Deep Science Publishing
https://doi.org/10.70593/978-93-49910-08-9_12

DeepScience
Open Access Books

Chapter 12: Exploring the future of cloud computing: Autonomous systems, edge artificial intelligence, and intelligent workload management

12.1. Introduction to Cloud Computing

The field of cloud computing is rapidly evolving, with profound implications for organizations, consumers, and everyday life. While computing capabilities have been evolving for decades, users traditionally interacted with dispersed resources bound in the physical local area surrounding the user, or at least in their logical domain of knowledge. This computational island of services and resource capacity consisted of local and on-premise local clusters or enterprise data centers, hosting servers with their various servers and storage for shared use in delivering time-sharing or best-effort services to users in the domain. But, with the transformation of the internet into a globalization infrastructure, which is high bandwidth, low latency, and highly resilient, the key layers of the cloud computing stack became services on demand, renting resources for the long run without worrying about the underlying infrastructure. For example, a service allowed policies-based bulk storage for images, and another offered machines on demand that provided physical resources to run programs on a best-effort and time-shared basis (Gill et al., 2022; Ibn-Khedher et al., 2022; Mishra et al., 2024).

Having become a layer of modern internet operations, cloud computing is now an integral part of many, if not most, consumers' daily lives and business operations. A working group was initiated with the goal of producing Systems Engineering Guidance for Cloud Computing. But, cloud computing is complex, with many layers and issues involved; it is not always easy to know where to start, or what issues to consider when making one or more decisions about cloud computing. This chapter presents some background discussion of what cloud computing involves and some guidance for

systems engineers, regardless of discipline or specialty area, who are faced with the decision or decisions about cloud computing.

Cloud computing is referred to as a huge pool of resources, which are not only shared by many users but also highly virtualized, so they can be conveniently accessed and used by heterogeneous end devices. Such devices would connect to the Internet via broadband wireless access since these huge pools of resources would be available worldwide. Data centers are physically distributed, and the only differences from general Internet servers are that they consist of a large number of machines with huge storage and broadband access to service a variety of user requests (Thota, 2024; Ramamoorthi, 2023; Walia et al., 2023).

Fig 12.1: Cloud Computing Autonomous Systems, Edge AI

12.1.1. Background and Significance

Cloud computing promises to provide computing resources on-demand, similar to other utility services while keeping management and coordinating computing resources a huge challenge for modern society. As individuals or companies start to rely more and more on clouds enabling important applications in our daily life like email, social networking, file storage, and photography, clouds receive heavy workload. In such cases, a data center can be viewed as an entity receiving requests from a variety of clients, processing jobs, and then returning results. The jobs of each user can be simple tasks, e.g., sending an email, printing a pest, uploading a photo etc. with quick responses needed from the

system; or they can be complicated jobs, e.g., constant exchanging of emails, searching for information, managing contacts etc. that leverage the system heavily. Ports in the data center deal with requests from various clients, while cloud machines respond to these requests and return results to clients via different ports.

Cloud computing can also be viewed as a generalized client-server computing paradigm. Servers can be either local workstations or resource clusters on the Internet running at various sites. Servers provide dynamic services on demand related to information retrieval and storage, database management, e-commerce transaction management, graphical processing, and other platform services. Clients like smart phones, PCs, and even automated sensors make continuous service requests.

12.2. The Evolution of Cloud Technologies

The term "cloud" was introduced into information technology (IT) usage after it emerged as an operating principle of multi-tenant IT service provisioning that enabled all types of organizations access to the latest IT without the burden of hardware or capabilities at their physical locations. However, the idea of shared IT services on demand is not new; it has been kicking around in some form for over 50 years. The origins of the idea of the cloud can be traced back to Large-Scale Information Systems that pioneered the concept of Shared Resources for Education and made computer resources available to schools across New York State for academic purposes. Significant advances in distributed computing and networking in the 1970s and the introduction of distributed workstation systems for interactive computing in the 1990s enabled the offer of relatively inexpensive data communications services to a much larger population and made the idea of local secure networked computing economically attractive at scale, enabling many scientific and creative users to collaborate on large shared computational problems. Later in its history, the Internet underwent a major shift to commercial services that offered aggregate capacity for web-based applications for all types of organizations. These shifts in the basic architecture of IT and telecommunications loomed over the emergence of cloud computing, which draws upon advancements in many areas of technology and service including virtualization of servers for on-demand shared capacities, utility pricing via service level agreements, global networked multi-tenant software-as-a-service designed to easily accommodate thousands of simultaneous users.

