

Article

AI and Computing Horizons: Cloud and Edge in the Modern Era

Nasif Fahmid Prangon ^{*,†}  and Jie Wu ^{*,†} 

Center for Networked Computing, Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA

* Correspondence: nasifprangon@temple.edu (N.F.P.); jiewu@temple.edu (J.W.);

Tel.: +1-267-666-5926 (N.F.P.); +1-561-809-2685 (J.W.)

† These authors contributed equally to this work.

Abstract: Harnessing remote computation power over the Internet without the need for expensive hardware and making costly services available to mass users at a marginal cost gave birth to the concept of cloud computing. This survey provides a concise overview of the growing confluence of cloud computing, edge intelligence, and AI, with a focus on their revolutionary impact on the Internet of Things (IoT). The survey starts with a fundamental introduction to cloud computing, overviewing its key parts and the services offered by different service providers. We then discuss how AI is improving cloud capabilities through its indigenous apps and services and is creating a smarter cloud. We then focus on the impact of AI in one of the popular cloud paradigms called edge cloud and discuss AI on Edge and AI for Edge. We discuss how AI implementation on edge devices is transforming edge and IoT networks by pulling cognitive processing closer to where the data originates, improving efficiency and response. We also discuss major cloud providers and their service offerings within the ecosystem and their respective use cases. Finally, this research looks ahead at new trends and future scopes that are now becoming possible at the confluence of the cloud, edge computing, and AI in IoT. The purpose of this study is to demystify edge intelligence, including cloud computing, edge computing, and AI, and to focus on their synergistic role in taking IoT technologies to new heights.

Keywords: 5G; AI; cloud computing; cloud service providers; edge computing; fog computing; IoT



Citation: Prangon, N.F.; Wu, J. AI and Computing Horizons: Cloud and Edge in the Modern Era. *J. Sens. Actuator Netw.* **2024**, *13*, 44. <https://doi.org/10.3390/jsan13040044>

Received: 1 June 2024

Revised: 25 July 2024

Accepted: 7 August 2024

Published: 9 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Computing, as well as artificial intelligence (AI), is under a fundamental shift that is driven by the emergence and integration of heterogeneous distributed computation paradigms—cloud, fog, and the edge—where the cloud is the orchestrator, fog is the middle layer, and the edge is the front-row seat for providing end users with cloud service offerings. These three specific models, with their kinds of properties and advantageous sides in the field [1], serve as the base foundation for the modern sort of tech progress, enabling the development and making of advanced AI and machine learning (ML) systems. The purpose is to understand the purpose and fundamentals of these three models, see their roles in the augmentation of cloud-focused AI and ML, and investigate how they interact with and complement each other to advance the field. Cloud computing, with massive storage space and formidable computational prowess, has made significant contributions to AI development. Due to technological developments, engineers can now operate with large amounts of data and accomplish complex tasks. Consequently, cloud-based systems are becoming an essential part of it all: modern AI and ML [2]. However, the rise of the paradigm known as the Internet of Things (IoT) results in demand for data processing in real-time and has uncovered some limitations of cloud computing, particularly in terms of latency and bandwidth constraints. The arrival of fog and edge computing signals a swift change in data processing, offering new answers to the issues created due to massive data and the need for processing in real-time [3].

Edge computing, located at the network's ends, comes out as an enticing solution to these restraints. Edge computing majorly reduces latency by handling data near its source, which makes immediate data analysis possible, and this is critical in time-sensitive applications. In the complex balance of cloud computing and AI, IoT serves as the dynamic partner, combining physical devices and digital data. The perfect harmony between IoT and edge computing is demonstrated by smart devices collecting and acting on data in real-time, helped by AI's analytical capabilities. This cooperation is vital for various applications, ranging from self-driving vehicles, which need immediate environmental assessments, to smart cities, with a myriad of sensors and actuators for optimized city management. The essence of this study is the incorporation of AI into these computing concepts, with specific emphasis on AI for Edge and AI on Edge. The term AI on Edge refers to the execution of AI processes right on edge devices, whereas AI for Edge refers to the deployment of AI models and algorithms in the central servers or upper layers to enhance edge computing capabilities. Previous works conducted on the edge–cloud continuum focused on many aspects of the domain. Researchers in [4] survey the current status of machine learning and data analytics frameworks, libraries, and paradigms enabling distributed intelligence across edge and cloud infrastructures. Challenges in ML workflow deployment upon such hybrid infrastructures are related to performance, reproducibility, and optimization of resources. The paper concludes by identifying open challenges in research and future directions toward the optimized deployment of AI workflows over heterogeneous edge-to-cloud environments. The paper [5] surveys the edge intelligence paradigm proposed as an alternative solution to the limitations of cloud computing for services supporting IoT. It provides a systematic analysis of the literature available on EI concerning definitions, architectures, essential techniques, and future research directions. More specifically, the present study attempts to provide an overall picture for both experts and beginners, showing the present state, challenges, and possible future improvements of EI. Survey [6] covers opportunities and challenges of integrating edge computing with cloud computing to come up with a DCCS. Then, it discusses how self-adaptive intelligence is required to manage the dynamic and heterogeneous nature of DCCS and how to use the MAPE-K framework. The paper identifies research opportunities and techniques that can help address the DCCS complexity and hopefully foster further development and collaboration.

In this paper, we give a fuller understanding of how AI not only enhances edge computing but also propels its evolution directly into a brand-new era of smart, independent systems capable of local decision-making. While edge computing and AI remain at the center of our concerns, attendant to them is the evolution of cloud service providers and their putative paths, particularly in terms of how their initiatives related to AI are remaking the horizon of the cloud. Based on our findings in the field, we provide an image for the framework to implement the concept of AI for Edge and AI on Edge on different aspects of edge cloud. Furthermore, we entail a closer look at the market strategies and technological advancement such service providers have undergone in pursuit of their AI-powered services. We deliver analytical insight into the changing scenario going on in cloud, fog, edge computing, and AI integration, whereby the authors focus on the significant role of IoT because it has sparked this revolution.

The paper outline is as follows. First, we define the basic concepts of cloud, fog, and edge computing and outline what makes them unique and their roles in today's computing infrastructure. We then focus on the convergence of AI with edge computing, considering topics from functional and application perspectives to eventually demonstrate how AI empowers edge computing with the capability for real-time, at-source data processing and decision-making. In the next section, we present the methodology, which includes a systemic literature review linked to this relevant study selection criteria. The second topic is the segmentation of cloud service into layers, with the contribution of each layer towards the architecture as a whole. Finally, edge intelligence introduces a new frontier: edge computing—we explore its importance and how it closes the gap between the cloud

and end devices. We explain the concepts of AI for Edge and AI on Edge, their distinctly played roles, their benefits, and each component of the framework. After that, we take the reader through the commercial cloud ecosystem, indicating key service providers and their AI-enabled offerings. Finally, we take a peek into the future by looking at ways in which this technology may evolve, therefore bringing a close to it by summarizing our findings and their implications for the field.

2. Methodology

The purpose of this study was to look into relevant information regarding AI integration into cloud and edge computing; we focused on conducting a systematic literature review. We had an exhaustive search approach across various academic databases such as Google Scholar, IEEE Xplore, and ACM Digital Library. The targeted keywords were, in particular, “AI on Edge”, “Edge Computing”, “Cloud Computing”, “AI for Edge”, “Edge Intelligence”, and “Cloud Service Providers”. Using boolean operators, attempts were then made to further develop the search by combining different search terms relevant to the subject matter, such as “AI AND Edge Computing” and “Cloud Computing OR Fog Computing”. All these searches were restricted only to publications written in the English language so that research is focused on text written in one language.

We carefully selected the inclusion and exclusion criteria to ensure that only relevant and good-quality studies were included. To observe recent trends, we included only peer-reviewed articles in journals or conferences that appeared in major technical reports within the last five years. Studies relating to integrating AI in cloud and edge computing systems or proposing novel methodologies or frameworks or that had contributed significantly to edge intelligence and AI-driven IoT applications were given a higher priority. Second, we included papers on cloud service providers and their AI-enabled services to understand the contributions from this area. We excluded non-peer-reviewed articles, white papers, unpublished theses, studies unrelated to the direct integration of AI in cloud or edge computing, and articles older than five years unless they were seminal works in the field.

The review process underwent multiple stages to ensure that the identified papers underwent a rigorous assessment. First, the titles and abstracts of the identified papers were screened for relevance to our study. Those for which the titles and abstracts did not correspond to our target inclusion criteria were excluded from further consideration. In the second stage, the complete text of each paper to be included based on the first screening was reviewed, which established a detailed assessment of the methodologies applied, findings presented, and relevance to our research objectives. Finally, the quality and contribution of each paper to the field were evaluated. Those offering high-impact findings or introducing new approaches made it into the shortlist of works to be included. Extracted data for each study included AI integration methodologies in edge and cloud computing, findings on performance improvement, latency reduction, energy efficiency, and applications that show practical implications of AI on Edge and AI for Edge.

We also included papers debating the services and technological advancement of large cloud service providers like AWS, Google Cloud Platform, and Microsoft Azure to see their impact on AI and edge computing. This was a systematic approach towards ensuring that all relevant literature had been read to provide a sound foundation for developing and evaluating our proposed AI for Edge and AI on Edge framework. Our careful process, hence, enabled us to retrieve relevant studies of the best quality for the research subject under investigation, ensuring that the results presented are robust and reliable.

