, H., Badawy, M., Elhosseini, M., & A. Elbaset, A. (2023). I2OT-EC: A framework for smart real-time monitoring and controlling crude oil production exploiting IIOT and edge computing. *Energies*, 16(4), 2023.
- [71]. Ranji, R., Mansoor, A. M., & Sani, A. A. (2020). EEDOS: an energy-efficient and delay-aware offloading scheme based on device to device collaboration in mobile edge computing: R. Ranji et al. *Telecommunication Systems*, 73(2), 171-182.
- [72]. Reyana, A., Kautish, S., Alnowibet, K. A., Zawbaa, H. M., & Wagdy Mohamed, A. (2023). Opportunities of IoT in fog computing for high fault tolerance and sustainable energy optimization. *Sustainability*, 15(11), 8702.
- [73]. Rostami, S., Broumandnia, A., & Khademzadeh, A. (2024). An energy-efficient task scheduling method for heterogeneous cloud computing systems using capuchin search and inverted ant colony optimization algorithm. *The Journal of Supercomputing*, 80(6), 7812-7848.
- [74]. Roy, A., Pachua, J. L., & Saha, A. K. (2021). An overview of queuing delay and various delay based algorithms in networks. *Computing*, 103(10), 2361-2399.
- [75]. Ruj, S., & Pal, A. (2020). Fault Tolerance and Reliability of Smart Grids. In *Encyclopedia of Wireless Networks* (pp. 468-478). Springer.
- [76]. Sadia, Z., & Maishar Mostari, Z. (2022). The Impact of Interactive Dashboard Visualization on Data-Driven Decision Making in Financial Services: A Qualitative Evaluation. *International Journal of Scientific Interdisciplinary Research*, 12(03), 01-35. <https://doi.org/10.63125/3jstsv75>
- [77]. Sahoo, S. S., Das, A., & Kumar, A. (2023). Fault tolerant architectures. In *Handbook of Computer Architecture* (pp. 1-44). Springer.
- [78]. Sahu, M., Damle, S., & Kherani, A. A. (2021). End-to-end uplink delay jitter in LTE systems. *Wireless Networks*, 27(3), 1783-1800.
- [79]. Scheibe, S., De Bloom, J., & Modderman, T. (2022). Resilience during crisis and the role of age: Involuntary telework during the COVID-19 pandemic. *International journal of environmental research and public health*, 19(3), 1762.
- [80]. Schroeder, P. A., Schwippel, T., Wolz, I., & Svaldi, J. (2020). Meta-analysis of the effects of transcranial direct current stimulation on inhibitory control. *Brain stimulation*, 13(5), 1159-1167.
- [81]. Shaiful, M., Anisur, R., & Md, A. (2022). A Systematic Literature Review on the Role of Digital Health Twins in Preventive Healthcare For Personal and Corporate Wellbeing. *American Journal of Interdisciplinary Studies*, 3(04), 1-31. <https://doi.org/10.63125/negjw373>
- [82]. Shaikh, S., & Jammal, M. (2024). Survey of fault management techniques for edge-enabled distributed metaverse applications. *Computer Networks*, 254, 110803.
- [83]. Sharma, Y., & Moulik, S. (2023). FATS-2TC: A fault tolerant real-time scheduler for energy and temperature aware heterogeneous platforms with two types of cores. *Microprocessors and Microsystems*, 96, 104744.
- [84]. Si, W., & Niu, L. (2024). Enhancing human reliability prediction in smart tower crane interfaces: a refined approach using simplified plant analysis risk-human reliability assessment and the decision making trial and evaluation laboratory-analytic network process. *Buildings*, 14(4), 1083.
- [85]. Simeone, A., Grant, R., Ye, W., & Caggiano, A. (2023). A human-cyber-physical system for Operator 5.0 smart risk assessment. *The International Journal of Advanced Manufacturing Technology*, 129(5), 2763-2782.
- [86]. Singh, P., & Singh, R. (2022). Energy-efficient delay-aware task offloading in fog-cloud computing system for IoT sensor applications. *Journal of Network and Systems Management*, 30(1), 14.
- [87]. Sobh, T., Turnbull, B., & Moustafa, N. (2023). A holistic review of cyber-physical-social systems: New directions and opportunities. *Sensors*, 23(17), 7391.
- [88]. Speyer, R., Sutt, A.-L., Bergström, L., Hamdy, S., Heijnen, B. J., Remijn, L., Wilkes-Gillan, S., & Cordier, R. (2022). Neurostimulation in people with oropharyngeal dysphagia: a systematic review and meta-analyses of randomised controlled trials – part i: pharyngeal and neuromuscular electrical stimulation. *Journal of Clinical Medicine*, 11(3), 776.
- [89]. Srivastava, A., & Kumar, N. (2024). An efficient firefly and honeybee based load balancing mechanism in cloud infrastructure. *Cluster Computing*, 27(3), 2805-2827.
- [90]. Steel, P., Beugelsdijk, S., & Aguinis, H. (2021). The anatomy of an award-winning meta-analysis: Recommendations for authors, reviewers, and readers of meta-analytic reviews: Piers Steel et al. *Journal of International Business Studies*, 52(1), 23-44.