12.2.1. Research design

The research design is exploratory in nature and is based on a longitudinal analysis of three clusters of web and electronic data that describe and depict the evolution of cloud

technologies, their functionality, utilization, and the related market. A commercial hosted hybrid collaboration service that enables both internal and external collaboration is used as an illustration and functionally representative example of the evolution of cloud technologies. A category of core enabling cloud functions eWork core services is provided their evolution is synthesized to introduce a new category of cloud technologies called next generation cloud technologies and to illustrate their emergence trend. The category of core enabling cloud functions is then applied to sample both service provision and technological support members of the hosted hybrid collaboration service product category. These core enabling cloud functions are then synthesized to introduce a first portrayal of the emerging core function set needed of the complete next generation cloud technologies product or system. Next, the results are utilized to construct two concept marketing maps that depict the market related with the cloud technology product category illustrated. The top-level map provides a marketing view into the conceptual provision of cloud functions and services and is primarily consumer-oriented that is directed towards the service consumption market. The second map addresses the supply markets and provides a macro view into the aspect of which are the leading cloud technology manufacturers enabling each core cloud functionality. These two marketing maps not only allow assessment of the maturity stage of the cloud technologies product category within the cloud computing umbrella, but can serve as a tool to forecast its evolution in the near-term future. The concept marketing maps follow the four marketing functions of information provision demand stimulation product/service delivery, and service consumption.

12.3. Understanding Autonomous Systems

The increasing complexity of the third IT wave—cloud, data intelligence, and business—presents unique challenges. Enterprise architecture is facing constant urgency to provide integrated IT services across disparate hybrid and multi cloud environments that seamlessly connect to on-premise technology stacks and business processes. This complexity necessitates a next-generation adaptive architecture that is driven by artificial intelligence but requires validated decision authority across cloud services. Revolutionary autonomous systems—next-generation cyber-physical solutions with closed-cyber loops—are enabling intelligent workload management to achieve closed decision cycles. Embedding smarter systems with superior cognitive intelligence empowers IT to operate with zero-touch infrastructure management and IT business services to better connect to the market and business. Autonomous systems will reimagine computing, data, network, and storage capabilities, transitioning from augmentation to autonomy—supporting human augmented intelligence to decision-making delegation, specializing business agent roles and capabilities to achieve superior outcomes across all capabilities, functioning with digitally enabled cognitive empathy.

Autonomous systems represent the completion of the second IT transition from on-premise computing services to information cloud services. Cloud computing democratizes access, capabilities, and enabling technologies; and autonomous systems ride the IT technology democratization for superior levels of automation completion, decision-making speed, business knowledge, and operational efficiency. Deploying next-generation transport for empathy-driven advanced enterprise services is no longer the privileged domain of mega-enterprises and powerful data-centric business models. Enabled by no-code low-code business computing with integrated decision agents that provide decision-making assistance and are trained from market and business knowledge, all enterprises can easily adapt with continuous and real-time business changes, providing superior market outcomes.As far as a cloud is concerned, the maximum possible level of automation requires that the user has no control of, nor input to, the execution of the cloud functions. Thus, in a fully autonomous cloud, the cloud-computing infrastructure has self-management capabilities. The cloud has the capability to satisfy a user request for a cloud service via the service execution system that requires the least possible resource effort in executing and maintaining the service and also ensures an acceptable level of performance and availability. The cloud infrastructure determines what resources are required for service execution and maintains the health of the infrastructure and the various cloud services executed concurrently on the same or shared resources.

