International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Edge Computing vs. Cloud Computing: A Comparative Analysis for Real-Time AI Applications

Bangar Raju Cherukuri

Senior Web Developer, Department of Information Technology, Andhra University, INDIA

Abstract

Real-time AI has been evidenced to be supported by two models, namely edge computing and cloud computing. This research seeks to compare the two methods: Latency, security features, and data processing capabilities will be among the most critical aspects of comparison. Real-time AI applications are envisioned to be robust through 2020, most of which will be used in applications such as autonomous automobile systems and IoT devices that demand low latency for data processing. As data processing is outsourced in clouds, it is elastic and centralized, but this deal faces latency problems when the distance between the source of data and the processing center is large. On the other hand, edge computing refers to conducting computation closer to the data source, which could reduce latency and enhance everyday real-time performance.

This research assesses the above models based on the literature review, technical papers, and case studies from industries that depend more on real-time AI. According to the research, edge computing is typically more effective for latency-sensitive workloads, while cloud computing outperforms throughput-intensive applications. Security concerns, however, present themselves as having dual effects; while advancing the privacy of handling data through edge computing, it elevates new risks. As such, this study concludes that no clear winner depends on the necessary application, therefore recommending a symmetric mode for the requirements, where both latency time and computational needs are required. Further studies should be conducted to establish the coordinated implementation approach of these models.

Keywords: edge computing, cloud computing, AI, Data Processing, ML

1. Introduction

1.1 Background to the Study

Machine learning has found its way into engineering artificial intelligence relative to complex processes such as learning, reasoning, and solving problems that require natural intelligence (Nilsson, 2014). The ability to meet social demands depends on processing large volumes of data, especially in real-time applications (Buyya & Dastjerdi, 2016). Mainstream centralized computing structures cannot address these demands due to latency problems and narrowly available bandwidth (Garcia Lopez et al., 2015).

Cloud computing began as a shift in the new self-service computing model that provides on-demand, on-demand services through the Internet (Mell & Grance, 2011). It permits AI applications to run within large compute-scale and storage capacities but requires minimal physical infrastructure (Armbrust et al.,

2010). This model is ideal for data-intensive activities such as training machine learning models and big data analysis. But as we know, most cloud computing is centralized, so there is a hood of latency due to the geographical location of data centers and end-user devices, which are very important for real-time applications such as AI self-driven cars and sensors (Garcia Lopez et al., 2015).

To overcome these restrictions, edge computing has been proposed as a distributed computing paradigm closer to the data source (Shi & Dustdar, 2016). Reducing the distance of computation and storage from the data acquisition and processing source, edge computing cuts latency and bandwidth consumption for AI applications (Satyanarayanan, 2017). This is especially important for those applications where the data should be processed as soon as possible, and decisions should be taken, for example, in a healthcare monitoring system or industry automation systems (Buyya & Dastjerdi, 2016).

The problem of understanding how cloud and edge computing are beneficial or not for implementing AI applications that require efficient data processing remains (Garcia Lopez et al., 2015). Given that AI is slowly finding its way into the various spheres of technology and human society, choosing the right computing model is particularly important in striving for the best results and fulfilling context-sensitive requirements of multiple applications (Buyya & Dastjerdi, 2016).

1.2 Overview

The definition of cloud computing is a model for the consumption of IT resources as services available over the internet where users can access applications, servers, storage, and other computing resources as a shared pool of configurable resources. They have scalability and flexibility since users can retrieve and store their data over the internet without substantial local resources (Chiang & Zhang, 2016). Hence, using AI, cloud computing offers the resources to process big data and computational algorithms (Li et al., 2018).

Edge computing, for example, is a decentralized model that focuses on processing data closer to the physical location of that data to enhance the raw response time and be careful with bandwidth utilization (Satyanarayanan 2017). Edge computing minimizes latency inherent in data transfer to a centralized hub, which is essential in real-time AI operations (Zhang, Chen, & Li, 2018). It is most suitable when IoT devices or autonomous systems demand decision-making based on data they collect in the first instance.

Another remarkable competency of AI is processing real-time data in self-driving cars, energy managing smart grids, and real-time security surveillance systems where delay is dangerous. Delayed responses are unallowable in many applications and situations, which results in high latency in data processing performance (Li et al., 2018). As a result, the decision between cloud and edge deployment models greatly influences how well AI is implemented in real-time scenarios (Zhang et al., 2018).

Introducing and defining the concepts of cloud and edge computing is crucial for decision-making concerning high-need AI applications and their implementation needs, coverage, data processing delays, and bandwidth (Chiang & Zhang, 2016). While cloud computing has a great capacity for data storage and complicated computation, edge computing has the capability of low latency required for real-time systems (Satyanarayanan, 2017).

1.3 Problem Statement

Arising from this, real-time AI applications like autonomous vehicles and the industrial Internet of Things must provide results and respond to data within similarly short intervals. However, a unique set of challenges that face cloud computing is latency due to the location of the source data and cloud

server. This delay is important in different systems that require the operation to happen in real-time, most especially because every millisecond is important. Also, moving huge data traffic to offsite cloud servers is dangerous as data is exposed during transfer, for instance, while dealing with sensitive data. Another disadvantage of the centralized cloud computing model is that when handling large data streams, they get bottlenecked at the central server, which could be more efficient.

Due to the above limitations, there is a need for other possible solutions that can fit the above challenges. One potential solution is edge computing, which processes data near its sources, helping to decrease latency and offer localized data processing, which is less risky from information security and needless bandwidth. However, since edge computing purports to do this in some ways, it may be important to determine whether it can outcompete cloud computing and its execrations in various real-time AI scenarios.

1.4 Objectives

1. To compare the latency performance of edge computing with that of cloud computing.
2. To examine the security trends and issues related to both computing paradigms.
3. To examine the performance of edge and cloud computing for real-time AI data processing applications in detail.
4. To contrast actual-usage scenarios involving self-driving cars and Industries IoT systems to determine which model is more effective.
5. To identify which classes of applications edge computing is more suited to provide a framework by which cloud and edge computing could be compared to help developers and businesses make informed decisions as to which model is more appropriate for use in their applications.

