

1

Why Do We Need Edge Computing?

What is edge computing? Why did it become popular after being proposed? What are the relationships between edge computing and IoT/Cloud Computing? In this chapter, we will answer these three questions by introducing the background, the evolutionary history, and the concept of edge computing.

1.1 The Background of the Emergence

To answer the question of this chapter, let us trace back to when edge computing was proposed, back to the big data era when the Internet of Things (IoT) and cloud computing were blooming.

The IoT technology [3] aims to connect physical objects to the Internet according to the communication protocols of IoT, utilizing technologies such as RFID (radio frequency identification), wireless data communication, and GPS (global positioning system). This enables information exchange for intelligent identification, positioning, tracking, monitoring, and management of Internet resources. IoT has significantly expanded with the advancement of computer and network communication technologies. It now encompasses the integration of almost all information technologies with computer and network technologies, facilitating real-time data sharing between objects and achieving intelligent real-time data collection, transmission, processing, and execution. The concept of “computer information perception without human intervention” has gradually been applied to fields such as wearable devices, smart homes, environmental sensing, intelligent transportation systems, and smart manufacturing [18, 36]. Key technologies involved in IoT include:

- **Sensor Technology:** This involves acquiring information from natural sources, processing (transforming), and identifying it. Sensor technology is a critical

aspect of computer applications, as it senses (or responds to) and detects specific information from the measured object, converting it into output signals according to certain rules.

- **RFID Technology:** This comprehensive technology integrates radio frequency and embedded technologies to automatically identify target objects and obtain related data through radio frequency signals. The identification process does not require human intervention and can operate in various harsh environments, with promising and broad applications in automatic identification, logistics management, and more.
- **Embedded System Technology:** This is a complex technology that integrates computer hardware and software, sensor technology, integrated circuit technology, and electronic application technology. Over the decades, intelligent terminal products characterized by embedded systems have become ubiquitous, ranging from smartwatches to aerospace satellite systems. Embedded systems are transforming people's lives, driving industrial production, and advancing the defense industry. If we make a simple analogy of the IoT to the human body, sensors are akin to human senses like eyes, nose, and skin; the network is the nervous system transmitting information, and the embedded system is the brain that classifies and processes the received information.

Later on, with the rapid development of IoT and the widespread adoption of 4G/5G wireless networks, the era of the Internet of Everything (IoE) [11] has arrived. Cisco introduced the concept of IoE in December 2012. It represents a new network architecture for future Internet connectivity and the evolution of IoT, enhancing the network's intelligent processing and security features. IoE employs a distributed structure, integrating application-centric networks, computing, and storage on a new platform. It is driven by IP settings, global higher bandwidth access, and IPv6, supporting hundreds of millions of edge terminals and devices connected to the Internet. Compared to IoT, IoE not only involves "thing-to-thing" connections, but also introduces a higher level of "human-to-thing" connectivity. Its distinguishing feature is that any "thing" will possess contextual awareness, enhanced computing capabilities, and sensing abilities.

Integrating humans and information into the Internet, the network will have billions or even trillions of connected nodes. The IoE is built on the physical network, enhancing network intelligence to achieve integration, coordination, and personalization among the "things" on the internet.

Application services based on the IoE platform require shorter response times and will generate a large amount of data involving personal privacy. For example, sensors and cameras installed on autonomous vehicles capture road condition information in real time; one car with five cameras can generate

more than 24 terabytes (TB) data per day [17]. According to the Insurance Institute for Highway Safety, there will be 3.5 million self-driving vehicles on U.S. roads by 2025 and 4.5 million by 2030 [21]. The Boeing-787 generates about 5 gigabytes (GB) of data per second and requires real-time processing of the data. In Beijing, China, the electric vehicle monitoring platform can provide continuous 7×24 -hour real-time monitoring for 10,000 electric vehicles and forward data to various enterprise platforms at a rate of one data point every 10 seconds per vehicle. In terms of social security, the United States has deployed over 30 million surveillance cameras, generating more than 4 billion hours of video data each week. China's "Skynet" surveillance network, used for crime prevention, has installed over 20 million high-definition surveillance cameras nationwide, monitoring and recording pedestrians and vehicles in real time.