3. Cloud and Extensions: Segmentation into Service Layers

The concept of cloud computing emerged as a service to allow users to rent computing services over the Internet. Cloud computing provides both the hardware and the software aspect as well as the systems in data centers that contribute to the service [7]. When the service, along with the resources, is being offered over an online platform to the people who use it, and the users compensate the service providers for availing these services, then

we consider it a public cloud. The infrastructure of cloud computing is made up of several components, like computing power, which comprises the scalable virtual machines in data centers that are needed for applications to run. The cloud also provides storage, serving up dependable, scalable data storage across distributed servers. Communication is a key part of cloud systems, whereby networking ensures efficient data transfer and communication between linked data centers and databases. In order to meet the rising demand to cater to more users, scalable database services for meeting the requirements of various data types are also provided as a service. Finally, management services, which consist of management and monitoring tools for cloud resources that primarily focus on security and compliance, are also among the offerings of cloud services. These components work as one to support the functionality and efficiency of cloud computing environments. The introduction of cloud computing triggered a potent shift in the way data are stored, processed, and managed, and this facilitated the delivery of scalable and elastic services over the Internet. However, the centralized nature of cloud services tends to often result in latency issues—especially for applications requiring real-time processing [8]. This certain challenge gave birth to the fog and edge computing paradigms, which bring computational resources closer to the data source, thereby reducing latency and bandwidth consumption. Fog computing serves as an intermediary layer between the cloud and edge devices; permitting computing, storage, and networking services to be delivered closer to end users. Furthermore, edge computing goes one step further by enabling data processing to take place at or near the source of data generation—like IoT devices. These paradigms have broadened the horizon of cloud computing and have provided a spectrum of solutions ranging from remote expanses of centralized cloud data centers to the immediacy of edge computing, catering to the rising demand for real-time, context-aware applications. Therefore, we see a new distributed computing ecosystem that fosters innovative cloud service models and a variety of services prioritizing proximity, immediacy, and efficiency [9]. The hierarchy and way of forming a cloud infrastructure are represented at a high level in Figure 1. The cloud layer consists of interconnected data centers through high-speed links. Fog is the layer beneath that serves a specific region that contains millions of nodes. Finally, edge serves the end users and devices residing in an area. Simply put, multiple edges are supported by a fog layer, and multiple fog layers are supported by a cloud layer, as depicted in Figure 2. In terms of cloud service offerings, they can be categorically separated into three categories: infrastructure as a service (IaaS) offers virtualized computing resources. Platform as a service (PaaS) provides a platform that enables customers to build, run, and manage applications without the hassle of infrastructure management. Software as a service (SaaS) provides software application services over the Internet on a subscription basis, therefore reducing the costs associated with managing resources related to running cloud applications [10]. In addition to these layers, there are sub-layers that are basically part of the three main layers. To enable event-driven computing, function as a service (FaaS) can be utilized within PaaS or SaaS. Backend as a service (BaaS) can complement SaaS applications by providing ready-made backend services like user authentication or database management. Container as a service (CaaS) can be used with IaaS environments to manage containerized applications more efficiently. Data as a service (DaaS) can augment SaaS applications by providing users access to external data sets and analytic services. Integration platform as a service (iPaaS) integrates the different SaaS, PaaS, and IaaS layer applications and services to streamline the workflow and data flow across different cloud services. These services and layers are marketed using different product names. However, the application and purpose of these layers remain the same at the core. Popular cloud service providers include the likes of Amazon Web Services (AWS) [11], Microsoft Azure [12], and Google Cloud Platform (GCP) [13], which is the public cloud infrastructure that is a part of Google Cloud and provides a plethora of cloud services and solutions across different layers and service offerings. Additionally, there are also plenty of open-source tools to implement cloud services and solutions locally on the user's premises using their own hardware; this is basically considered a private cloud, as the cloud is managed entirely by the user. In this

section, we discuss the cloud computing paradigm and evolution towards fog and edge to augment cloud services for the end users. We also discuss the key cloud service providers and the layers of their service offerings. In the next section, we discuss the evolution of the cloud in terms of AI and how the use of AI is revolutionizing the cloud paradigm.

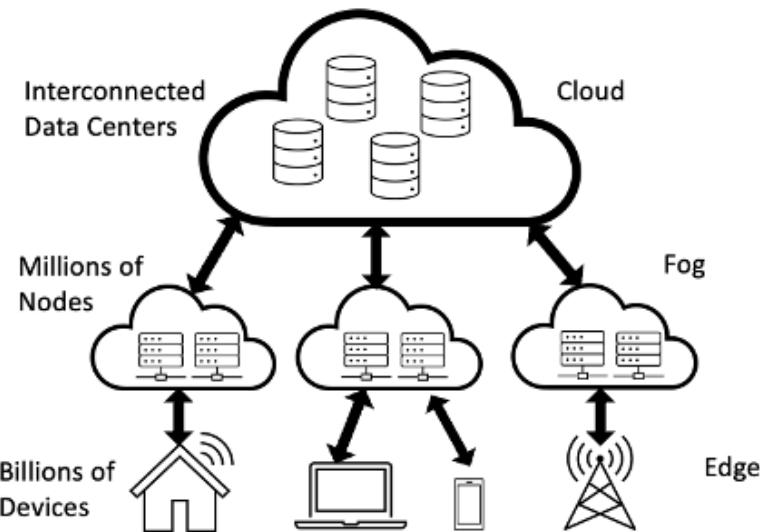


Figure 1. Hierarchy of distributed computing from cloud to edge.

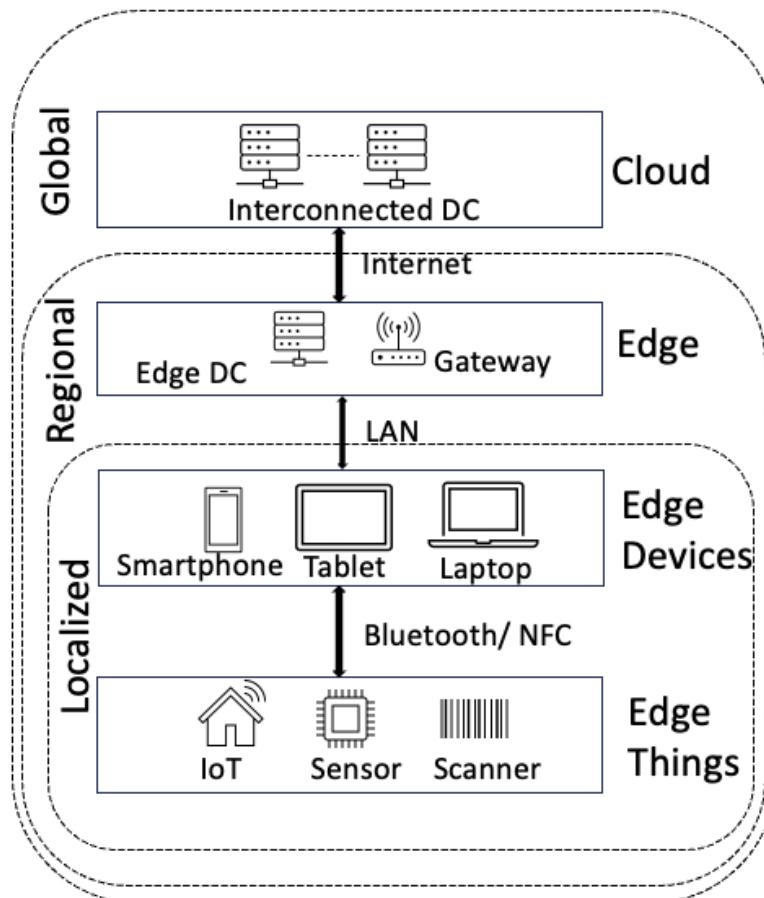


Figure 2. High-level view of cloud covering different zones.

4. AI and Cloud: How Cloud Is Becoming Smarter with AI

AI is transforming the cloud domain in the same manner as it has influenced various other sectors in terms of technological advancement. The use of AI has been practiced by both commercial service providers and researchers in academia, portraying the significance of the subject matter. From the commercial standpoint, AI is being implemented to improve various aspects of cloud services and their management. AI has been implemented in service automation, which results in increased productivity and reduced need for human intervention for different services. When it comes to resource management, AI algorithms are being implemented for allocating resources and scaling services based on user demands. To understand user demand and behavior, AI is being implemented for predictive analytics and provides more accurate forecasts that in turn help the services become more efficient. AI is also being implemented to detect security threats and attacks in cloud networks. With so much data being generated and stored in the cloud, AI is helping us visualize insights and provide major analyses of the data. This information, in turn, is helping us make better decisions and understand the users better. Thus, AI in the cloud is not only promoting a digital transformation in many industries but is also creating a smarter and more reliable self-governing cloud environment. All of this has been made possible because of the incorporation of AI into the cloud infrastructure. In addition to the interest in commercial cloud augmentation using AI, many researchers are also working on different aspects of the cloud and its challenges. Research is being conducted with regard to multiple aspects to make the cloud more efficient each day. Since the cloud is hosted on powerful hardware, energy consumption for the servers is a key field that can be made more efficient. There are many research works that focus on achieving energy efficiency in the cloud. For example, the research work discusses dynamic resource allocation in the cloud and task scheduling to distribute the load efficiently in order to achieve better power management in cloud data centers. AI is also being used to promote business growth for customers, as discussed above. Integrating AI models in the cloud infrastructure gives the users more visibility of their positions. Due to the rise of digital enterprise platforms that enable companies to rent out resources or services, the researchers in [14] focus on how AI- and ML-based technologies contribute to the innovation of business models and the changing dynamics of businesses. They also discuss the integration of AI in business models that are based on cloud systems. A single cloud hosts powerful computing resources, and AI is also being implemented in high-performance computing. The researchers in [15] investigate the role of AI in enabling high-performance computing and how this can transform the future of the cloud. From the security and privacy point of view, research work focusing on AI for the safekeeping of the cloud is also being practiced. Due to the distributed nature of cloud resources, fault tolerance is of key importance. The authors in [16] researched a cloud-based paradigm of predictive safekeeping using immediate information acquisition and utilization. The key argument of their study was to improve accuracy and reliability in fault diagnosis and maintenance scheduling, focusing on Industry 4.0, which is short for the Fourth Industrial Revolution. There is also a significant amount of research being conducted on cloud applications focused on the use of AI. Since we have a lot of applications that are not designed to be compatible with AI, application modernization is being researched by many. The authors of [17] stressed the need for advances in AI research to overcome the hurdles of cloud computing, like reworking large applications and catching runtime errors in application behavior. AI is also bringing changes to security. Load balancing and resource distribution for scaling are also a topic of interest for researchers. The study [18] looks at the use of different AI predictive techniques for proactive resource scheduling deployment in cloud, fog, edge computing, and networking systems. From the discussion above, we can see that the cloud is being transformed with AI by improving its efficacy in different aspects and service layers. This research is being pioneered in parallel by both the industry and researchers, which signifies the pace at which the cloud industry is transforming with the adoption of AI. In the next section, we discuss AI in a specific cloud paradigm, the edge, and learn about edge intelligence and its significance.