1.5 Scope and Significance

This research concerns the cloud and edge computing paradigms in real-time AI environments where low latency and secure data processing issues surface. The study will mainly focus on automobiles, smart factories, and IoT, as the information gathered must be processed and acted upon immediately. This study will evaluate the four proposed computing models based on these technicalities and then assess the strengths and weaknesses of each model in terms of latency, security, and data processing.

This comparative analysis is important because it will advise businesses or developers to choose the best method for implementing AI applications. As more AI enters modern technology and becomes a part of people's daily lives, knowing the best way to manage real-time data will result in more dependable systems that work faster. For illustration, in autonomous vehicles, decisions must be made quickly for safety purposes, while in industrial IoT systems, the response must be very accurate, prompt, and efficient. This research will also help fill the current literature gap on cloud and edge computing to improve the effectiveness and security of AI-based solutions and modernize real-time AI solutions across industries.

2. Literature Review

2.1 Definition and Evolution of Cloud Computing

Cloud computing is described as a style for delivering services through the internet, which can be accessed from any place, at any time, with minimum management intervention (Mell & Grance, 2011). This new computing model allows users to access and consume or use IT services via the internet and

only pay for what they consume, reducing the infrastructure costs required to support these infrastructures (Armbrust et al., 2010).

The evolution of cloud computing dates back to the utility computing concept in the mid-1960s, in which computation was expected to be like electricity (Parkhill, 1966). However, the practical realization began in the early 2000s with Virtualization and distributed computing ideas to manage resources effectively (Buyya et al., 2009). Public cloud service was announced into the commercial market in 2006 after the launching of Amazon Web Services Elastic Compute Cloud (EC2) (Armbrust et al., 2010).

Cloud computing evolved through various service models: IaaS-Infrastructure as a Service, which refers to virtualized computing resources offered over the Internet; PaaS-Platform as a Service, which supplies hardware and software tools over the Internet; SaaS-Software as a Service provides software applications over the internet in a subscription model (Zhang et al., 2010). These models offered on-demand resource provisioning, which helped them to grow, shrink, and adjust as needed based on the response to demands (Vaquero et al., 2008).

The trends currently experienced in cloud computing include Increased security, compatibility with other advancements, and optimization; due to the convergence of cloud computing with IoT devices, the platforms known as Cloud-IoT have been developed to handle large-scale data processing and analytics (Botta et al., 2016). Moreover, using containers also incorporates application deployment and scalability within cloud facilities, while microservices architectures enhance the cloud deployment of complicated applications (Pahl, 2015). Another architecture that has risen to address latency and bandwidth problems is edge computing, which focuses on processing data nearer the point where it is gathered (Shi & Dustdar, 2016).

Nevertheless, cloud computing also has certain limitations in terms of security and privacy, and facing several legal challenges, it needs continuous improvement in the research area (Zissis & Lekkas, 2012). Cloud computing remains a progressing concept that has changed to address the requirements of other real-time applications and interact with other emerging technologies.

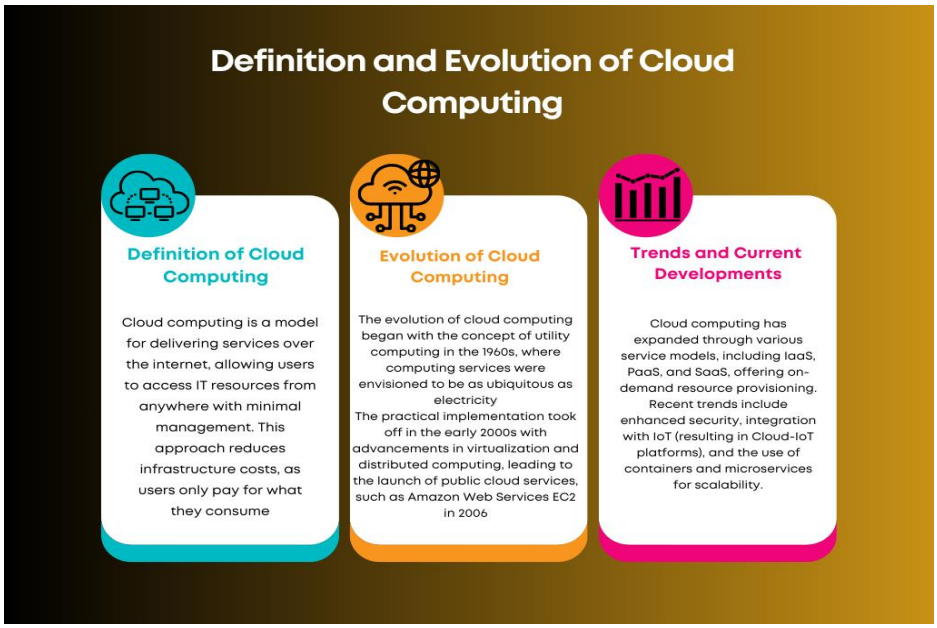


Fig 1: An image illustrating Definition and Evolution of Cloud Computing

2.2 Definition and Evolution of Edge Computing

Edge computing is a decentralized approach to computing that delivers computation and data storage services near the point where it is required, enhancing response times and reducing utilization of Bandwidth (Shi et al., 2016). Orchestrated edge computing that intervenes in the role of cloud computing occurred due to its inefficiency in handling issues such as latency, bandwidth consumption, and real-time information processing that the current applications like autonomous vehicles and IoT devices require (Satyanarayanan, 2017).

Edge computing was first conceived from the content delivery networks (CDNs) in the late 1990s, when computing infrastructure was placed on the boundaries of the networks to improve the delivery of web and videos (Shi & Dustdar, 2016). However, the increased use of IoT devices and the fast-increasing data generated at the network edge called for a more enhanced solution for processing and analyzing data at the network edge. This led to the formalization of edge computing around 2014, with the promise of computing on or near the data source (Satyanarayanan, 2017).

The disadvantages associated with the latency of cloud computing are solved by edge computing since it brings the data closer to where they are processed and made decisions et al., 2016). For instance, in the case of autonomous cars, it is crucial to process the data received from the car's sensors in real time to ensure the safety and effectiveness of the operation (Chiang & Zhang, 2016). Edge computing also increases data privacy and security, as data does not have to be transmitted over the network to central servers when processed locally (Shi & Dustdar, 2016).

