EDGE & CLOUD COMPUTING: A REVIEW OF ARCHITECTURE, CHALLENGES, AND FUTURE DIRECTIONS

Author Name
Affiliation
Email

***Abstract*—The proliferation of Internet of Things (IoT) devices and the escalating demand for low-latency and data-intensive services expose the limitations of conventional cloud-centric computing. While cloud platforms offer near-elastic compute and storage, their reliance on distant data centers introduces latency, bandwidth, and privacy constraints that impede real-time and mission-critical applications. Edge computing complements the cloud by relocating computation, storage, and networking closer to data sources and users to reduce latency and improve efficiency. The integration across the cloud–edge continuum enables dynamic workload placement between centralized and distributed resources, yielding performance, scalability, and reliability gains, while introducing new complexities in coordination, heterogeneity management, security, cost, and sustainability. This paper reviews the state of cloud and edge computing with emphasis on architectures, cross-cutting challenges, applications, and future directions. It presents a layered architectural view from devices through edge/fog to cloud, proposes a taxonomy of challenges spanning technical and socio-economic facets, examines emerging platforms and evaluation practices, and highlights trends including serverless edge, federated learning, confidential computing, WebAssembly-based runtimes, AI-driven orchestration, and carbon-aware scheduling.**

***Index Terms*—Cloud computing, edge computing, fog computing, hybrid cloud, orchestration, latency, federated learning, serverless, confidential computing, WebAssembly, sustainability, carbon-aware scheduling, 5G/6G.**

I. Introduction

Cloud computing provides on-demand access to a shared pool of configurable resources over the Internet, eliminating the burden of provisioning and operating physical infrastructure for consumers and enterprises [1], [2]. Its economies of scale and rich service models—IaaS, PaaS, and SaaS—have underpinned the rapid scaling of modern web services and data analytics. However, centralization can be at odds with the stringent latency, bandwidth, and sovereignty requirements of emerging applications (autonomous systems, immersive reality, industrial control), which mandate timely responses and local data handling.

Edge computing addresses these limitations by distributing computation and data storage toward the network periphery—on or near devices and gateways—thereby reducing round-trip delays and curbing core network traffic [3], [4]. 5G advances further lower the air-interface latency and provide bandwidth slices for critical services, yet place new demands on end-to-end orchestration across heterogeneous resources [5], [6]. The cloud–edge continuum integrates these paradigms so that latency-critical tasks execute close to data sources while cloud backends deliver large-scale analytics, learning, and global coordination [7], [10].

This review synthesizes the architectural landscape (§II–§III), interrogates challenges (§IV), surveys application domains (§V), and outlines future research directions (§VI). Concluding remarks are provided in §VII.

II. Background and Fundamentals

The evolution of computing reflects a pendulum between centralization and distribution, driven by technology maturation and application demands. Fig. 1 charts the progression from mainframes to edge computing, highlighting key enablers.

Key definitions used in this survey follow authoritative sources: cloud computing as standardized by NIST [1]; fog computing as the cloud extension toward the edge [8]–[10]; mobile edge computing (MEC) per ETSI [11]; and edge computing as the broader decentralization of compute/storage to the periphery [3], [4].

Concluding this section, the trend is toward a continuum where placement is fluid across devices, edge/fog, and cloud. This continuity motivates unified control, observability, and programming abstractions that hide heterogeneity while honoring locality and policy.

III. Architectures

A. Cloud-Only (Centralized) Architecture

In cloud-only designs, applications and data are hosted in centralized data centers operated by cloud providers. Users access services over the Internet, leveraging virtualization and multi-tenancy to elastically scale [2]. Fig. 2 depicts the canonical pipeline from end-users through the Internet to virtualized compute, storage, and networking planes, with IaaS/PaaS/SaaS service layers.