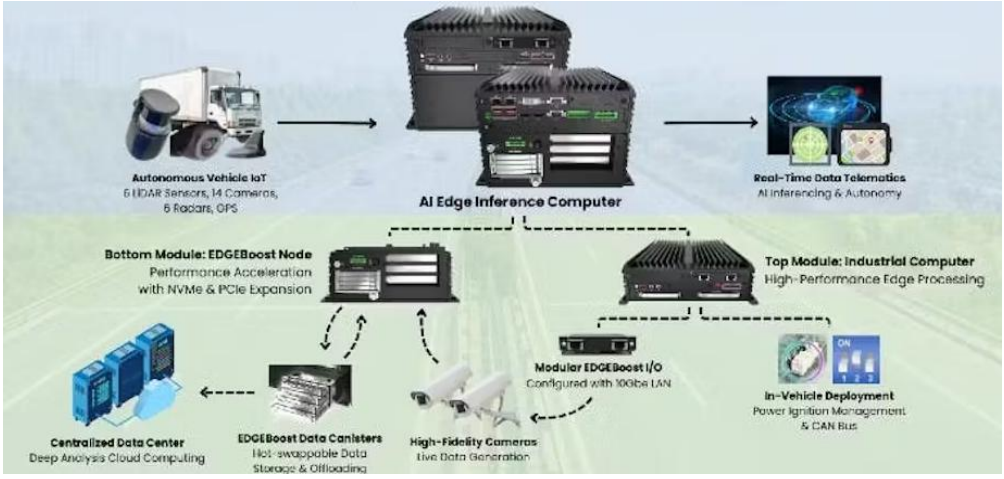

Fig 12.2: Autonomous Systems of Exploring the Future of Cloud Computing

12.3.1. Definition and Characteristics

An autonomous system is defined to be a system that has a partially or completely self-directed capability. The term "autonomous" is applied in a broad sense regarding the

level of automation provided to the system in the area of user control, guidance, and management. In the simplest case, the user has to initiate the execution of the operation sequence, and thereafter, the system executes the operation sequence with no additional user interaction. In a fully autonomous version, the system not only executes the operation sequence with no user interaction but also has the ability to determine and select the proper operation sequence to execute. The autonomous capability is only partial in those cases where the system provides some form of guidance to the user regarding the proper operation sequence to perform and/or requires user input at various points during system operation. The concept of guidance refers to the level of assistance provided to the user in complete control of the conduct of the operation. In the highest level of assistance, the system automates the execution of all but the definition of the operation sequence, which is provided by the user.

12.3.2. Applications in Cloud Computing

Cloud computing is much more than just a datacenter offering services to users. In the cloud realm, there are several applications of autonomous systems that can improve users' experience and resource management. One of these applications is to provide control and resource management of satellite and aerospace systems. The Space-Internet idea allows connectivity in all regions and is already being explored by several companies. It is expected to improve user latency because of the proximity of satellites relative to user devices. However, the always-on nature of the satellites still poses some challenges. For instance, a near-Earth titular requires an interplay of battery management, thermal control and other systems. These systems run software code that must have guaranteed behavior in order to properly steer the satellite to accomplish its mission. In addition to satellites, the study of how to use a constellation of CubeSats is also being explored, and how these CubeSats can offload Edge services.

Currently, most of the cloud datacenters are not using smart techniques to correlate the data that they already collect. With a considerable amount of heterogeneous appliances which are not being optimally utilized or even being batched-in or ignored. Using autonomous systems to monitor and control the datacenters is expected to improve resource utilization and also help reduce energy consumption in the datacenters. Virtualization environments are often used to support resource share and isolation in cloud execution. However, the optimal placement of heterogeneous workloads into the available resources is a complex problem that still relies on practical non-optimal heuristics. Using autonomous systems to manage the resource allocation process is being increasingly explored and is promising.