- [91]. Stewart, R., Chowdhury, M., & Arjoon, V. (2021). Bank stability and economic growth: trade-offs or opportunities? R. Stewart et al. *Empirical Economics*, 61(2), 827-853.
- [92]. Sudhakar, M., & Anne, K. R. (2024). Optimizing data processing for edge-enabled IoT devices using deep learning based heterogeneous data clustering approach. *Measurement: sensors*, 31, 101013.
- [93]. Suresh, P., Keerthika, P., Devi, R. M., Kamalam, G., Logeswaran, K., Sadasivuni, K. K., & Devendran, K. (2024). Optimized task scheduling approach with fault tolerant load balancing using multi-objective cat swarm optimization for multi-cloud environment. *Applied Soft Computing*, 165, 112129.
- [94]. Suslu, B., Ali, F., & Jennions, I. K. (2023). Understanding the role of sensor optimisation in complex systems. *Sensors*, 23(18), 7819.
- [95]. Syed Feroz, A. (2023). Quality-Management Systems for Resilient Energy Infrastructure Fabrication. *International Journal of Scientific Interdisciplinary Research*, 13(10), 37-91. <https://doi.org/10.63125/mbhhrm88>
- [96]. Tahmedur, R., & Rifat, C. (2023). Critical Path Analysis and Resource-Leveling Optimization for Multi-Project Engineering Portfolios. *American Journal of Interdisciplinary Studies*, 13(12), 01-47. <https://doi.org/10.63125/8eqnea46>

- [97]. Toczé, K., Lindqvist, J., & Nadjm-Tehrani, S. (2020). Characterization and modeling of an edge computing mixed reality workload. *Journal of Cloud Computing*, 9(1), 46.
- [98]. Torquato, M., Maciel, P., & Vieira, M. (2022). Model-based performability and dependability evaluation of a system with vm migration as rejuvenation in the presence of bursty workloads. *Journal of Network and Systems Management*, 30(1), 3.
- [99]. Udayakumar, P., & Anandan, D. R. (2024). Get started with iot network and security. In *Design and Deploy IoT Network & Security with Microsoft Azure: Embrace Microsoft Azure for IoT Network Enhancement and Security Uplift* (pp. 1-95). Springer.
- [100]. van Bussel, M. J., Odekerken-Schröder, G. J., Ou, C., Swart, R. R., & Jacobs, M. J. (2022). Analyzing the determinants to accept a virtual assistant and use cases among cancer patients: a mixed methods study. *BMC Health Services Research*, 22(1), 890.
- [101]. Wang, H., Li, D., Sheng, T., Sheng, J., Jing, P., & Zhang, D. (2023). A modeling of human reliability analysis on dam failure caused by extreme weather. *Applied Sciences*, 13(23), 12968.
- [102]. Wang, Y., Wang, H., & Cao, Y. (2024). Comprehensive review of storage optimization techniques in blockchain systems. *Applied Sciences*, 15(1), 243.
- [103]. Wei, D., Wang, H., Liu, X., Li, X., Gu, J., Zou, X., Tan, C., & Si, L. (2024). Adaptive tracking control method for shearer cutting trajectory based on the combined strategy. *Control Engineering Practice*, 153, 106116.
- [104]. Wen, J., Yang, J., Wang, T., Li, Y., & Lv, Z. (2023). Energy-efficient task allocation for reliable parallel computation of cluster-based wireless sensor network in edge computing. *Digital Communications and Networks*, 9(2), 473-482.
- [105]. Wu, G., Miao, Y., Zhang, Y., & Barnawi, A. (2020). Energy efficient for UAV-enabled mobile edge computing networks: Intelligent task prediction and offloading. *Computer Communications*, 150, 556-562.
- [106]. Wu, J. (2024). Introduction to functional safety and cyber resilience. In *Cyber resilience system engineering empowered by endogenous security and safety* (pp. 203-278). Springer.
- [107]. Xie, Z., Zhang, J., Zhang, Y., Xu, C., Chen, P., Qu, Z., & Song, W. (2023). An efficient fault tolerance strategy for multi-task mapreduce models using coded distributed computing. *International Conference on Algorithms and Architectures for Parallel Processing*.
- [108]. Xu, M., Zou, Y., & Cheng, X. (2024). Byzantine fault-tolerant wireless consensus. In *Wireless Consensus: Theory and Applications* (pp. 95-140). Springer.
- [109]. Zahid, F., Tanveer, A., Kuo, M. M., & Sinha, R. (2022). A systematic mapping of semi-formal and formal methods in requirements engineering of industrial cyber-physical systems. *Journal of Intelligent Manufacturing*, 33(6), 1603-1638.
- [110]. Zhang, Y., & Fu, J. (2021). Energy-efficient computation offloading strategy with tasks scheduling in edge computing. *Wireless Networks*, 27(1), 609-620.
- [111]. Zhang, Y., Tang, B., Luo, J., & Zhang, J. (2022). Deadline-aware dynamic task scheduling in edge-cloud collaborative computing. *Electronics*, 11(15), 2464.
- [112]. Zhao, Z., Song, W., Wang, H., Sun, Y., & Luo, H. (2024). Development and application of IoT monitoring systems for typical large amusement facilities. *Sensors*, 24(14), 4433.