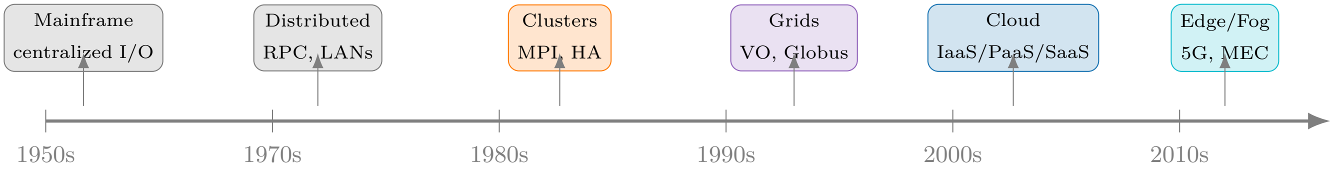

Fig. 1: Computing paradigms evolution timeline from mainframes to edge/fog and cloud.

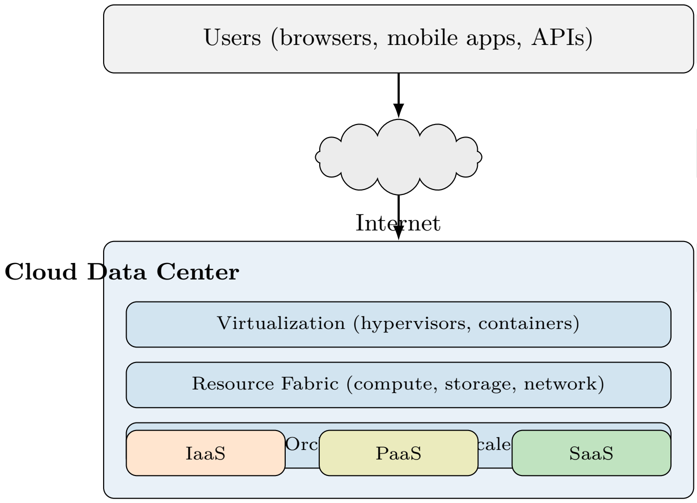
Users (browsers, mobile apps, APIs)
Internet
Cloud Data Center
Virtualization (hypervisors, containers)
Resource Fabric (compute, storage, network)
Orc
cale
IaaS
PaaS
SaaS

Fig. 2: Cloud-only architecture: centralized data centers with virtualization and service models (IaaS/PaaS/SaaS).

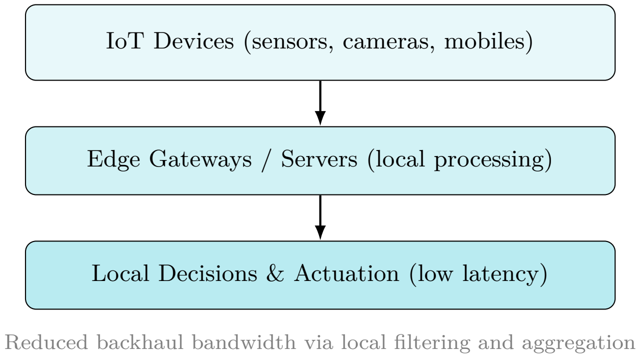
IoT Devices (sensors, cameras, mobiles)
Edge Gateways / Servers (local processing)
Local Decisions & Actuation (low latency)
Reduced backhaul bandwidth via local filtering and aggregation

Fig. 3: Edge-only architecture: decentralized processing near data sources for low-latency responses and bandwidth savings.

Discussion and concluding note: Cloud-only approaches maximize economies of scale and operational simplicity, but may violate latency and data residency constraints for time-critical workloads. They remain indispensable for global analytics and durable storage.

B. Edge-Only (Decentralized) Architecture

Edge-only designs shift computation, filtering, and short-term storage close to data sources (Fig. 3), minimizing wide-area traffic and enabling near-real-time actuation [3], [4].

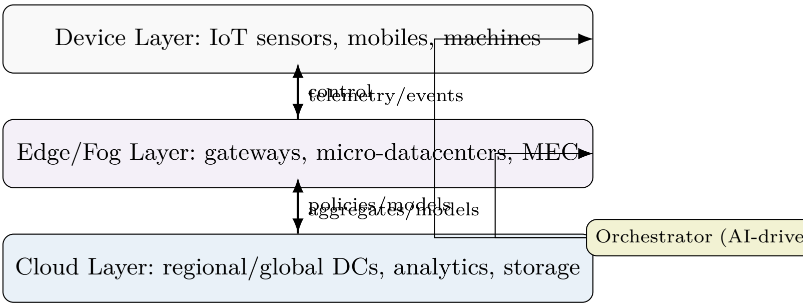
Device Layer: IoT sensors, mobiles, machines
control
telemetry/events
Edge/Fog Layer: gateways, micro-datacenters, MEC
policies/models
aggregates/models
Orchestrator (AI-drive
Cloud Layer: regional/global DCs, analytics, storage

Fig. 4: Hybrid three-tier architecture across device, edge/fog, and cloud layers with bi-directional data flows and AI-driven orchestration.