12.4. Edge AI: A Paradigm Shift

Advancements in existing data processing models, artificial intelligence (AI), and supported by a paradigm shift towards Edge Computing, promise a next stage revolution of cloud computing. Edge Computing is no longer just a concept, in just a few years considerable advances in Edge Computing implementation have been achieved. The combination of IoT and AI at the Edge creates the next evolution of Edge Computing, it acts as a facilitator to enable IoT in the Cloud and Smart IoT. Currently, thousands and millions of sensors are deployed in several areas and collecting enormous amounts of data. The data is then sent to the cloud to be processed, analyzed, structured, and queried to search for valuable information. Using the information that comes from the data collected by the IoT devices, services are created like Health Monitoring, Traffic Control, Smart Cities, Smart Energy, Weather Monitoring, Smart Health, and Security. The interaction between the Cloud and the end-users is done via the service application. In the last few years, the number of users with access to the internet has increased, increasing by this way the amount of information queried from the Cloud by the interface application of the services. With the IoT growing it is expected that the number of queries and requests received by the cloud will increase dramatically in the next few years. The associated threat with this increase is the infamous Cloud Bottleneck. A solution to these challenges is the implementation of Edge Computing, where processing and analytics operations are done by using Edge or Micro Data Centers located nearby the point of information generation. Moreover, both the data collection from sensors and the responses to the queries come from the Edge Points, reducing this way the need to send and receive information from the Cloud. With Edge Computing, users gain in terms of latency and response time and data providers companies save in terms of costs with bandwidth and network improvement. The question is not whether Edge Computing will be adopted.

12.4.1. Overview of Edge Computing

Emerging from the ubiquitous computing and cloud computing research agendas, edge computing represents a new architectural model in which services and content are moved from centralized data centers down closer to where these services and content are being used. Edge computing places small data centers closer to the customers to localize computing. In LTE networks, for example, small cell base stations are placed close to users' access points. Services are cached at the small cell base stations, so that they may be retrieved at low latency and little backhaul expenditure. Edge computing extends this principle further down the service stack from service delivery to content delivery to computing itself.

By pushing compute resources down to the edge, a number of real-time collaborative services become possible. In wireless networks, edge computing can facilitate the delivery of augmented reality and virtual reality applications. For these services, low latency, large bandwidth, and location-awareness are key. These characteristics can only be guaranteed by placing resources close to the users. For example, team collaboration applications which require sharing tools in real-time will require servers to be "nearby" to meet latency constraints. With augmented reality, moving computing to the edge allows instantaneous recognition of physical objects and provides forever available and accurate contextual search. A plausible model of computation for this is the edge collaborator, where every mobile user forms a client machine and data from the camera, GPS, accelerometer, and gyroscope are streamed to an edge server, which then sends back predicted results.

12.4.2. Integrating AI at the Edge

With advancements in computing, storage, and network technologies, many types of AI and machine learning algorithms can now be run locally on edge devices. Local execution provides many advantages for detecting, analyzing, and acting upon events since the data stays within the network perimeter boundary, eliminating privacy concerns. Furthermore, with the proper integration of AI and ML algorithms, low-cost and low-power devices can replicate some of the AI/ML workloads previously requiring remote, cloud-based systems. Emerging smart edge devices, such as lightweight cameras and visual sensors, microcontrollers with FPGAs or neuromorphic capabilities, and multimodal biometric sensors, blend perception, cognition, and action. Capable of communicating with one another and acting with a certain level of autonomy, intelligent edge devices are inextricably linked to the cloud—the two lie at the opposite extremes of the cloud-to-things continuum.

Traditionally, edge devices are designed with a specific purpose in mind, limiting their application to very niche areas. Integrating AI capabilities at the edge fosters computational intelligence, allowing devices to make intelligent decisions based on their situational awareness, thereby reaping the most benefits from edge computing without burdening the bandwidth-limited data channels. A good practical example of the advantage of moving AI to the edge with smart data acquisition is video surveillance. Smart AI-enhanced cameras with 3D facial recognition can be set up to trigger alarms only when they detect behaviors characteristic of action-at-a-distance, such as a person falling off a three-storey building, thereby restricting the transfer of terabytes of video data through bottlenecked fibers between the cloud servers and the video cameras. In this case, the enriched data on the right events is effectively delivered to the cloud studio

for higher-level analysis and long-term storage, where knowledge can be extracted through mining algorithms to model crowds, human behaviors, and events.