Since the concept was proposed in 2005, cloud computing has been widely applied, changing how people work and live. SaaS (Software as a Service) is commonly used in data centers of major IT companies like Google, Twitter, Facebook, and Baidu. Scalable infrastructure and processing engines supporting cloud services have significantly impacted application services such as Google File System (GFS), MapReduce programming model, Hadoop (a distributed system developed by Apache Foundation), and Spark (the in-memory computing framework designed by the AMP Lab at the University of California Berkeley). However, in the context of IoT and similar applications, data is geographically dispersed and demands higher response times and security. Although cloud computing provides an efficient platform for big data processing, the network bandwidth growth rate cannot keep up with the data growth rate. The cost reduction rate of network bandwidth is much slower than that of hardware resources like CPU and memory, and the complex network environment makes it challenging to significantly improve network latency. Therefore, the traditional cloud computing model will struggle to support application services based on IoE efficiently and in real time, requiring solutions to address the bandwidth and latency bottlenecks.

With the rapid development and widespread application of the IoE, edge devices are transitioning from primarily serving as data consumers to serving as both data producers and consumers. Simultaneously, network edge devices are gradually capable of utilizing the collected real-time data for pattern recognition, predictive analysis or optimization, and intelligent processing. In the edge computing model, computing resources are closer to the data source, and network edge devices now have sufficient computational power to process the raw data locally and send the results to the cloud computing center locally. The edge computing model not only reduces the bandwidth pressure in network transmission, speeding up data analysis and processing, but also lowers the risk of privacy leaks for sensitive terminal data.

Currently, big data processing is shifting from the centralized processing era centered on cloud computing (we refer to the years from 2005 to 2015 as the centralized big data processing era) to the edge computing era centered on the IoE (we refer to it as the edge-based big data processing era). During the centralized big data processing era, the focus was more on centralized storage and processing of big data, achieved by building cloud computing centers and leveraging their powerful computing capabilities to solve computational and storage issues centrally. In contrast, in the edge-based big data processing era, network edge devices generate massive real-time data. In 2018, Cisco's Global Cloud Index estimated that nearly 850 zettabytes (ZB) will be generated by all people, machines, and things by 2021. Yet only around 10% is classed as useful data; useful data is predicted to four times exceed data center traffic (21 ZB per year) [10]. From 2018 to 2023, the average number of devices owned per person worldwide increased from 2.4 to 3.6. Specifically, in North America, on average, one person owned eight devices in 2018 and 13 devices in 2023 [38]. According to Statista, the number of IoT devices connected to the network was 15.14 billion in 2023 and will reach 29.42 billion in 2030 [35]. This mismatch between data producing and data consuming requires the emergence of an alternation for cloud-based data centers. Instead of purely relying on cloud computing, data can be stored, processed, and analyzed at the network edge. These edge devices will be deployed on edge computing platforms supporting real-time data processing, providing users with numerous service or function interfaces, which users can invoke to obtain the necessary edge computing services.

Therefore, the linearly growing centralized cloud computing capacity can no longer match the exponential growth of massive edge data. Single computing resources based on the cloud computing model can no longer meet the demands for real-time processing, security, and low energy consumption in big data processing. Based on the existing centralized big data processing centered on the cloud computing model, there is an urgent need for edge big data processing technology centered on the edge computing model to handle the vast edge data. The two complement each other, applied to big data processing at both the cloud center and the edge end, addressing the inadequacies of cloud computing services in the IoE era.

When observing the data explosion in three dimensions: velocity, variety, and volume, we will find that the emergence and rapid development of edge computing is inevitable (Figure 1.1). The cloud-centralized computing paradigm performs well when the data is generated with a confined speed, size, and format. While in the IoE era, data is increasingly produced at the network's edge regarding velocity, variety, and volume. In terms of variety, different types of data (e.g., text, audio, and photo) are generated every day and every second from numerous devices (e.g., IoT, web browser, camera, and social media). These data are generated with different