5. Edge Intelligence: The New Frontier

As discussed in Section 1 of this paper, in order to integrate sensor networks efficiently with the cloud, new paradigms like edge and fog have been introduced. These new paradigms have redefined the way data are processed and managed in the realm of distributed computing, bridging the gap between the cloud and end devices. Edge computing permits data to be handled at or near its origin, reducing dependency on distant cloud servers. Because of the proximity to data sources, latency is reduced, bandwidth efficiency is increased, and real-time data processing capabilities are improved. This is critical in applications such as IoT, smart cities, and autonomous vehicles. Edge computing can be further extended to multi-access edge computing (MEC), formerly mobile edge computing, which is a subset of edge computing that specifically targets telecommunication networks and infrastructure. MEC brings computing resources closer to mobile users by directly embedding processing power within the cellular network infrastructure. This integration enables mobile applications to process data more quickly and efficiently, significantly improving user experiences in mobile environments. Edge computing and MEC both demonstrate fog computing principles by extending cloud capabilities to the network periphery, resulting in a more decentralized, responsive, and agile computing framework suitable for today's increasingly interconnected and data-intensive world. In this section, we discuss how the edge has introduced unique technologies that bridge the gap between end users and the cloud. We further extend the discussion to how AI is transforming the edge and creating an intelligent edge.

5.1. Edge as an Extension to the Cloud

In the following, we discuss the unique features and characteristics of edge computing. This helps us understand how the edge facilitates end users' integration into the cloud ecosystem. The discussion is focused on how communication and data are handled by the edge. We describe different communication protocols used by the edge and data management tools that are being utilized in edge to serve the end users and reduce the reliance on the cloud.

5.1.1. Communication in Edge

The edge layer comprises edge data centers with moderately powerful computing resource that directly connect to end devices and users. What makes the edge unique is that the devices it connects to are heterogeneous in nature. Most of these devices, like mobile devices, sensors, and smart home appliances, have limited computing resources and energy. Therefore, achieving efficient communication between the edge and these devices poses a challenge. Edge sensors are of a small form factor that houses communication modules that have limited range. Furthermore, they have limited energy sources: they are mostly powered by batteries. Therefore, the communication protocols used by these sensors are more focused on efficiency while utilizing the energy available. One of the most popular protocols is Message Queuing Telemetry Transport (MQTT), which is a messaging protocol. This protocol is well-suited for compact sensors and mobile devices because of its minimal power usage and small-sized data packets. End devices that do not require constant communication with the edge utilize Constrained Application Protocol (CoAP), as it is well-suited for constrained nodes that have limited connectivity windows. This protocol allows the devices to store the data locally and share them asynchronously with the edge servers when connectivity is available. For secured communication that handles sensitive data, Advanced Message Queuing Protocol (AMQP) ensures reliability and security among the edge and devices by implementing authentication, authorization, and encryption of data. For edge devices that require real-time data processing and streaming services, Data Distribution Service (DDS) is widely used. This protocol ensures quality of service (QoS) and guarantees the reception of data packets, making it suitable for real-time applications. The researchers from [19] discuss more detailed implementations and functionality of these and other protocols for edge communication.

5.1.2. Data in Edge

Collecting, storing, and processing data generated by the end devices is a prime objective of the edge servers and devices. To serve low-latency applications hosted on edge devices that have limited energy and connectivity to remote edge servers, edge databases are built to function without constant reliance on the core cloud. The edge servers process the data locally using the computing resources available at the edge. MEC also enables the utilization of resources used by the telecom providers to be used as an extension of edge, which also helps process data without relying on the cloud. The processing of data is not the only key factor considered in data management across the edge: there are other factors such as risk and security. Edge devices that produce sensitive data have to be isolated from other devices and layers to ensure the privacy preservation of user data. While maintaining data isolation, their synchronization among devices is equally important to ensure reliability across the system. Data replication and synchronization and the risk associated with it have been researched from the early age of distributed computing [20] and are still a key research topic in the field of edge database systems. These edge database systems are built on different platforms suited for different purposes. Since the edge servers are not as powerful as the cloud, SQLite is an ideal serverless and self-contained database, making it suitable for the resources. Relational databases like MySQL and MariaDB are ideal for applications that require structured data management, which is suitable for handling complex queries. For secure data handling and data integrity, PostgreSQL is used, as it is robust and supports advanced SQL features. Non-structured databases, on the other hand, are more suitable for building flexible and scalable systems that can adapt to dynamic network requirements. MongoDB is a database platform that stores data in JSON-like form. Couchbase and Cassandra are also popular solutions focused on high availability and scalability, with Cassandra being very popular in distributed edge computing environments. These edge database systems are critical components of the robust and responsive data management ecosystem needed for edge computing environments.

5.1.3. Data in Upstream and Downstream Applications

A massive amount of data is communicated across various layers using the communication protocols and database platforms discussed previously. Data flow in the edge paradigm is bidirectional, which means data can flow from the cloud to the end, defined as downstream data, and from the devices to the cloud, which is called upstream data. Data management is critical in upstream applications within edge computing environments and can be segmented into different classes. Redundant or short-term irrelevant data, such as the last five minutes of temperature readings, are distinguished from long-term valuable data, such as the average temperature over a week, or critical short-term data that trigger immediate actions, such as turning on a heater when the room temperature falls below a certain threshold. Proper labeling of these data is important, since the decision as to where and how to process them is based on the labels. Downstream data deal with providing feedback or decisions to lower levels and assist them with their computing processes. Downstream applications are concerned with optimizing data delivery and service responsiveness. There are many methods to make downstream data handling more efficient, like caching mechanisms that are used to speed up data access. As we already know, edge services prioritize latency-sensitive tasks to improve performance; therefore, proper planning and integration of communication strategies and database management systems improve data flow from and to the edge, thereby improving the overall system efficiency and user experience.

5.2. Edge Intelligence and Benefits

This section looks into the concept of intelligent edge and highlights the corresponding AI in edge's effect on performance, service cost reduction, and privacy against efficacy. The incorporation of AI into the cloud computing regime represents a recent paradigm shift that transforms the technology's many core components. It combines the rapid

evolution of computational intelligence and cloud infrastructure's massive capacity to cut costs, automate resource management, boost system reliability through predictive analytics, and guarantee new flexibility in privacy secured by new security sensor data insights from the system. AI will make the edge more independent while also significantly enhancing it into a robust and versatile intelligent seat. Hence, the edge will evolve from an organization back into a system that is more user-centric and is able to provide customized solutions that dynamically match ever-changing user demands and needs. AI in edge has influenced different aspects of the service. The study [21] verifies that AI has pervaded the workings of edge computing in terms of performance, cost, privacy, efficiency, and reliability. It enhances the reduction in computational overhead on the resources when low computationally intensive AI models are implemented at the edge. This results in easier and more rapid decision-making and data handling. It is quite important for applications requiring low latency, such as autonomous driving, live health monitoring, and periodic data analysis. It has been proven by applying efficient AI models to process the resulting data. In turn, this results in less energy used in computation, better device lifetimes, and improving system performance. AI running on edge computing allows one to reduce costs associated with data transfer and cloud resource usage. It makes data processing locally more reliable and decreases the necessity for sending large amounts of data to the cloud—not only in terms of saving associated costs of data, but also bandwidth. Resource placement, task scheduling, and offloading planning with AI will greatly reduce the cost of communication with end devices at the edge. In most cases, the application and user data is the major concern that has to be maintained in terms of privacy. Edge intelligence can play a crucial role in this application domain. In fact, processing sensitive data right there on edge devices as opposed to the need to transmit them over to cloud servers sharply reduces the risks of data breaches and violations of privacy. This gains high significance in scenarios such as health care and finance, where personal information needs stringent security. More so, federated learning (FL), which bases decisions on the local dataset, can avoid data breaches and help with privacy maintenance because decisions can be made without recourse to any central server [22]. AI optimizes the performance of edge devices since, in other words, it increases their operational efficiency and effectiveness. For instance, AI algorithms can be used to dynamically adjust the computational burden or energy consumption of a given device in response to the prevailing demands, thus conserving energy and extending the lifetime of the device. Reliability can also be ameliorated through AI for Edge computing systems. It can forecast potential system failures and work in advance on their mitigation by spotting patterns in data that could show some anomalies, which might spell trouble. This proactive approach can reduce downtime and guarantee continuity of operation—issues that are very critical in industrial automation and, more so, in smart city infrastructures, which are always considered mission-critical applications.