12.5. Intelligent Workload Management

Data centers are taking an increasing share of the overall energy consumption in the world. Deploying and adjusting heavy infrastructure costs too much energy and human resources. In addition, the current cloud services are designed to support agile and flexible services delivery. Often, such workloads cannot be planned beforehand, which leads to the surging resource demand asking for proper scalability design. The evolution from the Internet of Things to the Edge cloud also brings new workload distribution strategies, edge device resource management, as well as scheduling engines or middleware design principles.

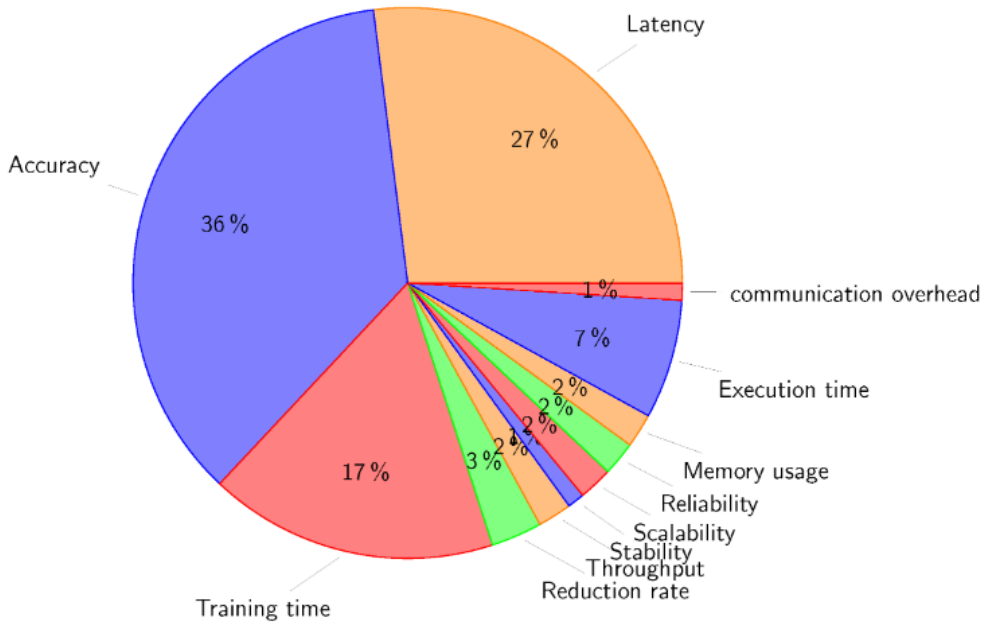

Fig : At the Confluence of Artificial Intelligence and Edge Computing

An edge cloud is a cloud deployed at the network side near the data source. Compared to traditional data centers, edge clouds have advantages like reduced latency, lower bandwidth cost, reduced bottleneck risk, etc. These advantages come from the fact that design edge virtualization requires relatively less design considerations. Such advantages make edge clouds promising for time-critical services, like the ones run on vehicles, drones, or sensitive production lines. Similar to traditional data centers, the Intelligent Workload Management system for an edge cloud has several functions: workload distribution strategy, dynamic resource allocation, scheduling middleware, etc.

There are two types of workload distribution strategies. Proactive strategies, like edge servers, edge devices, and Backbone cloud resource planning, try to deploy edge cloud services beforehand. Heuristic-based strategies build models from historical data to predict the execution distribution. Dynamic strategies try to remove idle workers while collecting more data from the source for scope expansion. They can balance the distribution process within the edge cloud according to the message transfer delay between the cloud and entities that generate data, as well as the processing speed of data at the edge. Choosing a middleware should depend on various criteria, like energy efficiency, whether device profiling is allowed, or whether edge cloud management information feedback is supported.