Today's advancement includes adopting technologies such as 5G networks to offer higher bandwidth and lower latency to uplift edge computing capabilities (Taleb et al., 2017). However, edge AI allows devices to perform large computation tasks, including image recognition and natural language processing, without Cloud resources (Li et al., 2018). There are some issues, such as how to deal with the heterogeneity of edge devices, provide security measures for edge computing, and provide a reference model/framework for edge computing deployment (Shi et al., 2016).

Edge computing must stand still as it is now constantly under development to fit the new requirements for nearly real-time processing and a continuously growing number of connected gadgets. It is an extension of cloud computing since it transfers the computations, relieves the network traffic, and offers real-time analysis and response to data (Satyanarayanan, 2017).

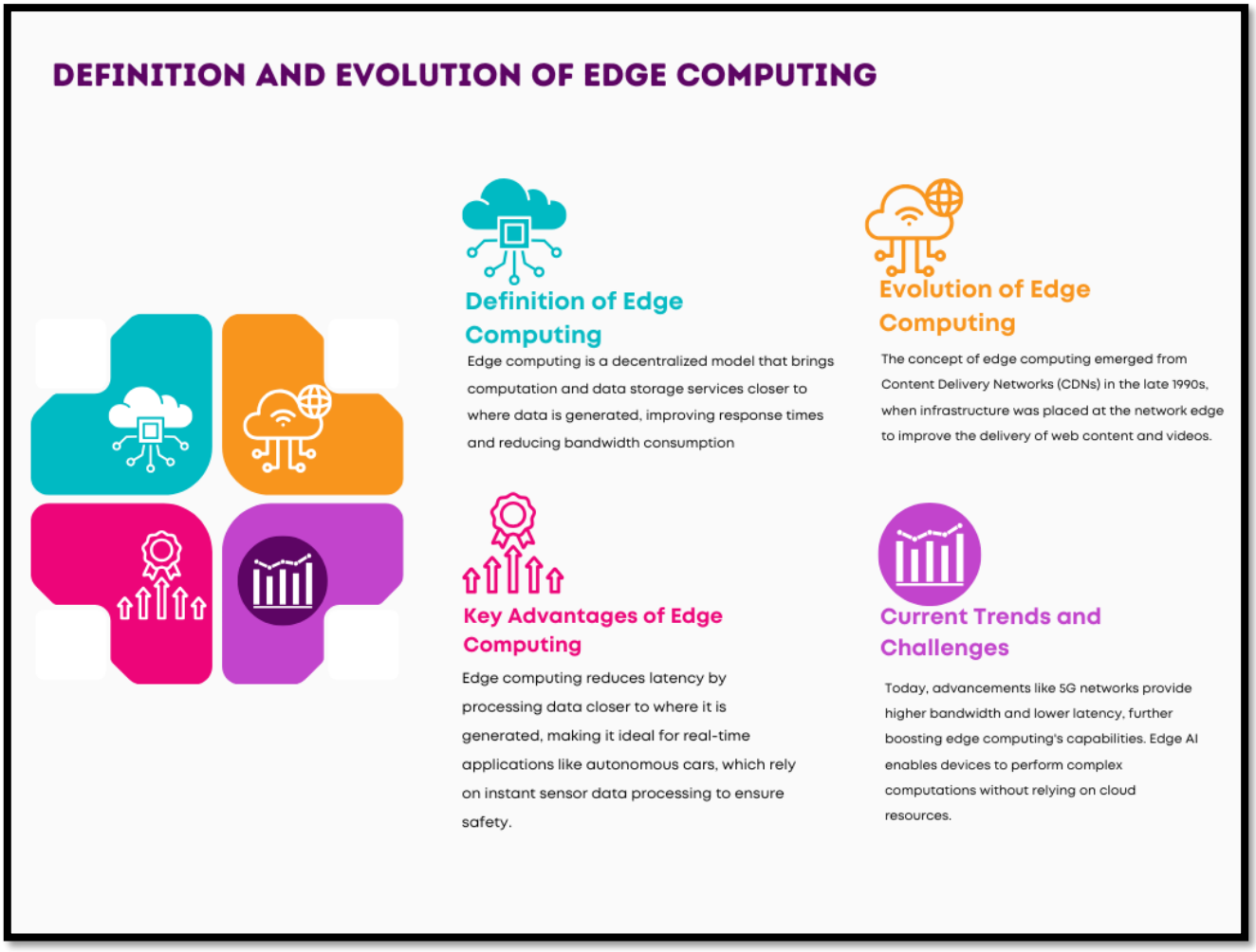


Fig 2: An image illustrating Definition and Evolution of Edge Computing

2.3 Latency Issues in Cloud vs. Edge Computing

For real-time applications, there can be no talk of AI without mentioning latency issues since data processing and response delays become a factor in efficiency and the system's usability (Shi, Zhang, & Li, 2018). Regarding cloud computing, data generated at end devices must be sent over the network to distant or centralized server warehouses for processing, which could cause a significant increase in latency due to competition for the shares that may stem from distance (Mao et al., 2017). For example, in self-driving cars, even in infrequent cases, their data processing output should be obtained immediately to support real-time decisions to avoid creating dangerous situations that could arise when an autonomous car receives a signal that it is safe to proceed through an intersection just a couple of milliseconds after a human driver comes to that conclusion (Shi et al., 2018).

These latency problems are solved by edge computing since it enhances data processing near the source, and thus, we reduce the distance data must travel to achieve this goal (Shi et al., 2018). By pre-processing some data at edge nodes, the completion of data analysis and subsequent actions can be accomplished almost in real time; this is especially important for AI applications that have stringent response time requirements, such as surveillance and industrial control (Wang et al., 2017). For instance, in smart manufacturing, edge computing helps identify faulty machine behavior in operation, leading to prompt rectification and little downtime (Shi et al., 2018).