Discussion and concluding note: Edge-only deployments excel for safety-critical control loops and privacy-preserving analytics but face constraints in aggregate compute capacity, manageability, and cross-edge coordination.

C. Hybrid Cloud–Edge (Three-Tier) Architecture

Most real systems adopt a hybrid model with device, edge/fog, and cloud layers (Fig. 4). Time-critical tasks run near the edge; heavy compute, training, and global optimization are offloaded to the cloud. Intermediate fog nodes aggregate and coordinate among edges [10], [11].

Discussion and concluding note: Hybrid architectures reconcile latency and scale while enabling progressive deployment. They necessitate robust placement, state synchronization, and policy enforcement across administrative domains.

D. Comparative Analysis of Computing Models

Fig. 5 summarizes cloud, fog, and edge trade-offs along latency, compute/storage capacity, cost, and typical use cases.

IV. Challenges

The cloud–edge continuum introduces interdependent challenges spanning systems, networking, and operations (Fig. 6).

A. Scalability and Resource Management

Edge environments are resource-constrained and geographically distributed. Containers provide lightweight isolation compared to VMs, but orchestration at scale across heterogeneous sites remains immature relative to

Model	Latency	Compute/Storage	Cost	Typical Use Cases
Cloud	High	Very High	Pay-as-you-go	Global analytics, training, batch processing
Fog	Medium	Medium	Moderate	Aggregation, regional coordination, pre-processing
Edge	Very Low	Low–Medium	Hardware-heavy	Real-time control, privacy-preserving inference

Fig. 5: Comparative view of cloud, fog, and edge computing characteristics.

cloud counterparts [5], [6]. Scheduling must jointly optimize latency, throughput, and energy under fluctuating demand and failures. Concluding, there is a need for intent-driven, hierarchy-aware orchestration that spans device/edge/fog/cloud with robust multi-tenancy.

B. Networking, Connectivity, and Reliability

Intermittent wireless links, variable backhaul, and diverse access technologies complicate consistent QoS. 5G/MEC reduce radio latency but end-to-end guarantees require placement-aware routing, congestion control, and service function chaining [11]. In conclusion, cross-layer codesign is essential to bridge application SLOs and network realities.

C. Latency and Quality of Service

Ultra-low-latency applications (e.g., industrial control, AR/VR) are sensitive to jitter and tail latency. Predictive models for QoS can guide proactive scaling and prewarming of functions, particularly for serverless platforms. Concluding, robust percentile-tail management and SLA prediction are first-order design goals.

D. Security and Privacy

Decentralization enlarges the attack surface. Confidential computing (TEEs), remote attestation, encryption at rest/in transit, and zero-trust access are vital [16]. Privacy-preserving analytics (federated learning and secure aggregation) mitigate data exfiltration risks [15], [21]. Concluding, verifiable trust and privacy budgets should be built into edge data pipelines.

E. Heterogeneity

Hardware diversity (CPUs, GPUs, TPUs, NPUs), instruction sets, and accelerators, alongside protocol variety, complicate portability. WebAssembly and container images for multiple architectures improve portability, but performance portability remains an open challenge [17]. Concluding, standardized execution substrates and interface descriptions are needed.

F. Energy Efficiency and Sustainability

Battery-powered devices and thermally constrained sites necessitate energy-aware computing (DVFS, adaptive offloading) and carbon-aware scheduling that shifts delay-tolerant workloads to greener regions/times [18]. Concluding, sustainability should be a first-class objective across placement, networking, and model selection.

G. Cost Optimization

While edge reduces bandwidth costs via local processing, it introduces CAPEX (hardware) and OPEX (remote management). Transparent costing models and joint optimization of placement and data movement are required. Concluding, economics must be integrated with technical scheduling decisions.