6. AI for Edge

The discussion in this section considers how the other layers help the edge layer with decision-making, data processing, and other such assignments. In line with the above discussion, one of the great contributions of AI to edge with the purpose of making it possible to perform better is introducing optimization tools that are specifically designed to improve the efficiency of edge. The work presented by the researchers in [23] discusses various concepts and tools to make the edge more efficient regarding performance. The reason is that the tools are trained based on data from past experiences, which can prove useful since they can help learning to be facilitated in the edge layer toward a better environment. This way, the edge devices are enabled to process data locally, without relying on any heavy hardware, thereby nullifying the need to constantly shuttle big volumes of data to and from the centralized servers. This eliminates processing time and, most importantly, reduces latency to enable real-time decision-making. In many cases, such efficiency gains can mean the difference between life and death: for instance, in autonomous vehicles or emergency response systems. A classic example is the supervised learning ML technique, which

finds use in cloud computing. For instance, stochastic gradient descent does not get the parameters right on the first try, but it iteratively modifies them toward a direction that tries to minimize the loss function required to train neural networks. In other words, supervised learning may take place on the device only if the cloud allows the device to learn and adapt to new data in real time. The model generated by SL in the cloud may be shared with the edge, thereby reducing constant backend communication for decisions: for instance, when it is important for facial recognition in security or to inculcate a much more personalized touch to retail suggestions. Edge AI also utilizes reinforcement learning, on which its base is built, akin to how humans learn from the environment. RL algorithms learn to make decisions by the interaction of actions with their feedback. Such a method becomes very handy in dynamic systems marked with changing states: for example, a traffic management system that requires the algorithm to learn real-time conditions. Recent advances in AI also include the multi-arm bandit theory and multi-agent learning, which provide a decision framework with choices bearing uncertain rewards. This, for example, optimally decides which data that the application should process locally and which to send over to the cloud on edge computing so that performance is optimized without saturating the bandwidth. A deep Q-network is a form of reinforced learning technique that unifies Q-learning with deep neural networks. DQNs guide making elaborate decisions by pitting numerous possible actions against the consequences one should expect from them. An example of using these is aiding devices at the edge to take intelligent, autonomous decisions over huge data sets; an example is the predictive maintenance of industrial machinery [24]. The impacts of AI for Edge can be divided into multiple sections, as depicted in Figure 3. These sections represent “how AI implemented outside edge layer impacts edge layer”. In these layers, the different tools and methods discussed above are implemented. The topology section is the base for an AI for Edge architecture. It includes the physical and network design acting as the backbone of the edge computing infrastructure. AI for orchestrating the edge sites will provide optimization for data acquisition and proper network planning, where instant changes in resource allocation are made to match the predicted network demand using ML-based algorithms. On the other hand, AI in network planning and predictive maintenance will really help improve performance in wireless networks. The content section focuses on data efficiency. AI makes this section better through the intelligent provisioning of data from remote cloud services and those located at the edge. MAB and RL help place services, allowing for dynamic decisions on the placement of services for optimal performance [25]. Fast design of the caching pool becomes a predictive task, since, based on the prediction of traffic, AI can analyze the usage pattern so as to cache content intelligently in such a way that reduces latencies and congestion. The maximum AI impact is noted in the service section. The Lyapunov optimization techniques balance system stability with resource utilization to ensure that decisions on DQN computational offloading are both effective and efficient at the same time [26]. Therefore, AI now governs the best time and nodes for such computational offloading in order to better the response time and lessen edge device energy consumption. It is primarily not about independent optimizations in each section but striving to create a coherent system wherein the AI capabilities in each section can inform and power the others. For instance, the AI-based data acquisition in the topology section has impacts on decisions regarding computational offloading in the service section. In the same way, strategies for service placement in the content section can impact network planning in the topology section.

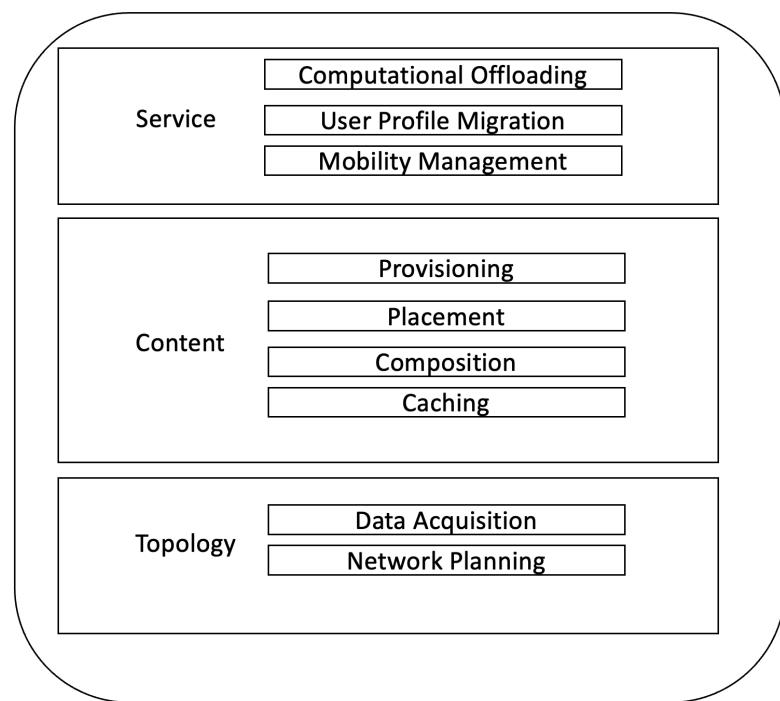


Figure 3. AI for Edge implementation framework.

7. AI on Edge

AI on Edge is the topic that discusses how AI is implemented directly at the edge. Edge computing heralds a versatile and flexible deployment platform for AI over a wide variety of scenarios linked to infrastructure setups, where this AI can take the lead in performance and function increases. The smart features of edge computing derive from the diversity of sources, such as data from networked vehicles in smart transportation, smart devices in a home, and other types of smart city infrastructure—all part of the rich tapestry of data that is ready for input to let AI work its magic. AI at the edge can be implemented at a couple of different levels of abstraction, both in software and hardware. AI-compatible chips are highly necessary for this ecosystem to be successful. GPUs, TPUs, and NPUs are processors that are made to accelerate AI workloads in their computations. Making instantaneous analytics and decision-making possible at the edge is derived from the power of rapid computation of sophisticated algorithms through the use of these AI chips. Enabling AI to work in perfect integration with scenarios where data processing cannot afford any latency is breaking the classic barriers of computation. This synergy between AI and edge computing breeds innovation towards smarter, better-reacting technologies that are more aligned with the fabric of our environment. The layered approach to how AI can be implemented on edge makes the integration of edge computing with AI powerful in three quite different sections that are interlinked to serve different roles. Each has a specified role at the architecture level, where AI is put into place at the edge directly to contribute to increased system efficiency when it comes to matters of privacy and improved performance. The first layer in Figure 4 focuses on model adaptation for AI models in edge conditions. FL enables the conditioning of the models on training and inference frameworks, which would otherwise leak user data. However, it does so without leakage, as mentioned in [27]. Model compression through quantization, dimensionality reduction, and pruning techniques is core to the cost reduction associated with computation and increased robustness while maintaining system performance. The toolset applied in this section discusses that the AI models created follow the rules for being lightweight and efficient, meeting the diverse requirements of certain scenarios included in edge applications. The following layer is centered on the framework design that will host such adapted models. Privacy under

deep neural networks (DNNs) is kept in very high order because of how FL can split data, and also, the model is implemented in a fashion whereby data confidentiality for the data sent over the nodes is preserved [28]. The key points of this layer include frameworks for AI workload partitioning across the edge and core, ensuring that model inference is performed efficiently, and support for task splitting between the edge and core to optimize low latency and bandwidth usage. Processing accelerator tools center on processor acceleration [29], answerable to the want of AI computations through hardware support in designing DNN-specific instruction sets and leveraging highly parallel computing paradigms. The system equally reduces latency and boosts speeds by bringing computation close to data through near-data processing. Therefore, it offers real AI-time processing capabilities at the edge. Using hardware accelerators like GPUs, TPUs, and NPUs brings sufficient computational power for serious AI tasks right at the data source. Together, these three sections build a powerful AI-enabled edge computing architecture—each layer builds on the previous one, ensuring that AI on Edge is not just a fad but is practical and effective. Issues around this layered architecture involve practical concerns: privacy, cost, and real-life performance, in addition to the specific requirements of AI and edge computing.

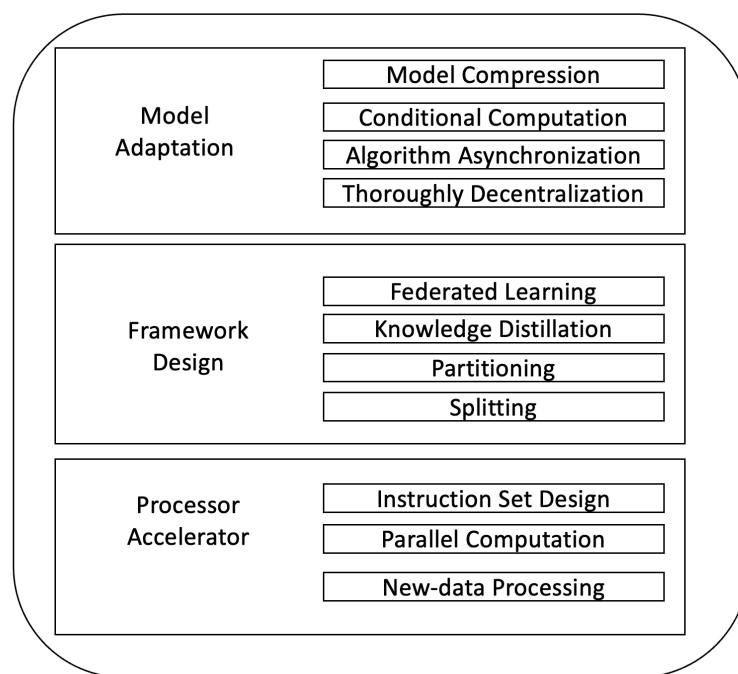


Figure 4. AI on Edge implementation framework.