According to the research, while implementing edge computing, the latencies can be brought down from the hundreds of milliseconds to mere milliseconds, which is important in today's applications such as virtual reality, online gaming, and more, where real-time response is important (Mao et al., 2017). Nevertheless, as a result of utilizing distant data centers, cloud computing is less efficient in providing the low latency that is necessary for certain AI applications (Wang et al., 2017). Further, it helps to release a load off the network through local computation, hence pulling less data to the cloud (Shi et al., 2018).

Moreover, edge computing also supports distributed artificial intelligence models with machine learning and inference at the edge, resulting in less time on interference and fater data capacity and dependability (Mao et al., 2017). Isolating data processing closer to the data source reduces latency and network traffic and is particularly important as the model complexity and data volume growth (Shi et al., 2018). Consequently, edge computing is a suitable solution to the latency issues in cloud computing, especially with applications of deep learning that require almost real-time predictions and processing.

2.4 Data Security in Cloud vs. Edge Computing

One of the biggest concerns in cloud and edge computing models is data security issues, which are different in both models, and other types of threats and challenges are involved (Roman, Lopez, & Mambo, 2018). In adopting cloud computing, the information is stored and processed in data centers, thus meaning that cybercriminals have a massive repository of mutual information to attack (Abhishek, Rani, & Singh, 2018). Some risks associated with using the cloud include leakage and dishonest access to the data stored in the cloud since the attackers easily exploit the hosts by hacking the virtualization technologies or intercepting the data in transit (Roman et al., 2018).

There are newer security concerns due to the decentralized nature of edge computing and because many devices are installed at the edges of the network (Roman et al., 2018). These edge devices are characterized by their low processing capabilities, and therefore, it becomes hard to incorporate efficient security measures like intricate encryption and efficient anti-incursion systems (Zhao et al., 2019). Moreover, the attack surface is very large due to the complexity and many devices at the edge. This means a bigger possibility for an edge device to get infected by malware or be targeted by a DDoS attack (Roman et al., 2018).

A key challenge in edge computing specifically relates to the possibility of physical attacks because the devices are commonly placed in insecure or even physically inaccessible areas (Deng et al., 2019). Such exposure can cause a breach of the system and unauthorized alteration of data, which is undesirable. Additionally, with different manufacturers and platforms currently coming up and working under different strategies, there might be an inconsistent approach to security, which may complicate the general handling of security for the overall edge network (Roman et al., 2018).

Nevertheless, edge computing can help to improve data privacy as many computations can be done at the edge and do not require sending data through potentially hazardous networks to centralized servers (Roman et al., 2018). Such processing reduces the exposure of data to transmission, which is a vulnerable point in cloud computing models (Zhao et al., 2019). Since edge computing devices come with limited resources, lightweight security solutions that suit such devices are highly recommended to secure the edging computing environments without compromising the devices (Deng et al., 2019).

2.5 Data Processing Capabilities: Cloud vs. Edge

Throughput is important when comparing cloud and edge computing and how they execute real-time AI applications. Cloud computing has a high computation capability, flexible storage, and accommodating a large amount of data and computation (Shi et al., 2016). Large data centers can bootstrap tremendous resources required to undertake complex data processing, such as training various machine learning algorithms or other large-scale analytical jobs. However, this centralized approach needs to improve in that there is always a time delay because data has to go from the source to the cloud and then back (Yu et al., 2018).

On the other hand, Edge computing structured data as close as possible to the source, which makes it easier for it to be processed and gives a faster response to data that otherwise would travel long distances before being processed. (Shi et al., 2016) This localized processing capability eliminates the crucial latency in applications such as self-driving cars and smart factories. For example, by describing data at the network's edge, the systems can provide analytics or make decisions more quickly and improve the system's efficiency (Satyanarayanan, 2017).

Even though a cloud-computing environment is designed for handling enormous quantities of data, edge-computing outperforms it when fast response is important. This increasing demand for real-time AI applications implies that edge computing can work in synergy with cloud computing, where the latter can handle tasks that require processing other than in real-time (Shi et al., 2016).

2.6 Case Studies: Autonomous Vehicles

Self-driving cars require integrating, analyzing, and analyzing massive volumes of information within a limited time to execute their functions safely and optimally. Most self-driven cars employ cloud computing to store and process data collected by the vehicle since it has been the preferred computation paradigm in the automotive industry (Hou et al., 2016); however, it has some limitations, such as high latency due to centralized data centers. For example, an executive self-driving car has to determine what to do when it finds a barrier; it has to analyze data from several sensors and cams simultaneously. Even if it relies on cloud computing for this process, it can cause extra delay in sending back and forth data to a separate server before anything is done.

To overcome these latency issues, Edge computing is gradually being incorporated into the technologies that underpin self-driving cars, allowing data to be processed near the source of the data(Kumar et al., 2018). When edge nodes are put in the vehicle or other connected structures, the autonomous systems can gather data and respond faster to changes occurring on the roads. Such local handling capacity is critical for functions like object recognition, lane tracking, or immediate path guidance, whose execution can be impaired by delays and is, therefore, fatal (Hou et al., 2016).

An example of the practical implementation of the discussed approach is vehicular fog computing VFC, where vehicles are considered mobile data centers that perform computations and then exchange processed information with nearby vehicles (Zhou et al., 2018). Besides, it helps decrease latency and increase data privacy because when it is necessary to transmit imperative data, it doesn’t have to be transmitted over a long distance. Besides, it relieves network congestion by lowering the data transmitted to cloud servers, making the entire system work more effectively (Hou et al., 2016).

As technology advances, it may become even more common to use edge and cloud societies where edge nodes will perform the first tier of analytics. In contrast, cloud systems will look for a non-time-sensitive second tier. This way, the hybrid approach allows us to achieve real-time response while accessing big

computational resources, which can be effectively implemented in numerous automotive applications (Kumar et al., 2018).

2.7 Case Studies: Internet of Things (IoT)

The Internet of Things (IoT) is a system that includes numerous interconnected devices that contribute data at an increasing rate, so processing models must be well coordinated for effective functioning. In particular, general IoT data has been processed and stored through cloud computing, which offers platforms where information from different sensors and devices can be analyzed (Xu, He, & Li, 2014). However, centralized cloud infrastructure causes latency, which is unsuitable for real-time IoT applications like smart health care and other industries.