H. Mobility Support

User and device mobility requires session continuity, stateful migration, and location-aware routing. Techniques such as predictive handover and micro-state replication can preserve QoE for vehicular and mobile AR use cases [20]. Concluding, mobility must be elevated to a first-class dimension in orchestration.

V. Applications

A. Smart Cities and Public Infrastructure

City-scale sensing and actuation benefit from local inference (traffic signals, public safety) with cloud-based planning and policy optimization. Healthcare monitoring leverages wearables and home sensors for continuous telemetry and edge filtering, while cloud aggregates support population-level analytics.

B. Autonomous Vehicles and Intelligent Transportation

Autonomous systems rely on multi-modal sensing (LiDAR, radar, cameras) and near-sensor inference for safety. Edge sites at RSUs and road-side cabinets support collaborative perception; the cloud supports fleet-scale learning and map updates.

C. AR/VR and Real-Time Media

Interactive applications require motion-to-photon latency below human perceptual thresholds. Edge rendering and encoding coupled with cloud asset management deliver immersive experiences with constrained backhaul.

D. Industry 4.0

Industrial IoT uses digital twins, predictive maintenance, and closed-loop control. Edge gateways ensure deterministic response on the factory floor, while the cloud provides model training and supply-chain integration.

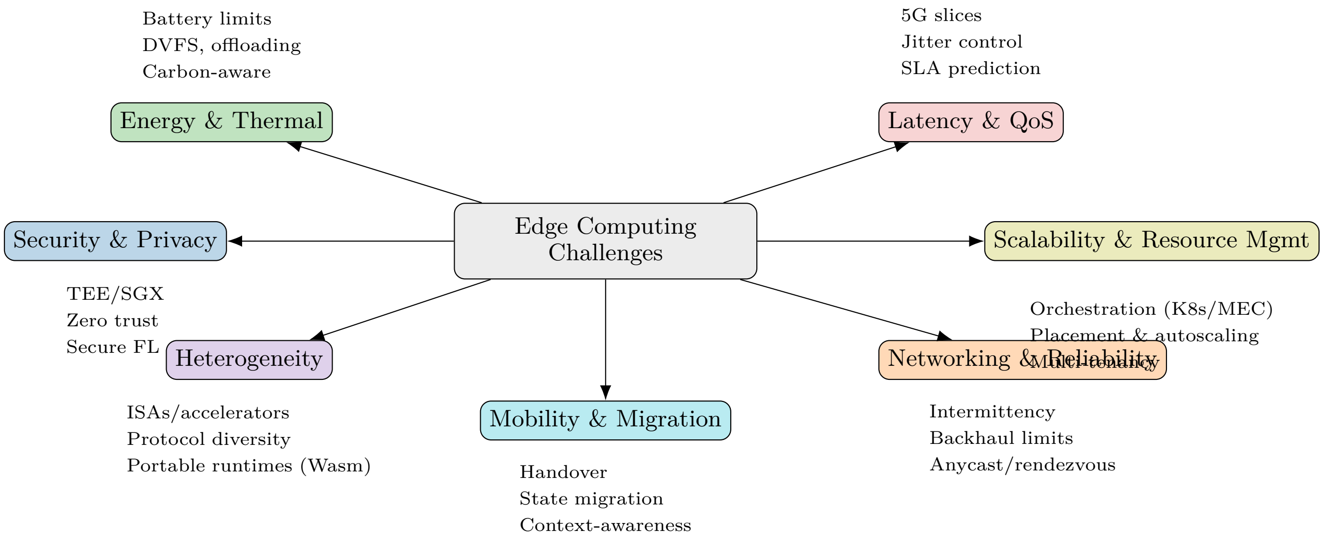

Fig. 6: Mind map of key challenges for edge computing across systems, networking, and operations.

E. Edge AI

Deploying ML at the edge reduces latency and preserves privacy. Federated learning trains global models without centralizing raw data, complemented by secure aggregation and TEEs for confidentiality [15], [16], [21]. Concluding, model compression and hardware-aware NAS are crucial for sustainable edge AI.