12.5.1. Workload Distribution Strategies

Workload distribution strategies are based on different criteria that affect how quantum workloads are distributed to the different edge devices and how these different workloads are managed over time. Workload allocation operates using the following basic strategies: 1) workload replication; 2) workload load balancing; and 3) workload co-allocation. Workload replication can be done in two approaches: intrusive and non-intrusive. The intrusive strategy involves replicating all the workload into all devices given that the devices have enough capacity and the management is able to intercept the results. In this strategy, every cloud edge service manages the portion of clients assigned to such service. The invoked task will publish the same output result to a common manager that will publish to the subscribers at a certain rate. In contrast, during non-intrusive distribution, the workload portion is replicated in the cloud service instead of inside the devices, and results are not guaranteed to be published at the same time or to be similar but the clients can accomplish their interactions.

The balancing strategy consists of minimizing the differences in load among devices often using association rules. However, this fails to consider time and does not guarantee service availability for a minimum number of users. This strategy is useful at times when accessing columns that are controlled by more than one device. Load balancing can be performed in the time sense, scaling tasks by limiting device connections on temporal columns associated with balance rules. Such actions ensure that devices will be available to all users and that suitable results will be obtained at the same time and that they will be similar in a certain range. Phased devices can be utilized in temporally reduced range intervals.

12.5.2. Dynamic Resource Allocation

Autonomous systems may engage local computing resources and begin parallel operations involving numerous micro tasks. So far as there is adequate processing bandwidth, task execution by these systems can be almost instantaneous. However, if operational demands exceed local computing capability, task execution will slow down. Delays in task completion can bring about undesirable consequences in mission accomplishment by the autonomous systems. It will likewise adversely impact the workload balance which intelligent workload management seeks to maximize. Therefore, it is imperative that the local computing load be continuously monitored through real-time data exchange between cloud and edge. Based on this load data, dynamic resource allocation will track surges in task execution delays, fixing those in a proactive way through the engagement of spare capacity at cloud resources.

To go deeper into dynamic resource allocation, a strategy for buffering micro task queuing at the edge can be considered. The micro tasks involved in these queuing scenarios are usually composed of stimulus data coming from mobile edge analytics at autonomous platforms such as airborne drones or federations of these. Real-time tracking of autonomous platform motion provides the locus of activity for mobile edge analytics. This involves the autonomous platforms circling over a common area for emitted data monitoring. For instance, suppose that these platforms are monitoring the emission of hazardous pollutants by an urban building. In this case, it is preferable for efficient pollutant sensing that drone circling radius be a small circle located just overhead the building or an aircraft flaring over the building edge. In such cases, the buffering of accumulated stimulus data is carried out for these horizontal buffer zones defined by flashing motion maneuvers of the autonomous platforms. Micro tasks are then dispatched for execution according to locally defined edge rules. These dispatches will be triggered by the availability of cloud processing resources.

12.6. Conclusion

This book addressed the new cloud trends, reflecting and analyzing the relevant technology scenarios. It presented the expansion to the Edge verticals, exploring the more flexible deployment of Distributed Cloud solutions – whose services can be located anywhere, from a Data Center to the proximity of Edge-based workloads – and the new Data-Driven Economy, the foundation of the Digital Transformation movement. The emergence of more intelligent solutions, the AI/ML-based services that embrace the promises of Autonomous Systems and Intelligent Workload Management, which enhanced the capability of managing the entire ecosystem, from Data Centers to Edge Gateways and Devices.

AI is an innovative and disruptive technology, bringing an enormous potential to almost any industry vertical ecosystems. However, its growth is severely impeded by the lack

of appropriate infrastructure. AI well-functioning is dependent on fast data collection and transfer for training the models, on appropriate platforms to run the model training processes, and on similarly powerful infrastructures to execute the inferences for the real-time actions. Clouds were designed to deal with these needs, but the more demanding latency requirements are hampering its role. The Edge model came in to solve this problem, but lacks capabilities. The solution is therefore the integration of Cloud and Edge infrastructures, creating a Distributed Cloud model capable of catering to the entire infrastructure. For this architecture to effectively work, it is paramount to have intelligent management capabilities, enabling the workloads management within the entire ecosystem – Cloud and Edge.