These problems have been solved by edge computing, a process of performing analysis where the digital devices are located to minimize the required time to analyze and perform useful actions on the collected data (Shi & Dustdar, 2016). For instance, in the industrial IoT setting, edge computing helps to process data in real-time at the factory level, consequently identifying faulty equipment and fixing them quickly, which would mean minimal time offline (Xu et al., 2014). The processing at the local level also reduces the bandwidth requirement since only the processed data are transmitted to the cloud for additional analysis or storage.

Furthermore, edge computing provides short response times between individual devices and applications in consumer IoT applications such as smart homes (Garcia Lopez et al., 2015). This is because edge computing guarantees that data processing is done at the collection site, depending on the network requirements, and minimizes delays and privacy concerns since some data may not need to cross a local network. Cloud and edge computing make the IoT ecosystems effective while guaranteeing the capability and security of the data (Xu et al., 2014).

3.0 METHODOLOGY

3.1 Research Design

The research methodology used in this study involves evaluating cloud and edge computing with regards to pre-specified performance indicators. The ones seen as limitations are latency, data processing, data security, as well as scalability. This will make it easy for us to evaluate the dynamic behavior of every computing model in real-time real-time AI applications such as self-driving cars and the internet of things. The comparative framework will entail comparing the results obtained from the various use case scenarios and comparing and contrasting the quality of the inferential and subordinate results derived from both qualitative and quantitative results. By making such comparisons systematically across these metrics, the study hopes to reveal the circumstances that may make one model more advantageous and, hence, understand when pursuing one particular type of computing for a specific type of AI application is beneficial.

3.2 Data Collection

Sources of information for this study will be from theoretical and empirical perspectives, and key sources of information will include studies, cases, and technical reports. The literature review will give an initial appreciation of cloud computing, how it has evolved, and where it stands, and the same is true for edge computing. Each computing model will be described using the scenario-based approach, and case studies will demonstrate how they work in practice. Secondary data will be collected by reviewing

technical reports from infamous industry players and organizations to gather current information on innovations, constraints, and production best practices. As a result, the study will use theoretical and empirical data to synthesize the two paradigms to provide the best comparison.

3.3 Case Studies/Examples

3.3.1 Tesla's Autonomous Driving Systems

Edge computing is important in Tesla automobiles because it improves well-known self-driving recognition systems. Tesla cars are fitted with autonomous processors such as the Full Self-Driving (FSD) computer that computes data collected from sensors, including cameras, radar, and ultrasonic devices within the car (Tesla, n.d.). This local processing is particularly important for functions like obstacle detection and the subsequent decision-making that needs to occur in real-time to avoid crashes, which cannot be done if the data has to be uploaded to cloud servers (Lee, 2019).

It has low latency, which is important in leadership decision-making, especially for autonomous vehicles in ever-changing road conditions (Hu et al., 2015). Such data processing at the Edge allows Tesla vehicles to perform real-time object recognition and navigation tasks. Data privacy and security are also improved because critical data is processed within the vehicle without exposing it to network risks (Rao & Selvamani, 2015).

Similarly, Amazon additionally employs IaaS to exploit the cloud for tasks that involve mass data processing, including fleet learning and application updates, in Tesla (Cunningham, 2019). Information collected from the individual vehicle is transmitted to the cloud for central processing to improve the AI algorithm in use and the general performance of the whole system. This model integrates the latency properties in edge computing with the scale of cloud computing to enhance the functionality and security of autonomous driving systems (Hu et al., 2015).

3.3.2 Smart Home IoT Devices

It is used in Smart Homes where various things like thermostats, lighting, security cameras, etc., work more efficiently and with better reliability as they use edge computing. These devices perform data analysis locally, and an immediate response is given to a user’s actions and changes in the environment compared to cloud processing. For instance, an intelligent thermostat can raise or lower temperature depending on the sensors, thus enhancing efficiency and ensuring human comfort without consulting a server (Gubbi et al., 2013).

In smart homes, edge computing also increases privacy and security since the data is kept and processed within the home's local area network rather than sending information online. Reidentification can be done at the on-board camera unit if the system is designed to execute face and motion detection features (Siciar et al., 2015).

Nevertheless, there is always a role for cloud computation where high computation or raw data storage levels are needed, such as long-term data analysis and integrating new machine learning models into the edge environment (Marinissen & Taubenblatt, 2018). Using the edge and cloud compartments gives pro-home smart systems the benefits of real-time operations response and enhanced features.

3.4 Evaluation Metrics

The evaluation of cloud computing and edge computing models will be based on four key performance metrics: Delay time, data accuracy, rate of operation, and growth capability. In its turn, the term latency is used to address the time between the formation of data and the provision of result and is critical in

real-time AIS. In general, edge computing should decrease the latency associated with the data collected since much of this information is analyzed directly on devices. In contrast, cloud computing can have high latency due to the need to send the data to central servers.

The fourth area of concern is data security, which is a crucial metric because both models have their difficulties. In effect, cloud computing must mitigate the dangers of centrally storing vast volumes of data that can be breached and intercepting sensitive data in transit. In contrast, edge computing requires a means to secure numerous distributed devices, which presents an even larger attack surface. Security issues are vital in executing both models to ensure that sensitive information is properly protected.

Throughput pertains to the capacity of receiving and responding to the data in the least amount of time possible. At the same time, cloud computing works best when big data are involved and likely to overload local devices and networks. Edge computing is real-limited to local networks where real-time data requires a fast response.

Flexibility evaluates the ability of the system to expand as the company grows. Clouds are fundamentally elastic environments that can increase resources when necessary, while Edge often needs help finding more auxiliary devices to employ. These will be the basics through which the efficiency of one style of computing will be matched with the inefficiency of another.