VI. Future Directions

A. AI-Driven Orchestration

Learning-based schedulers can predict demand, prewarm functions, and optimize placement across the continuum in real time.

B. Blockchain-Assisted Security

Smart contracts for decentralized access control and audit can complement traditional PKI for IoT/edge trust, though performance and cost trade-offs remain open.

C. Confidential Computing and Privacy

TEEs and emerging confidential GPUs enable verifiable processing on untrusted sites; combining with differential privacy and federated methods can strengthen end-to-end confidentiality [16].

D. Serverless at the Edge

Function-as-a-Service adapted for constrained nodes must mitigate cold starts, enable fine-grained autoscaling, and expose latency SLOs. Portable runtimes (e.g., Wasm) are promising substrates [17].

E. WebAssembly-Based Runtimes

Wasm provides a secure, lightweight, and portable execution format suited for heterogeneous hardware, enabling language-agnostic deployment with near-native performance.

F. 6G and AI-Native Networks

6G envisions AI-native control loops and integrated sensing/communications, amplifying the need for co-designed compute-network orchestration [19].

G. Carbon-Aware Scheduling and Green Computing

Shifting delay-tolerant compute to low-carbon regions and periods, model sparsification, and right-sizing resources can reduce environmental impact without sacrificing SLOs [18].

H. Standardization and Interoperability

Open interfaces across orchestration, telemetry, and security (attestation) are required to tame heterogeneity and promote portability across vendors and domains [9], [11].

VII. Conclusion

This review articulated the architectural foundations, systemic challenges, and application opportunities spanning the cloud–edge continuum. Cloud-only designs deliver scale, edge-only deployments deliver immediacy, and hybrid models reconcile both through layered placement and orchestration. Progress hinges on unifying abstractions, verifiable trust, AI-driven operations, and sustainability-aware optimization. With continued advances in 5G/6G, confidential computing, serverless runtimes, and federated learning, future systems can be both performant and privacy-preserving, efficient and sustainable.

Acknowledgment

This work synthesizes publicly available research and standards. Author and affiliation placeholders should be replaced prior to submission.

References

[1] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," NIST Special Publication 800-145, 2011.

[2] M. Armbrust *et al.*, "A View of Cloud Computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[3] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge Computing: Vision and Challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[4] M. Satyanarayanan, "The Emergence of Edge Computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017.

[5] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[6] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.

[7] B. Varghese and R. Buyya, "Next Generation Cloud Computing: New Trends and Research Directions," *Future Generation Computer Systems*, vol. 79, pp. 849–861, 2018.

[8] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog Computing and Its Role in the Internet of Things," in *Proc. MCC*, 2012, pp. 13–16.

[9] OpenFog Consortium Architecture Working Group, "OpenFog Reference Architecture for Fog Computing," 2017.

[10] M. Chiang and T. Zhang, "Fog and IoT: An Overview of Research Opportunities," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 854–864, 2016.

[11] ETSI, "Mobile Edge Computing (MEC); Framework and Reference Architecture," ETSI GS MEC 003 V1.1.1, 2016.

[12] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The Case for VM-Based Cloudlets in Mobile Computing," *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14–23, 2009.

[13] A. Yousefpour *et al.*, "All One Needs to Know About Fog Computing and Related Edge Computing Paradigms," *Journal of Systems Architecture*, vol. 98, pp. 289–330, 2019.

[14] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, "Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7457–7469, 2020.

[15] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. AISTATS*, 2017.

[16] F. Schuster *et al.*, "VC3: Trustworthy Data Analytics in the Cloud Using SGX," in *Proc. IEEE S&P*, 2015, pp. 38–54.

[17] A. Haas *et al.*, "Bringing the Web up to Speed with WebAssembly," in *Proc. PLDI*, 2017, pp. 185–200.

[18] Z. Liu, M. Lin, A. Wierman, S. Low, and L. L. H. Andrew, "Geographical Load Balancing with Renewables," *IEEE/ACM Transactions on Networking*, vol. 23, no. 6, pp. 1954–1967, 2015.

[19] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020.

[20] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile Edge Computing: A Survey," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 450–465, 2018.

[21] K. Bonawitz *et al.*, "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in *Proc. ACM CCS*, 2017, pp. 1175–1191.