8. Navigating the Commercial Cloud Ecosystem

This section focuses on the commercial cloud and its service providers to get a real-world overview of the industry. With the emergence of cloud computing's largest players—GCP, AWS, and Microsoft Azure—the world of technology has become totally revolutionized. Amazon Web Services lit the beacon of the cloud revolution sometime back in 2006. It gives scalable and affordable computing, storage, and analytics facilities to businesses of varying sizes. Not long after Amazon, Microsoft launched Azure in 2010, effectively extending its enterprise knowledge into the cloud. Meanwhile, Google Cloud has leveraged its ginormous infrastructure that was built initially to bolster its search engine and YouTube to set its foot in the cloud market in 2011. These cloud mammoths are no longer merely providers of storage and computing services. Actually, now they are all-encompassing platforms instigating innovations across a number of industries. AWS remains at the top of the market with a wide range of services and tools aimed at machine learning, quantum computing, and IoT. Azure consolidates its strong position by integrat-

ing and handling Microsoft's software offerings, like enterprise solutions, and placing a careful focus on hybrid cloud, AI, and enterprise services. Talking about its data analytics and machine learning skills, Google Cloud is using AI to transform industries, solidifying itself as the AI-first cloud platform. These cloud service providers are continuing to innovate. AWS is stretching the limits of cloud innovation with a forever-expanding suite of services, which are a strong foundation for the rising generation of cloud-native businesses [30]. Meanwhile, Azure is channeling heavy investment into AI and edge computing, looking to democratize AI and giving power to both developers and organizations [31]. Google Cloud is putting more chips into AI [32] and quantum computing, betting big that this advanced tech will redefine the possibilities in the cloud.

8.1. Cloud Service Layers

The cloud services landscape is thus enriched with large diversity; key market players offer a wide range of services to satisfy different requirements and technical needs. The following provides us with an overview of three principal service model offerings, as organized in Table 1.

- IaaS: This involves the fundamental pieces of cloud computing and generally deals with networking capabilities, hardware (either virtual or dedicated), and storage space for data. In this layer, AWS provides services such as the Amazon Elastic Compute Cloud (EC2) for computing capacity, Amazon Simple Storage Service (S3) for scalable storage, and Amazon Virtual Private Cloud (VPC) for isolated cloud resources. All Google services, likewise, have a counterpart for robust and scalable computing choices: Google Compute Engine, Google Cloud Storage, and Google Cloud Virtual Network. In Microsoft Azure, the services available are Azure Virtual Machines, Azure Blob Storage, and Azure Virtual Network.
- SaaS: This layer represents end-user applications that are exposed through the Internet. Instant deployments are offered by AWS, GCP, and Azure; some of the temporarily similar services that they offer are Amazon Chime for communication, Amazon Work Mail for email, and Amazon Connect for setting up contact centers. GCP provides Google Workspace for productivity, Google App Engine for app hosting options, and Firebase for developing mobile and web apps. Microsoft Azure is in step with Microsoft 365 for productivity and collaboration, Azure Active Directory for identity services, and Azure Communication Services for building rich communication experiences.
- PaaS: This layer interacts with infrastructure that is responsible for providing developers with a base on which to deploy and, even further, be responsible for the governance of their applications. AWS provides services in this layer to include AWS Elastic Beanstalk for easy deployment of apps, AWS Lambda for serverless computing, and Amazon RDS, coupled with Amazon Redshift, to provide a fully managed, petabyte-scale data-warehousing service that can automate tasks associated with provisioning, configuring, securing, scaling, and self-healing of a data warehouse. An example is Amazon, where the integrated analytics service vision has been set to proliferate data processing across warehouse and big data systems with GCP. The other services provided are managed, like the fully furnished platform known as App Engine, Cloud Functions for event-driven computing, and Cloud SQL for managed database services. And, finally Microsoft Azure provides Azure App Service to host applications, Azure Functions for serverless computing, Azure SQL Database for managed databases, and Azure Logic Apps for application integration and workflows.

Table 1. Popular offerings by cloud service providers based on layers.

Vendor	Layer	Offerings
AWS	IaaS	Amazon EC2, Amazon S3, Amazon VPC, Amazon EKS
	SaaS	Amazon Chime, Amazon WorkMail, Amazon Connect, AWS Marketplace
	PaaS	SaaS Subscriptions
GCP	IaaS	AWS Elastic Beanstalk, AWS Lambda, AWS RDS, AWS Fargate
	SaaS	Compute Engine, Google Cloud Storage, Google Cloud Virtual Network, Google Kubernetes Engine
	PaaS	Google Workspace, Google App Engine, Google Cloud Identity, Firebase
Azure	IaaS	Google App Engine, Cloud Functions, Cloud SQL, Cloud Run
	SaaS	Azure Virtual Machines, Azure Blob Storage, Azure Virtual Network, Azure Kubernetes Service
	PaaS	Microsoft 365, Azure Active Directory, Azure Communication Services, Azure Virtual Desktop
	IaaS	Azure App Service, Azure Functions, Azure SQL Database, Azure Logic Apps

8.2. Cloud Service Providers and Their Services

These large cloud service providers are critical to the very value of providing certain tools that allow businesses to be more innovative and transformative in their business process within the AI domain. The biggies have created formidable silos of service in AI; among these are machine learning platforms, cognitive computing capabilities, and advanced analytics—capabilities that can support a fresh order of applications from natural language processing to intricate data pattern recognition. AI for every player with AI-enabled tools enables business opportunities with strategic competitive advantages and operational excellence, as shown in Table 2.

Table 2. Tools specialized for AI and edge by popular cloud service providers.

Application	Amazon Web Services	Google Cloud Platform	Microsoft Azure
Platform	Amazon SageMaker	AI Platform	Azure Machine Learning Service
Image Analysis	AWS Rekognition	Google Cloud Vision API	Azure Cognitive Services
Deep Learning	AWS DeepLens	AutoML	ML.NET
Edge AI	AWS IoT Edge	Cloud IoT Edge	Azure IoT Edge

8.2.1. AWS

Some of the key tools from AWS include:

- Amazon SageMaker is a fully managed service for building, training, and deploying models with ML. It allows developers to quickly create models and scale them up to the cloud's full capacity without being hampered by the undifferentiated heavy lifting traditionally inherent in hardware.
- AWS DeepLens is the world's first video camera to use deep learning and was purpose-built for developers. It enables the ability to gain hands-on experience in AI but without using ready models in SageMaker or custom model construction, hence leading to practical experimentation with AI.
- Amazon Rekognition supports powerful image and video analysis so that developers can include image recognition features in their applications with no prior in-depth knowledge of ML or computer vision.
- Amazon Lex: Utilizing the service of Amazon Lex, one can create conversational interfaces, both voice and text, in any application. It is highly advanced deep learning technology that underlies the Amazon Alexa services. It same comprises capabilities of deep learning for automatic speech recognition and natural language understanding in a very sophisticated manner [33].

8.2.2. GCP

The tools from GCP include:

- AI Platform is a whole-pack tool to host and run ML models, from their conception and training to actual usage in production, and it also incorporates support for inferencing tailored for different types of ML frameworks.
- TensorFlow: This is an all-in-one ML framework that is widely popular in the open-source environment and is equally synonymous with versatile development. It is a product pioneered by Google that now has many developers working on it worldwide; it uses deep learning and neural network capability.
- AutoML: Google's AutoML takes the guesswork out of machine learning with its completely automated training and deployment of the model. It is appropriate for both experienced practitioners and those new to ML.
- Google Cloud Vision API analyze images on the cloud, thus deriving insights through image recognition capabilities. It enables applications to understand the content of an image without doing any processing on the device itself [34].

8.2.3. Microsoft Azure

- The tools provided by Microsoft Azure are the following: Azure Machine Learning Service is a managed cloud service provided by Azure that enables developers to train, deploy, automate, and manage ML models. It is developed for agility and has tools straddling the entire lifecycle of machine learning.
- Azure Cognitive Services: These allow developers to use a suite of APIs, as well as services, to build functionalities into applications that may relate to cognitive computing or artificial intelligence, such as computer vision and natural language processing.
- ML.NET is an open, cross-platform machine learning framework designed for .NET developers. ML.NET democratizes machine learning, bringing the established benefits of repeatability, transparency, and interpretability into the hands of .NET developers using the set of toolboxes.
- Azure Databricks is an Apache-Spark-based analytics platform with a perfectly optimized environment for Azure, allowing collaboration and big data processing in the support of ML activities [35].

8.2.4. Other Key Cloud Service Providers

Apart from the established giants of this field, such as Amazon Web Services, Google Cloud Platform, and Microsoft Azure, the cloud computing landscape actually further comprises a number of other key firms that offer quite a wide range of services and capabilities. Among the popular names within this area, IBM Cloud has a strong suite of offerings that also come with features for AI, ML, and IoT; it is especially commended for its emphasis on solid enterprise solutions and security [36]. Another major provider in the industry is Oracle Cloud itself, specifically recognized for its database services and other fully functional cloud functions in various areas of enterprise operations: application development, data management, and business analytics [37]. VMware Cloud is a platform that offers businesses unique solutions in virtualization and the use of cloud infrastructure by organizations looking to move or extend already-existing on-premises or data center infrastructures onto the cloud [38]. Another interesting player is Digital Ocean, which has gained popularity for its sheer simplicity and developer-friendly approach; it offers cloud services that are particularly appealing to small businesses and startups. Others become pioneers in their own regions for offering these cloud services. Notable examples market-wise come from Asia, mainly from China, due to their regulations. These Chinese companies unite their efforts in the provisioning of cloud services to regions with colossal user bases. One of these major players is Alibaba Cloud, which belongs to the Alibaba Group and is one of the largest cloud service companies in Asia; it offers variations in cloud services. It has made commendable progress with AI and e-commerce solutions. This scalability and reliability go into serving small as well as large companies [39]. On the other hand, Salesforce Cloud

is identified with the CRM provisions; Salesforce has a vast range of applications that are cloud-based for sales, services, and marketing. Tencent Cloud is identified with Tencent and offers a vast array of cloud services, ranging from cloud hosting and AI to big data analytics. It has been deeply integrated into the host of applications and services by the company and includes social media platforms, gaming, and online payments [40]. Huawei Cloud: The cloud service unit of this giant in telecommunication avails various cloud services like cloud servers, cloud databases, and AI services. Huawei features a very solid cloud infrastructure with corresponding proper security implementations; hence, it is the core of the international telecommunications services in this company [41]. Often referred to as the Google of China, Baidu presents Baidu Cloud [42], which is broadly similar to a range of cloud computing services. The very strong AI and ML services in Baidu Cloud are from immense previous experience in search engines and AI research by Baidu. Other emerging CSPs, apart from the three key CSP firms, have also been at the forefront of innovating infrastructure with solutions targeting challenges for their own research and development facilities.