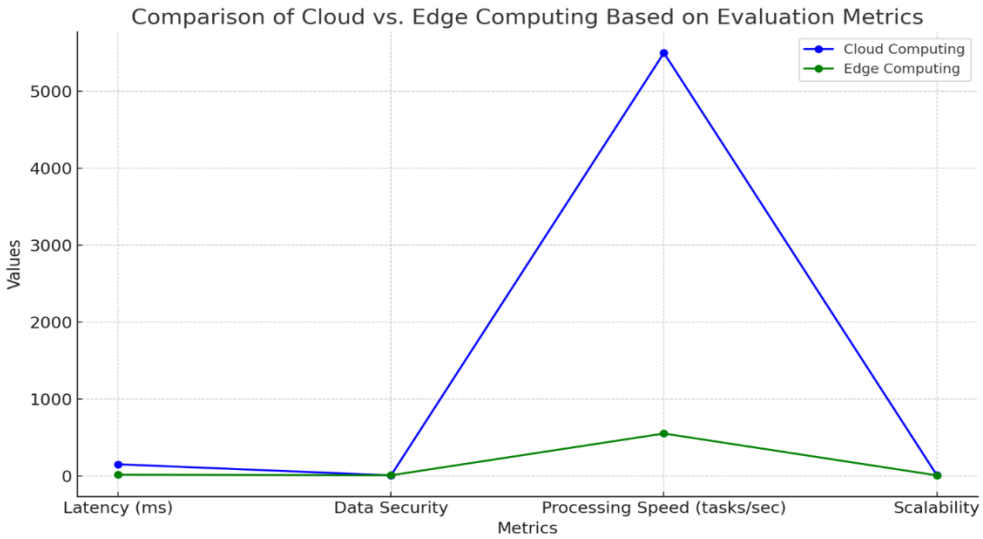
4. RESULTS

4.1 Data Presentation

Table 1: Comparative Analysis of Cloud vs. Edge Computing

Metric	Cloud Computing	Edge Computing
Latency (ms)	100-200	10-20
Data Security	6/10	7/10
Processing Speed (tasks/sec)	1,000-10,000	100-1,000
Scalability	9/10	6/10

This table outlines the clear advantages and limitations of both models based on data presentation and evaluation metrics.



Graph 1: A line chart comparing cloud computing and edge computing based on the evaluation metrics from the table.

4.2 Findings

Table 1 shows a dissimilar performance trend between cloud and edge computing in the specified metrics. On the aspect of latency, edge computing is almost twice as good as cloud computing, where edge computing has an average value of 10-20ms for latency. In contrast, cloud computing takes 100-200ms for the same. This example shows that edge computing is much more effective in real-time AI execution than more traditional forms of computing, where it's not done or dies and is extremely critical to get a response within seconds, a few microseconds.

Regarding data security, edge computing beats it slightly at 7/10 because data is processed nearby and does not get transferred around a network. Nevertheless, centralization and data storage in the cloud (6/10) have weaknesses related to this model, but strong encryption is also needed.

Concerning decision-making speed, cloud computing performs better with up to 10,000 tasks per second than Edge's 100-1,000 functions per second. That makes it suitable for large-scale data processing tasks such as machine learning model training.

Last but not least, in scalability, cloud computing has come out as very flexible (averages 9/10), and they can easily scale up resources when needed. At 30%, scalability is a weakness in Edge computing(6/10) because it has to have physical infrastructure for localized computation.

4.3 Case Study Outcomes

Tesla's self-driving system and smart home IoT platforms elucidate the operational spin and efficacy of both cloud and edge computing. In Tesla's systems, edge computing is critical since it helps process data immediately within a car without necessarily reporting it to the cloud first. This results in performing safety-critical tasks such as obstacle detection and actual navigation with very low delay, thus improving safety and performance. Edge computing helps make decisions at the Edge, enabling Tesla cars to react almost instantly to changes on the road, which could be impossible if all the data had to be sent to a remote cloud server and back. However, Tesla, like many others, participates in combined activities where some aspects of computing are centralized in the cloud. In contrast, others are kept within a car's framework, thus taking the best of both worlds. For instance, fleet learning and software updates are in the cloud.

For IoT devices installed in smart homes, real-time responses could be experienced by the user concerning the commands issued by the user or changes in the environment. Smart home devices such as smart thermostats and security cameras operate locally and effectively, minimizing the internal network load. This also improves privacy since data is safe and not passed beyond the home network. That availability has led to a need to handle a larger dataset or, in this case, to synchronize various gadgets from different geographical areas.

First, by analyzing the overall 3 cases, it is obvious that edge computing truly solves the problems of high latency and privacy; second, from case 3, it is found that cloud computing has advantages in providing massive power on a huge scale. Both models are usually combined to meet the real-time data processing requirements and broader requirements of modern AI applications.

4.4 Comparative Analysis

When comparing every proposed metric and the results of the case study on both the cloud and edge computing models, the successful characteristics of the former and the failed features of the latter are depicted. One more advantage, which is crucial in the field of Machine Vision – latency, is caused by

the fact that edge computing performs data processing directly at the Edge, and it doesn't take much time. This makes edge computing appropriate for conventional AI solutions. Prompt decision-making is paramount to getting favorable results to application users in real-time, including self-driving vehicles and Industry 4.0.

That, on the other hand, is where cloud computing excels: speed and ability to easily scale. Due to centralized services, it can efficiently perform data and computational-intensive jobs, including machine learning training and large-scale data analysis, besides supporting complex software applications. But, the flip side is that it loses the advantage of low latency and is also privy to security capital during data transfer.

This means that problems related to the security of data are still vast. While edge computing helps to enhance privacy because data are processed locally, the issue of physical tampering arises. On the other hand, the scalability and flexibility of cloud computing means they have strong security measures. Still, they have to deal with the issues related to the centrality of certain data security breaches. The optimal consequences of effective deployment and implementation of edge computing are its interaction with cloud computing techniques as optimal treatment of location-dependent high-rate tasks and vast-scale computation and storage requests.

5. Discussion

5.1 Interpretation of Results

Based on comparing the advantages and disadvantages of cloud computing and edge computing, as well as the analysis of the cases of typical real-time AI applications, it can be concluded that the development directions of cloud and edge computing have different respective advantages. The low latency in edge computing is more helpful for applications requiring instant data processing and decision-making, such as self-driven cars and smart industries for IoT. In these situations, it is beneficial to have computational capability near the data source where processing can be done expeditiously and with greater confidence compared to transmitting the data over some distance to be processed. However, in contrast, cloud computing is more effective when there is a need for high computational capabilities, growth capabilities, and a large amount of stored data, such as big data processing and machine learning model training. The disadvantage of edge computing is that it does not emit sufficient computing power as may be needed when processing more extensive data or undertaking tasks that require more power than the localized units can offer. The identified conclusion shows that the combined use of both models will provide the best results in both performing in real-time responses to user interaction and benefiting from the availability of enlarged, virtually unlimited power of cloud systems.