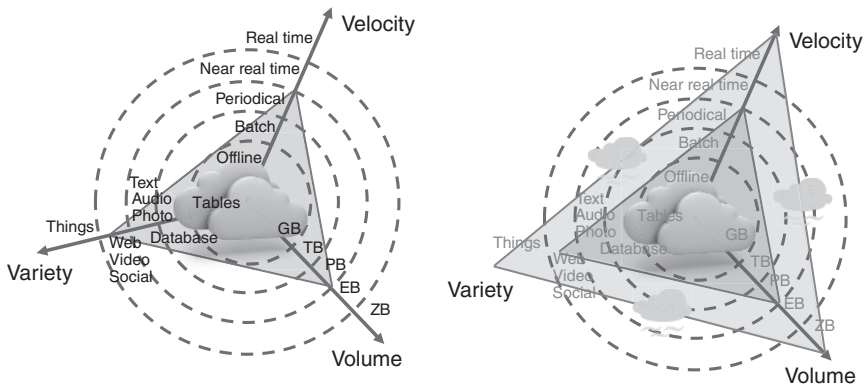


Figure 1.1 The increase of data pushes the evolution of edge computing.

velocities (e.g., real time, near real time, periodic, and batch generated). Therefore, storage and process requirements for these data will be different. With the tremendous number of devices and the frequent speed of data generation, there is no surprise that the volume of data generated is increasing dramatically. Megabytes (MB) and TB have become the typical units. In fact, as of 2024, the amount of data generated per day is around 328.77 million TB, which equals 0.33 ZB [13]. Given these factors, relying purely on the cloud for all data processing is impossible. This is not just because of the computing pressure brought to the cloud computation center but also because the bandwidth capability required for transmitting this amount of data is challenging. The only solution to process the data reliably and in a timely manner is edge computing, which ensures a shorter response time, more efficient processing, and smaller network pressure.

Compared to cloud computing, edge computing can better support mobile computing and IoT applications, offering the following distinct advantages:

- Greatly Alleviates Network Bandwidth and Data Center Pressure:** With the development of IoT, global devices will generate massive amounts of data. However, only a small portion of this data is critical, while most of it is temporary and does not need long-term storage (the amount of data generated by devices is two orders of magnitude higher than the amount of data that needs to be stored). Edge computing can fully utilize geographically distributed network edges to process a large amount of temporary data, thereby reducing the pressure on network bandwidth and data centers.
- Enhances Service Responsiveness:** The inherent limitations of mobile devices in computing, storage, and power resources are evident. Cloud computing can provide services to mobile devices to address these deficiencies. However, network transmission speeds are constrained by the development of

communication technologies, and in complex network environments, issues such as unstable connections and routing further exacerbate latency, jitter, and slow data transmission speeds, severely affecting the responsiveness of cloud services. Edge computing offers services near the user, ensuring low network latency through proximity and reducing network jitter with more straightforward routing. With the development of 5G and 6G, the diverse application scenarios and differentiated service requirements pose challenges to 5G/6G networks regarding throughput, latency, number of connections, and reliability. Edge computing and 5G/6G technologies complement each other, with edge computing leveraging localization, proximity, and low latency to drive 5G/6G architectural changes, while 5G/6G technology is essential for reducing data transmission latency and enhancing service responsiveness in edge computing systems.

- **Protects Privacy Data and Enhances Data Security:** Data security has always been a critical issue in IoT applications. Surveys show that approximately 86% of the U.S. general population is concerned about data privacy [23]. In the cloud computing model, all data and applications are stored in data centers, making it difficult for users to have fine-grained control over data access and usage. Edge computing provides the infrastructure for storing and using critical privacy data, restricting the operation of privacy data within firewalls and thereby enhancing data security (for more detailed information, see Chapter 8).

1.2 The Evolutionary History

The field of edge computing has developed rapidly since 2014. We categorize the development process into three stages: technology preparation period, rapid growth period, and steady development period. We use “edge computing” as the keyword to search the number of articles published per year in Google Scholar. As shown in Figure 1.2, before 2015, edge computing was in the technology preparation period. Since 2015, the number of papers related to “edge computing” has grown tenfold. Edge computing has entered a rapid growth period. The number of papers has been increasing and reaching a steady development period since 2020. In this period, the development is focused on integrating academia and industry, bringing the product into the business, and finally facilitating peoples’ daily lives. Figure 1.3 illustrates typical events in the development process of edge computing.