The promise of new technologies transforming our ecosystem is always there. However, hype alone is not enough to push their adoption and leverage their potential. Cloud computing started addressing the need for accepted and transparent foundations to support these new technologies in the first cloud wave. We are now seeing the arrival of new paradigms, bringing to our attention the requirements for a second wave. Will the coming years show the maturity and acceptance of this new wave, preparing Cloud infrastructures to support the new and exciting technologies?

12.6.1. Future Trends

Cloud computing has demonstrated its usefulness in the past decades by catalyzing the computing service economy and enabling many new synergies and applications. Today, it is at a crossroad - in order to support complex, near real-time workloads, very-low-latency, and very-low-power applications, as well as ultra-large infrastructure scale, we need to rethink the very basic tenets of several areas, such as the underlying infrastructure and data centers, and data and application management. It is clear, present-day central cloud computing is reaching the limits of its benefits, and it is time to create more advanced models of edge computing and intelligent source management. We envisage that in the following few years, there will be several innovations along three trends. First, we will see autonomous hyperscale cloud infrastructures, data centers and hardware technologies. Today’s cloud infrastructure is mainly assembled and operated by humans, albeit with heavy reliance on automation tools. Clouds are expected to be operating in a more autonomous and intelligent manner, with less human intervention in facilities operations, hardware orchestration, and fault diagnosis and recovery. Autonomously-managed clouds would allow better scalability, flexibility, performance, efficiency, availability, and reliability. With this trend, we expect to see breakthroughs in cloud-scale heterogeneous infrastructures, hardware-in-loop telemetry, near-zero-PUE cooling techniques, component-level fault tolerance and heartbeating, and acceleration-chip-modes-aware load balancing.

Secondly, we foresee the explosive growth of low-overhead distributed edge cloud systems for ultra-low-latency interactive workloads supporting transforming aspects of the economy, such as social, entertainment, commerce, smart city, and automotive. Over the years, the edge has been neglected in both innovation and investment. Each company invested in edge cloud capabilities for their own products - but the edge as a cloud environment has not matured. Most of the new capabilities are limited to specific ecosystems and brands. It is time for an open cloud edge environment where mobile and IoT-centric workloads, services, and products can flourish. With this trend, we expect to see breakthroughs in low-overhead single-board server products, transparent distribution, failover, and autoscaling of edge cloud services, and edge system interoperability at the hardware and software levels.

References

Thota, R. C. (2024). Optimizing edge computing and AI for low-latency cloud workloads. International Journal of Science and Research Archive, 13(1), 3484-3500.

Ibn-Khedher, H., Laroui, M., Moungla, H., Afifi, H., & Abd-Elrahman, E. (2022). Next-generation edge computing assisted autonomous driving based artificial intelligence algorithms. IEEE Access, 10, 53987-54001.

Ramamoorthi, V. (2023). Exploring AI-Driven Cloud-Edge Orchestration for IoT Applications.

Gill, S. S., Xu, M., Ottaviani, C., Patros, P., Bahsoon, R., Shaghaghi, A., ... & Uhlig, S. (2022). AI for next generation computing: Emerging trends and future directions. Internet of Things, 19, 100514.

Walia, G. K., Kumar, M., & Gill, S. S. (2023). AI-empowered fog/edge resource management for IoT applications: A comprehensive review, research challenges, and future perspectives. IEEE Communications Surveys & Tutorials, 26(1), 619-669.

Mishra, A. K., Ravinder Reddy, R., Tyagi, A. K., & Arowolo, M. O. (2024). Artificial intelligence-enabled edge computing: Necessity of next generation future computing system. In IoT Edge Intelligence (pp. 67-109). Cham: Springer Nature Switzerland.