8.3. Cloud Services Focusing on Edge Applications

There are some unique properties and demands associated with edge computing as a paradigm of cloud computing. This is the topic of our next section. We then give a general overview of the services most targeted by CSPs to capture the edge market. Examples of such cloud service providers include Amazon Web Services, Microsoft Azure, and Google Cloud; together, they provide a collection of edge services and tools that extend cloud capabilities to the edge of the network in a manner that improves the performance and response time of the intended end users. For example, AWS Outposts deliver AWS cloud services on-premises with processing and storage options in order to serve applications requiring low-latency responses or compliance related to data processing. AWS Local Zones extend AWS infrastructure into geographic locations where people have significant populations and need very low-latency workloads. Designed for low latency, AWS Wavelength extends infrastructure to the edge of the telecom carrier network, thus enabling applications on mobile and connected devices to benefit from ultra-reliable and low-latency communications. The Azure IoT Edge service and many others enable built cloud solutions to be extended down to the edge. This provides computing, storage, and advanced analytics at the edge by using controlled devices. Specialized services like the Azure Data Box help with data transfer at a massive scale into Azure, while the Azure Network Function Manager helps deploy network functions at the edge. Azure Sphere allows device connection to IoT, while Azure SQL Edge supports databases at the edge. Azure Front Door facilitates additional protection from threats. Distributed Cloud Edge is, in turn, a service by Google Cloud. It is a managed hardware and software stack that enables customers to run applications at the edge. The Edge TPU is used to accelerate the machine learning interface for devices on the edge. Extended cloud functionalities for edge devices are provided by Cloud IoT Edge. Cloud CDN, on the other side, is responsible for delivering content in a performance-optimized way that minimizes latency. It also provides network edge services for network solutions, along with Google Kubernetes Engine for the management of containerized applications.

8.4. Open-Source Cloud Solutions

One significant development of the modern cloud computing environment is the open-source nature of different cloud tools that have been developed and delivered under an open-source license. There is no associated licensing charge, and there is freedom to modify, distribute, and freely use the available tools. Open-source cloud tools take care of a large number of features and might be of help in different parts of cloud computing, such as cloud storage, infrastructural management for the cloud, and platforms that lead to the development of cloud-based applications. Notable examples include Kubernetes, Apache CloudStack, and OpenStack. The reason for the great use of these tools is mainly in business

and in development through the benefits that they bring: particularly, affordability and customizability, which mean an easier platform for research and innovation. There are quite a number of widely accepted and feature-rich open-source cloud computing tools, where Kubernetes is celebrated for orchestrating containerized applications [43], and OpenStack ranks as the favorite for building private clouds [44]. On the other hand, Envoy wears the crown for networking, while Apache Zipkin sees broader adoption for distributed tracing and Jaeger for tracing in Kubernetes environments [45]. Prometheus guarantees the monitoring and alerting of cloud applications [46]. Terraform was developed by HashiCorp and is identified by its possibilities for infrastructure as code. Grafana is oriented towards metric visualization and tracking. Eucalyptus allows the building of private and hybrid clouds that are AWS compatible [47]. Apache Mesos allows efficient job control in distributed environments [48]. GitSecret allows structuring and encryption of secrets on a Git repository. They are making cloud computing environments more effective to meet needs from orchestration and monitoring to infrastructure management and security. This debate is thus quite complicated at present, being that it has its roots in the philosophies of innovation through collaboration and control; that is to say, whether cloud technologies should be open-source or proprietary is still subject to debate. Open-source cloud technologies do provide a collaborative environment, with developers from all corners of the globe given an opportunity to contribute to the code; this automatically drives innovation and development. The model actually encourages a high level of transparency, security due to much scrutiny, and flexibility in ways that organizations can fine-tune solutions to their specific requirements. Open-source mitigates the chance of vendor lock-in even further, based on the increased degree of freedom and adaptability to changes in the technological landscape. On the other hand, these proprietary cloud technologies carry their own share of returns. Especially in terms of specialized support and controlled released updates, companies running proprietary clouds can deliver dedicated customer service. To ensure features that are dependable any time some level of assurance is needed, companies with proprietary clouds can ensure dedicated customer service that suffices to meet the needs of businesses. Their features allow a business to protect its intellectual property in order to remain competitive by providing situational awareness. This showdown is representative of something greater in the tech industry: a conflict between cooperation and competition. Proprietary systems place more emphasis on individual inventiveness and monetization schemes, while open-source models emphasize group expansion and availability. The real decision to be made between propriety and open-source technologies lies with requirements, assets, and tactics, plus the very different views various firms take on the issues of cooperation, security, and innovation.

9. Paving the Future

Cloud, fog, and edge computing have a dynamic and quickly changing future, particularly when it comes to AI and ML applications. This section discusses the current trends and also draws future outlines for these technologies based on the current research scene. We separately discuss different paradigms and end with regulatory and ethical concepts comprising all the paradigms. The same has been summarized in Table 3.

9.1. Cloud Computing

For businesses and research institutions, commercial cloud computing will enable the use of the computational performance and scalability necessary to run intricate AI and ML models. Implementing AI models would be simplified by requiring more sophisticated levels of AI services on cloud platforms, where hybrid cloud solutions enable increased flexibility and optimization in the deployment of AI/ML. This trend is expected to extend to more levels of complex AI implementations, such as extended reality, which has been discussed in detail by the research work in [49].

Table 3. Future of AI and cloud.

Category	Aspect	Focus
Cloud Computing	AI/ML Integration	Enabling sophisticated AI/ML model deployment with high performance and scalability.
	Hybrid Solutions	Offering increased flexibility and optimization for complex AI implementations.
Fog Computing	AI Convergence	Driving real-time data analysis, decision-making, and system efficiency.
	Industry Impact	Promoting intelligent IoT applications and responsive networks in various sectors.
Edge Computing and IoT	Research Focus	Focus on scalability, management, resource allocation, and edge-cloud orchestration.
	Network Advancements	5G and future standards to enhance connectivity and service delivery.
5G and Future Networks in IoT	Industry Transformation	Enabling real-time communication and smart industrial environments.
	Research Focus	Developing protocols for sensor-edge communication and improving QoE.
General Trends With AI	Standardization	Push for standardization and interoperability among cloud, edge, and fog computing.
	Hardware Innovation	Development of custom AI hardware and new AI frameworks.
Regulatory and Ethical Concerns	Ethical Implementation	Addressing biases, potential misuse, transparency, and accountability in AI systems.
	Data Privacy	Developing laws and standards for data privacy and ethical AI application.

9.2. Fog Computing

The convergence of fog computing with AI is expected to drive revolutionary breakthroughs in the upcoming future. It is likely to revolutionize several sectors by improving data management right at the edge of the network for more effective analysis. The efforts to couple AI and fog computing are feasible in the real-time data analysis spectrum, decision-making processes, and overall system efficiency. Primary sectors of focus are AI-based fog and edge computing, the augmentation of AI in fog and edge, and finally, cloud-fog automation. There are existing studies that provide an outline of the future of fog and AI. The study by [50] investigates and classifies AI-directed fog and edge computing in a structured manner. Augmenting fog with AI is approached in [51], where different aspects of fog and how AI can make them better are discussed in terms of performance, reliability, and efficiency. The study by [52] concerns itself with the main challenges and advancements in AI-augmented edge and fog computing that exist today. In turn, the survey implies a set of goals, supportive technologies, and possible directions for possible future investigations in the automation of cloud fog. Based on these studies, it can be inferred that the world of fog will promote the realization of more intelligent IoT applications, autonomous systems, and more adaptive and responsive networks. As a result, various innovations in fields such as industrial automation, smart cities, healthcare, and transport are anticipated. However, the principal complication remains the integration of AI algorithms with fog computing architectures, taking into account data security and privacy, processing delays, and, at the same time, supporting innovations in distributed computing. It can be inferred that research in these sectors will give the fog AI era a boost.

9.3. Edge Computing and IoT

Major concerns related to edge computing and IoT sensor networks are robust security, privacy protections in distributed environments, scaling, and efficient management of sprawling networks of devices in real-time to cope with the challenges of data management and storage at the edge. Network connectivity and reliability are also among the top con-

cerns. However, in most edge use cases, downtime must be minimal and latency sparsely imposed to meet the real-time processing requirements of edge applications. Other key considerations for edge are energy efficiency and cost; solutions for edge computation are supposed to be financially good and in line with the ecosystem. Furthermore, introducing AI at the edge makes it even more challenging to maintain computational efficiency and model accuracy. Future research will focus on improved scalability and management through autonomic systems, as hinted at in the study by [53], and improved resource allocation, along with refined edge–cloud orchestration, as mentioned in the study by [54]. Other advances that will follow are distributed data analytics, edge computing architectures, and innovative data processing models. Generally, what follows in this research area—the optimization of storage solutions—are questions related to caching strategies, data synchronization, and data consistency. Network resilience, low-latency networking solutions, network slicing, and virtualization are key enablers for robust connectivity, and 5G and future telecommunication standards will unlock new potential for remote connectivity and service delivery, as outlined in research from [55]. In brief, advancements are moving toward performance- and sustainability-enhanced development in the areas of analytics infrastructure, network infrastructure, and edge-to-cloud orchestration while solving major challenges such as security, scalability, data management, network reliability, energy saving, and AI integration.

9.4. 5G and Future Networks in IoT

The introduction of 5G and MEC technology is a pivotal change in the world of IoT. It is way more than just the small sensors in our homes or the big machines in our factories—it is everything. This is not just wishful thinking, but in fact, it is really fast due to the rapid speeds and ultra-low latency with 5G. The most exciting part about this adaptation is how 5G will enable the evolutionary growth of the industrial IoT. This is not merely another booster in the speed of connectivity but rather a technology that puts intelligence into industrial environments to make them smart, efficient, and flexible. In other words, real-time communication for 5G networks with each other and edge infrastructure has to have the capability to make automated decisions that make productivity better. Equally important is the synergy that exists between edge computing and 5G, as mentioned in the research work [56] that discusses processing data near its source, going to another extreme of reduction in dependence on central servers, a decrease in delay, and new possibilities of conducting real-time analytics and making decisions. New protocols dedicated to sensor-edge communication are a key area of research. As discussed in [57], the integration of 5G and IoT is creating a more connected world; from smart cities to autonomous vehicles, the applications are endless and exciting. Research into smooth implementation and operation in terms of quality of experience (QoE) for the users is of significant importance. Lastly, as we step into the future, papers like [58] remind us that the evolution of technology is constant and that developing sustainable and flexible solutions to challenges is important for future safe systems.