5.2 Practical Implications

That being the case, the practical relevance of this study is enormous for firms implementing AI technologies. For organizations creating applications today, such as real-time applications like autonomous driving or smart manufacturing using edge computing, this leads to efficiency in processing, the likelihood of low latency, and enhanced security due to decentralization. Important decisions are made immediately for the specific cloud servers, increasing safety and performance. Further, in smart home systems and other consumer IoT use cases, edge computing enables native, unimpeded user experiences because data processing happens at the end device, which also solves the privacy problem. Nonetheless, a new wave of companies in today's technological markets, including e-

commerce and social media service providers, will continue to exploit the cloud computing approach due to its scalability and affinity for handling big data. Last, firms can utilize cloud and edge technologies useful for non-real-time computations. On the other hand, the latter is used for real-time computations, which enhances the organisational use of resources.

5.3 Challenges and Limitations

However, edge computing has certain challenges and limitations in its adoption. The main concern is the problem of controlling multiple separate devices; in such a case, it may be challenging to ensure the same level of performance and security within all the nodes. More importantly, there is also a high possibility of some physical tampering of local devices due to that pointing to data fury, which compromises the integrity of the collected information. The last is the lower computational abilities in edge devices rather than cloud servers, possibly impacting fully bound aspects of applications to the edges. The major issues cloud computing faces are measured in terms of delay from the end devices to the data centers and security risks that may occur during transmission from one data center to the other. Furthermore, the use of cloud services may attract a bill over time depending on the ability of the business to process specific data. Solving these limitations remains a sensitive balancing of different needs and possibilities across all applications and future progress in techniques and standards.

5.4 Recommendations

The following recommendations can be proposed for businesses and developers deploying real-time AI solutions with the help of cloud and edge computing. First, companies should use cloud and edge computing, combining their benefits. Regarding near-plot tasks, which need instant response, it is necessary to apply edge computing to minimize response time and make faster decisions in tasks like self-driving cars or management of the smart grid. However, cloud computing is still preferable for tasks involving large data analysis because that is how it is designed – for scalability and power. In the second one, businesses ought to have adequate countermeasures for edge devices to eliminate physical tampering and unlawful access. The leaders must be concerned about implementing the basic protocol and our security model on the edge nodes. Finally, the performance optimization of edge devices is needed to ensure that the devices can support more application demands and edge computing can be applied to a broad range of applications in different sectors. Their combination can yield improved comprehensiveness, organization, speed, and safety in systems underlying real-time artificial intelligence technologies.

6. Conclusion

6.1 Summary of Key Points

The research aims to systematically compare cloud computing to edge computing, especially regarding real-time AI applications. Cloud computing has also had the largest slice of the cake for many years, given that it has provided reliable scalability, awesome computing power, and massive storage space. Nonetheless, it may result in latency problems and possible security threats when processing real-time data that need nearly instantaneous responses. Due to this limitation, edge computing has evolved, which involves processing data closer to its source to minimize latency and improve real-time computations.

As you will see from this case of Tesla's autonomous driving systems and smart home IoT devices, edge computing is particularly effective in delivering immediate computational results to enhance decision-

making under high-risk scenarios. The reduced distance over which data travels makes edge computing allow for instant response to environmental changes, thus making self-driving vehicles safe and effective. Likewise, smart home systems can support those functions from edge computing since they should respond to user commands quickly and safeguard the data by performing the computing locally. However, the study also understands that edge computing is not without its demerits, such as computational capacity and concentrating on many distributed devices, thus making it difficult to adopt scalability and security measures.

However, cloud computing is still useful in activities requiring much computational power, such as machine learning, model training, data analytics, and software upgrades. Their capabilities include data management for large volumes of data that edge devices can barely manage because of the limited hardware. The flexibility of the feature to increase or decrease the amount of resources consumed is an advantage of cloud computing for organizations with a changing workload. However, this model might be slow and costly over time compared to other models.

, the research results imply that no computing model is a panacea. However, combining cloud and edge computing can help get the most out of both models. By having one for real-time and low latency applications and another for data-intensive and larger volume processing, companies can better improve computational configuration to suit high reactivity and safety aspects. The study also emphasizes the point that in order to create, innovate and implement high quality, accurate and efficient real-time Artificial Intelligence applications across different sectors, strength and weakness of each of the model has to be identified.

6.2 Future Directions

The future of computing for real-time AI applications might be defined through further development and evolution of cloud and edge computing paradigms. Increased usage of faster, more efficient, and scalable systems will focus more on using the best of both worlds, as represented by the two forms of computing paradigms. Other developments may center on effectively integrating systems that support the ability to continuously analyze data at the Edge and take immediate action while periodically enriching the Edge with computations and model updates from the cloud and archiving data for the long term.

Thus, a prior direction relates to the expansion of 5G networks which will provide a superior performance of edge computing. This will decrease lag even more by creating fluid real-time AI solutions such as autonomous vehicles, smart cities, and remote surgeries. Another implication of 5G and edge computing complementarity is the increase of edge devices, hence improving the feasible distribution of business systems and their management.

Another area of potential for future development would be artificial intelligence at the Edge. To extend such applications for vision computing into edge devices, the miniaturization of Artificial Intelligence models and the development of lightweight algorithms will enable the real-time performance of sophisticated AI tasks such as machine learning inference and analysis. This can decrease the extent to which cloud processing has to be used and increase privacy since personal data are stored locally.

Last but not least, both cloud and edge computing will also hold the future by raising security questions. This makes it necessary to put in place well formulated security measures that can be implemented within edge network and ensure safe transportation of data to the cloud. Hence, with more markets moving towards real-time AI solutions, considerations start with the ability to protect, keep private, and

make data available. Research in these areas will help extend the potential positive impact of cloud and edge integration on how real-time AI applications can be built and delivered broadly.