9.5. General Trends with AI

To smoothly integrate AI/ML applications, there will probably be a push for standardization and interoperability among these computing paradigms. Performance at the cloud, edge, and fog levels will be improved by the further development of custom hardware for AI, such as processors designed specifically for machine learning tasks. The efficiency and viability of AI/ML applications, particularly in edge and fog computing, will be greatly impacted by the introduction of 5G and subsequent telecommunications standards. In [59], an interesting concept called a “Cloud-Edge-Terminal Collaborative Network” (CETCN) is examined as a new paradigm suited for emerging applications. It emphasizes the benefits of deep reinforcement learning (DRL), which offers flexibility such as not requiring accurate modeling of the CETCN, effectively responding to high-dimensional and dynamic tasks, and enabling collaboration between different vendors. Innovation at the hardware level

will also continue to make the cloud suitable for collaborating with AI-related technologies. Hyperscale data centers are massive facilities that are specially designed to meet the extensive needs of large-scale data processing and storage. Innovation using specialist chips like GPUs and field programmable gate arrays (FPGAs) have opened doors to new frameworks to support AI [60]. Further studies are being done on architectures that enable collaboration between cloud and edge AI. The researchers in [61] address complex topics such as pretraining models, graph neural networks, and the principles of reinforcement learning and also identify possible future directions and challenges in the field.

9.6. Regulatory and Ethical Concerns

It is expected that massive ethical and regulatory problems at the border of cloud computing and AI, with regard to ethical implementations through regulated data privacy, strong guarantees of security, and clear accountability while considering what impact decisions that AI systems make have on society, will be discussed more in the near future. This borders on possible biases in AI algorithms, the potential misuse of AI, and the requirement for AI systems to be transparent and intelligible. The regulations should aim at developing standards that support laws on data privacy, encourage the moral application of AI, and protect from risks associated with the merger between AI and cloud computing. The authors of [62] are concerned about the unethical issues raised or, rather, hiccups in regulations for AI, and they focus on the need for moral standards and laws. It is clear that fresh ethical frameworks have to be designed to revisit computer ethics, more specifically in terms of AI systems and their legislative frameworks. The use cases of these computing paradigms will be tempered by stiffer data privacy laws across these computing paradigms, as emphasized in [63,64]. All those issues need to be addressed for the growth of AI in an ethical and sustainable manner on a cloud platform. In the future, more consideration will be given to developing transparent, fair, and ethical AI systems in research directions regarding AI and cloud paradigms.

10. Conclusions

In this paper, we have discussed how cloud computing, edge intelligence, and AI are converging to further transform IoT. The central insight from our study is that we have specified a framework for AI for Edge and AI on Edge by conducting a literature review. The computational power and scalability that complex AI and ML models require can be delivered through cloud computing, while hybrid cloud solutions enhance flexibility and deployment optimization. When linked with fog and edge computing, AI will supply real-time analytics for arriving data, better decision-making, and higher system efficiency. Edge AI reduces latency and enhances the efficiency and response time of IoT networks; this is rather crucial for those applications requiring real-time processing. Further, 5G and MEC technologies support further real-time communication with reduced reliance on central servers. Future research will focus on developing standardized protocols and interoperability frameworks in a rhythm that goes well with AI/ML applications across cloud, fog, and edge computing paradigms. It will be supplemented by next-generation AI hardware such as GPUs, TPUs, and NPUs for edge AI performance. Second, edge–cloud orchestration techniques like federated learning and multi-agent systems will be essential; the assurance of ethical AI deployment is indispensable. Individual levels of privacy and security will have to be catered to during the analysis of aggregate data through differential privacy techniques, while secure transmission protocols answer for the integrity and confidentiality of the data throughout all layers of computation. Addressing these areas will fully harness the potential of cloud, edge, and AI convergence to create a brighter, more efficient, and more secure digital future. These topics require extensive discussion that is beyond the scope of our current research work; however, the importance of studying these topics is of prime significance for the growth and development of AI and cloud computing paradigms.

Author Contributions: Conceptualization, N.F.P. and J.W.; methodology, N.F.P.; validation, N.F.P.; formal analysis, N.F.P.; investigation, N.F.P.; resources, N.F.P.; data curation, N.F.P.; writing—original draft preparation, N.F.P.; writing—review and editing, N.F.P. and J.W.; visualization, N.F.P.; supervision, J.W.; project administration, J.W.; funding acquisition, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by NSF grants SaTC 2310298, CNS 2214940, CPS 2128378, CNS 2107014, and CNS 2150152.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Giordanelli, R.; Mastroianni, C. The cloud computing paradigm: Characteristics, opportunities and research issues. *Ist. Calc. Reti Ad Alte Prestazioni (ICAR)* **2010**, *5*, 11.
2. Khan, T.; Tian, W.; Zhou, G.; Ilager, S.; Gong, M.; Buyya, R. Machine learning (ML)-centric resource management in cloud computing: A review and future directions. *J. Netw. Comput. Appl.* **2022**, *204*, 103405. [[CrossRef](#)]
3. Sriram, G.S. Edge computing vs. Cloud computing: An overview of big data challenges and opportunities for large enterprises. *Int. Res. J. Mod. Eng. Technol. Sci.* **2022**, *4*, 1331–1337.
4. Rosendo, D.; Costan, A.; Valduriez, P.; Antoniu, G. Distributed intelligence on the Edge-to-Cloud Continuum: A systematic literature review. *J. Parallel Distrib. Comput.* **2022**, *166*, 71–94. [[CrossRef](#)]
5. Barbuto, V.; Savaglio, C.; Chen, M.; Fortino, G. Disclosing edge intelligence: A systematic meta-survey. *Big Data Cogn. Comput.* **2023**, *7*, 44. [[CrossRef](#)]
6. Pujol, V.C.; Donta, P.K.; Morichetta, A.; Murturi, I.; Dustdar, S. Edge intelligence—Research opportunities for distributed computing continuum systems. *IEEE Internet Comput.* **2023**, *27*, 53–74. [[CrossRef](#)]
7. Armbrust, M.; Fox, A.; Griffith, R.; Joseph, A.D.; Katz, R.; Konwinski, A.; Lee, G.; Patterson, D.; Rabkin, A.; Stoica, I.; et al. A view of cloud computing. *Commun. ACM* **2010**, *53*, 50–58. [[CrossRef](#)]
8. García-Valls, M.; Cucinotta, T.; Lu, C. Challenges in real-time virtualization and predictable cloud computing. *J. Syst. Archit.* **2014**, *60*, 726–740. [[CrossRef](#)]
9. Buyya, R.; Srivastava, S.N. *Fog and Edge Computing: Principles and Paradigms*; John Wiley & Sons: Hoboken, NJ, USA, 2019.
10. Sowmya, S.K.; Deepika, P.; Naren, J. Layers of cloud–IaaS, PaaS and SaaS: A survey. *Int. J. Comput. Sci. Inf. Technol.* **2014**, *5*, 4477–4480.
11. Wittig, A.; Wittig, M. *Amazon Web Services in Action: An In-Depth Guide to AWS*; Simon and Schuster: New York, NY, USA, 2023.
12. Copeland, M.; Soh, J.; Puca, A.; Manning, M.; Gollob, D. Microsoft azure and cloud computing. In *Microsoft Azure: Planning, Deploying, and Managing Your Data Center in the Cloud*; Apress: Berkeley, CA, USA, 2015; pp. 3–26.
13. Bisong, E. *Building Machine Learning and Deep Learning Models on Google Cloud Platform*; Apress: Berkeley, CA, USA, 2019.
14. Mishra, S.; Tripathi, A.R. AI business model: An integrative business approach. *J. Innov. Entrep.* **2021**, *10*, 18. [[CrossRef](#)]
15. Ahamad, S.; Mohseni, M.; Shekher, V.; Smaisim, G.F.; Tripathi, A.; Alanya-Beltran, J. A detailed analysis of the critical role of artificial intelligence in enabling high-performance cloud computing systems. In Proceedings of the 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 28–29 April 2022.
16. Li, Z.; Wang, Y.; Wang, K.-S. Intelligent predictive maintenance for fault diagnosis and prognosis in machine centers: Industry 4.0 scenario. *Adv. Manuf.* **2017**, *5*, 377–387. [[CrossRef](#)]
17. Dasgupta, G.B. AI and its Applications in the Cloud strategy. In Proceedings of the 14th Innovations in Software Engineering Conference (Formerly Known as India Software Engineering Conference), Bhubaneswar, India, 25–27 February 2021.
18. Ikhlasse, H.; Benjamin, D.; Vincent, C.; Hicham, M. An overall statistical analysis of AI tools deployed in cloud computing and networking systems. In Proceedings of the 2020 5th International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech), Marrakesh, Morocco, 24–26 November 2020; pp. 1–7.
19. Sharma, Y.; Khan, M.G.; Al-Dulaimy, A.; Khoshkhohghi, M.A.; Taheri, J. Networking models and protocols for/on edge computing. *Edge Comput. Model. Technol. Appl.* **2020**, *33*, 77.
20. Gray, J.; Helland, P.; O’Neil, P.; Shasha, D. The dangers of replication and a solution. In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, QC, Canada, 4–6 June 1996; pp. 173–182.
21. Deng, S.; Zhao, H.; Fang, W.; Yin, J.; Dustdar, S.; Zomaya, A.Y. Edge intelligence: The confluence of edge computing and artificial intelligence. *IEEE Internet Things J.* **2020**, *7*, 7457–7469. [[CrossRef](#)]
22. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
23. Surianarayanan, C.; Lawrence, J.J.; Chelliah, P.R.; Prakash, E.; Hewage, C. A survey on optimization techniques for edge artificial intelligence (AI). *Sensors* **2023**, *23*, 1279. [[CrossRef](#)] [[PubMed](#)]