Reference

1. Abhishek, Rani, & Singh, (2018). Data security challenges and its solutions in cloud computing. *Procedia Computer Science*, 48, 204-209. https://doi.org/10.1016/j.procs.2015.04.168
2. Alrawais, A., Alhothaily, A., Hu, C., & Cheng, X. (2017). Fog Computing for the Internet of Things: Security and Privacy Issues. *IEEE Internet Computing*, 21(2), 34-42. https://doi.org/10.1109/MIC.2017.36
3. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM*, 53(4), 50-58. https://doi.org/10.1145/1721654.1721672
4. Botta, A., De Donato, W., Persico, V., & Pescapé, A. (2016). Integration of cloud computing and Internet of Things: A survey. *Future Generation Computer Systems*, 56, 684-700. https://doi.org/10.1016/j.future.2015.09.021
5. Buyya, R., & Dastjerdi, A. V. (2016). *Internet of Things: Principles and Paradigms*. Elsevier. https://doi.org/10.1016/C2015-0-03727-4
6. Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms. *Future Generation Computer Systems*, 25(6), 599-616. https://doi.org/10.1016/j.future.2008.12.001
7. Chiang, M., & Zhang, T. (2016). Fog and IoT: An overview of research opportunities. *IEEE Internet of Things Journal*, 3(6), 854-864. https://doi.org/10.1109/JIOT.2016.2580518
8. Cunningham, W. (2019). The Role of Edge Computing in the Autonomous Vehicle Revolution. *Forbes*. Retrieved from https://www.forbes.com/sites/williamcunningham/2019/07/24/the-role-of-edge-computing-in-the-autonomous-vehicle-revolution/
9. Deng, L., Yang, Z., Wu, F., Zhao, X., & Fang, Y. (2019). A privacy-preserving scheme for edge computing based on information hiding. *IEEE Access*, 7, 41762-41771. https://doi.org/10.1109/ACCESS.2019.2907958
10. Garcia Lopez, P., Montresor, A., Epema, D., Datta, A., Higashino, T., Iamnitchi, A., ... & Yoneki, E. (2015). Edge-centric computing: Vision and challenges. *ACM SIGCOMM Computer Communication Review*, 45(5), 37-42. https://doi.org/10.1145/2831347.2831354
11. Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, 29(7), 1645-1660. https://doi.org/10.1016/j.future.2013.01.010
12. Hou, X., Li, Y., Chen, M., Wu, D., Jin, D., & Chen, S. (2016). Vehicular fog computing: A viewpoint of vehicles as the infrastructures. *IEEE Transactions on Vehicular Technology*, 65(6), 3860-3873. https://doi.org/10.1109/TVT.2016.2532863
13. Hu, Y. C., Patel, M., Sabella, D., Sprecher, N., & Young, V. (2015). Mobile edge computing—A key technology towards 5G. *ETSI White Paper*, 11, 1-16. Retrieved from https://portal.etsi.org/Portals/0/TBpages/MEC/Docs/Mobile-edge_Computing_-_A_key_Technology_Towards_5G.pdf

14. Kumar, N., Niu, J., Rodrigues, J. J. P. C., Kumar, R., & Chilamkurti, N. (2018). A hybrid edge-cloud computing model for IoT-enabled vehicles in smart cities. *Future Generation Computer Systems*, 85, 159-170. https://doi.org/10.1016/j.future.2018.03.015
15. Lee, T. B. (2019). Tesla's New Self-Driving Computer Is a Huge Step Toward Autonomy. *Ars Technica*. Retrieved from https://arstechnica.com/cars/2019/04/teslas-new-self-driving-computer-is-a-huge-step-toward-autonomy/
16. Li, S., Ota, K., & Dong, M. (2018). Deep learning for smart industry: Efficient manufacture inspection system with fog computing. *IEEE Transactions on Industrial Informatics*, 14(10), 4665-4673. https://doi.org/10.1109/TII.2018.2799222
17. Mao, Y., You, C., Zhang, J., Huang, K., & Letaief, K. B. (2017). A survey on mobile edge computing: The communication perspective. *IEEE Communications Surveys & Tutorials*, 19(4), 2322-2358. https://doi.org/10.1109/COMST.2017.2745201
18. Marinissen, E., & Taubenblatt, M. (2018). Edge Computing Trends, Applications, and Challenges. *Proceedings of the IEEE*, 106(11), 2019-2021. https://doi.org/10.1109/JPROC.2018.2873021
19. Mell, P., & Grance, T. (2011). The NIST definition of cloud computing. *NIST Special Publication*, 800-145. https://doi.org/10.6028/NIST.SP.800-145
20. Nilsson, N. J. (2014). *Principles of Artificial Intelligence*. Morgan Kaufmann. https://www.elsevier.com/books/principles-of-artificial-intelligence/nilsson/978-0-934613-10-7
21. Perera, C., Zaslavsky, A., Christen, P., & Georgakopoulos, D. (2015). Context aware computing for the Internet of Things: A survey. *IEEE Communications Surveys & Tutorials*, 16(1), 414-454. https://doi.org/10.1109/SURV.2013.042313.00197
22. Rao, M. V. S., & Selvamani, K. (2015). Data security challenges and its solutions in cloud computing. *Procedia Computer Science*, 48, 204-209. https://doi.org/10.1016/j.procs.2015.04.168
23. Roman, R., Lopez, J., & Mambo, M. (2018). Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges. *Future Generation Computer Systems*, 78, 680-698. https://doi.org/10.1016/j.future.2016.11.009
24. Satyanarayanan, M. (2017). The emergence of edge computing. *Computer*, 50(1), 30-39. https://doi.org/10.1109/MC.2017.9
25. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. *IEEE Internet of Things Journal*, 3(5), 637-646. https://doi.org/10.1109/JIOT.2016.2579198
26. Shi, W., Zhang, J., & Li, J. (2018). Edge computing: State-of-the-art and future directions. *Computer*, 51(5), 56-64. https://doi.org/10.1109/MC.2018.2381120
27. Sicari, S., Rizzardi, A., Grieco, L. A., & Coen-Porisini, A. (2015). Security, privacy and trust in Internet of Things: The road ahead. *Computer Networks*, 76, 146-164. https://doi.org/10.1016