24. Xue, M.; Wu, H.; Peng, G.; Wolter, K. DDPQN: An efficient DNN offloading strategy in local-edge-cloud collaborative environments. *IEEE Trans. Serv. Comput.* **2021**, *15*, 640–655. [[CrossRef](#)]
25. Cao, B.; Wu, T.; Bai, X. Stochastic programming based multi-arm bandit offloading strategy for internet of things. *Digit. Commun. Netw.* **2023**, *9*, 1200–1211. [[CrossRef](#)]
26. Lu, H.; Gu, C.; Luo, F.; Ding, W.; Liu, X. Optimization of lightweight task offloading strategy for mobile edge computing based on deep reinforcement learning. *Future Gener. Comput. Syst.* **2020**, *102*, 847–861. [[CrossRef](#)]
27. Khan, L.U.; Pandey, S.R.; Tran, N.H.; Saad, W.; Han, Z.; Nguyen, M.N.H.; Hong, C.S. Federated learning for edge networks: Resource optimization and incentive mechanism. *IEEE Commun. Mag.* **2020**, *58*, 88–93. [[CrossRef](#)]
28. Arumugam, S.; Shandilya, S.K.; Bacanin, N. Federated learning-based privacy preservation with blockchain assistance in IoT 5G heterogeneous networks. *J. Web Eng.* **2022**, *21*, 1323–1346. [[CrossRef](#)]
29. Li, W.; Liewig, M. A survey of AI accelerators for edge environment. In *Trends and Innovations in Information Systems and Technologies: Volume 2*; Springer: Cham, Switzerland, 2020; pp. 35–44.
30. Frank, R.; Schumacher, G.; Tamm, A. The cloud transformation. In *Cloud Transformation: The Public Cloud Is Changing Businesses*; Springer: Wiesbaden, Germany, 2023; pp. 203–245.
31. Peterson, L.; Anderson, T.; Katti, S.; McKeown, N.; Parulkar, G.; Rexford, J.; Satyanarayanan, M.; Sunay, O.; Vahdat, A. Democratizing the network edge. *ACM SIGCOMM Comput. Commun. Rev.* **2019**, *49*, 31–36. [[CrossRef](#)]
32. Arif, T.M. *Introduction to Deep Learning for Engineers: Using Python and Google Cloud Platform*; Springer Nature: Berlin/Heidelberg, Germany, 2022.
33. Elger, P.; Shanaghy, E. *AI as a Service: Serverless Machine Learning with AWS*; Manning Publications: Shelter Island, NY, USA, 2020.
34. Ravulavaru, A. *Google Cloud AI Services Quick Start Guide: Build Intelligent Applications with Google Cloud AI Services*; Packt Publishing Ltd.: Birmingham, UK, 2018.
35. Salvaris, M.; Dean, D.; Tok, W.H. Microsoft AI platform. In *Deep Learning with Azure: Building and Deploying Artificial Intelligence Solutions on the Microsoft AI Platform*; Apress: Berkeley, CA, USA, 2018; pp. 79–98.
36. Dixit, A.K. How IBM is changing the world with cloud computing. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2021**, *7*, 355–360.
37. Islam, R.; Patamsetti, V.; Gadhi, A.; Gondu, R.M.; Bandaru, C.M.; Kesani, S.C.; Abiona, O. The future of cloud computing: Benefits and challenges. *Int. J. Commun. Netw. Syst. Sci.* **2023**, *16*, 53–65. [[CrossRef](#)]
38. Gallagher, S. *VMware Private Cloud Computing with vCloud Director*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
39. Zhang, G.; Ravishankar, M.N. Exploring vendor capabilities in the cloud environment: A case study of Alibaba Cloud Computing. *Inf. Manag.* **2019**, *56*, 343–355. [[CrossRef](#)]
40. Chandel, S.; Ni, T.Y.; Yang, G. Enterprise cloud: Its growth and security challenges in China. In Proceedings of the 2018 5th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2018 4th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), Shanghai, China, 22–24 June 2018; pp. 144–152.
41. Alkhawajah, W. Huawei: An information and communications technology company. *J. Inf. Technol. Econ. Dev.* **2019**, *10*, 1–10.
42. Jiang, M. The business and politics of search engines: A comparative study of Baidu and Google’s search results of internet events in China. *New Media Soc.* **2014**, *16*, 212–233. [[CrossRef](#)]
43. Shah, J.; Dubaria, D. Building modern clouds: Using Docker, Kubernetes, Google Cloud Platform. In Proceedings of the IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 7–9 January 2019.
44. Baset, S.A. Open source cloud technologies. In Proceedings of the Third ACM Symposium on Cloud Computing, San Jose, CA, USA, 14–17 October 2012.
45. Janes, A.; Li, X.; Lenarduzzi, V. Open tracing tools: Overview and critical comparison. *arXiv* **2022**, arXiv:2207.06875.
46. Turnbull, J. *Monitoring with Prometheus*; Turnbull Press: Haddington, Scotland, UK, 2018.
47. Nurmi, D.; Wolski, R.; Grzegorczyk, C.; Obertelli, G.; Soman, S.; Youseff, L.; Zagorodnov, D. The eucalyptus open-source cloud-computing system. In Proceedings of the 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, Shanghai, China, 18–21 May 2009; pp. 124–131.
48. Frampton, M.; Frampton, M. Apache mesos. In *Complete Guide to Open Source Big Data Stack*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 97–137.
49. Reifert, R.J.; Dahrouj, H.; Sezgin, A. Extended reality via cooperative NOMA in hybrid cloud/mobile-edge computing networks. *IEEE Internet Things J.* **2023**, *11*, 12834–12852. [[CrossRef](#)]
50. Iftikhar, S.; Gill, S.S.; Song, C.; Xu, M.; Aslanpour, M.S.; Toosi, A.N.; Du, J.; Wu, H.; Ghosh, S.; Chowdhury, D. AI-based fog and edge computing: A systematic review, taxonomy and future directions. *Internet Things* **2023**, *21*, 100674. [[CrossRef](#)]
51. Tuli, S.; Mirhakimi, F.; Pallewatta, S.; Zawad, S.; Casale, G.; Javadi, B.; Yan, F.; Buyya, R.; Jennings, N.R. AI augmented edge and fog computing: Trends and challenges. *J. Netw. Comput. Appl.* **2023**, *216*, 103648. [[CrossRef](#)]
52. Jin, J.; Yu, K.; Kua, J.; Zhang, N.; Pang, Z.; Han, Q.L. Cloud-fog automation: Vision, enabling technologies, and future research directions. *IEEE Trans. Ind. Inform.* **2023**, *20*, 1039–1054. [[CrossRef](#)]
53. da Silva, T.P.; Neto, A.R.; Batista, T.V.; Delicato, F.C.; Pires, P.F.; Lopes, F. Online machine learning for auto-scaling in the edge computing. *Pervasive Mob. Comput.* **2022**, *87*, 101722. [[CrossRef](#)]
54. Liu, B.; Chen, C.H. An adaptive multi-hop branch ensemble-based graph adaptation framework with edge-cloud orchestration for condition monitoring. *IEEE Trans. Ind. Inform.* **2023**, *19*, 10102–10113. [[CrossRef](#)]
55. Singh, R.; Gill, S.S. Edge AI: A survey. *Internet Things Cyber-Phys. Syst.* **2023**, *3*, 71–92. [[CrossRef](#)]

56. Mahmood, A.; Beltramelli, L.; Abedin, S.F.; Zeb, S.; Mowla, N.I.; Hassan, S.A.; Sisinni, E.; Gidlund, M. Industrial IoT in 5G-and-beyond networks: Vision, architecture, and design trends. *IEEE Trans. Ind. Inform.* **2021**, *18*, 4122–4137. [[CrossRef](#)]
57. Mani, V.; Kavitha, C.; Srividhya, S. Edge computing enabled by 5G for computing offloading in the industrial internet of things. In *Information Security Practices for the Internet of Things, 5G, and Next-Generation Wireless Networks*; IGI Global: Hershey, PA, USA 2022; pp. 210–228.
58. Patel, S. Advancements in Edge Computing: Fundamentals, Survey, Trends, and Future. *Authorea Prepr.* **2023**. [[CrossRef](#)]
59. Gu, H.; Zhao, L.; Han, Z.; Zheng, G.; Song, S. AI-enhanced cloud-edge-terminal collaborative network: Survey, applications, and future directions. *IEEE Commun. Surv. Tutor.* **2023**, *26*, 1322–1385. [[CrossRef](#)]
60. Pasumarty, R.; Praveen, R.; Mahesh, T. The future of AI-enabled servers in the cloud: A survey. In Proceedings of the 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 11–13 November 2021; pp. 578–583.
61. Yao, J.; Zhang, S.; Yao, Y.; Wang, F.; Ma, J.; Zhang, J.; Chu, Y.; Ji, L.; Jia, K.; Shen, T. Edge-cloud polarization and collaboration: A comprehensive survey for AI. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 6866–6886. [[CrossRef](#)]
62. Jacobs, M.; Simon, J. Reexamining computer ethics in light of AI systems and AI regulation. *AI Ethics* **2023**, *3*, 1203–1213. [[CrossRef](#)]
63. Nasim, S.F.; Ali, M.R.; Kulsoom, U. Artificial intelligence incidents and ethics: A narrative review. *Int. J. Technol. Innov. Manag. (IJTIM)* **2022**, *2*, 52–64. [[CrossRef](#)]
64. Hacker, P.; Engel, A.; Mauer, M. Regulating ChatGPT and other large generative AI models. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago, IL, USA, 12–15 June 2023; pp. 1112–1123.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.