The development of edge computing is closely linked to the evolution of data-oriented computing models. As the scale of data increases, the demand for performance and energy efficiency in data processing continues to grow.

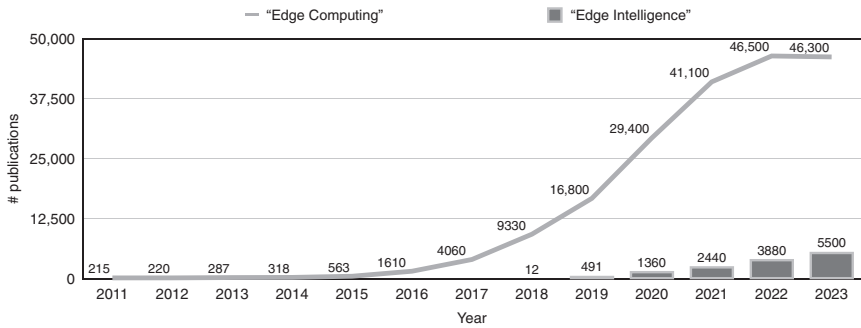


Figure 1.2 Number of publications searched by keyword “edge computing” and “edge intelligence.”

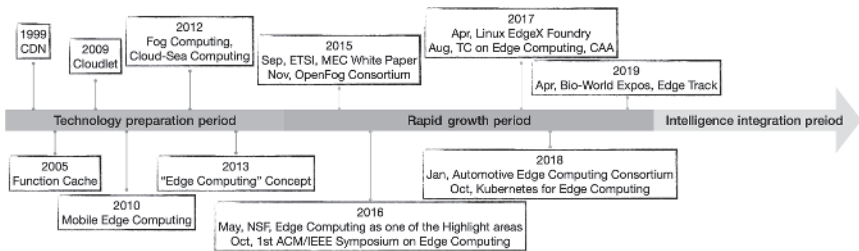


Figure 1.3 The evolution of edge computing and key milestones.

To address the issues of computational load and data transmission bandwidth in data transfer, computation, and storage processes, researchers explored ways to enhance data-processing capabilities near data sources even before the advent of edge computing. This involves shifting computational tasks from centralized computing centers to the network edge. The main typical models include distributed database models, peer-to-peer (P2P) computing models, content delivery network (CDN) models, mobile edge computing models, fog computing models, and cloud-sea computing models. We will explain these different models in the order of their emergence and also introduce the history of edge computing.

1.2.1 Technology Preparation Period

During the technology preparation period, edge computing went through the development process of dormancy, presentation, definition, and generalization.

1.2.1.1 Distributed Database Models

The distributed database model results from combining database technology and network technology. In the era of big data, the growth in the variety and quantity

of data has made distributed databases a core technology for data storage and processing. Distributed databases are deployed on self-organizing network servers or dispersed across the Internet, enterprise networks, Internet, and other independent computers in self-organizing networks. Data is stored on multiple machines, and operations are not limited to a single machine but allow transactions to be executed across multiple machines to improve database access performance.

Distributed databases have become a core technology for big data processing. Based on their structure, distributed databases include homogeneous and heterogeneous systems. The former has database instances running in environments with the same software and hardware, featuring a single-access interface. The latter operates in environments where hardware, operating systems, database management systems, and data models vary. Based on the types of data processed, distributed databases mainly include relational (such as Structured Query Language, SQL), nonrelational (such as NoSQL), extensible markup language (XML)-based, and NewSQL distributed databases. Among these, NoSQL and NewSQL distributed databases are the most widely used [16]. NoSQL distributed databases, designed to meet the demands for high concurrency, efficient storage access, high reliability, and scalability in big data environments, are divided into key-value stores, column stores, document-oriented databases, and graph databases. NewSQL distributed databases, characterized by real-time processing, complex analysis, and fast querying, are relational distributed databases designed for massive data storage in big data environments, including Google Spanner, Clustrix, and VoltDB. SQL-distributed databases are relational distributed databases, with typical examples including Microsoft's and Oracle's distributed databases. XML-based distributed databases mainly store data in XML format and are essentially document-oriented, similar to NoSQL distributed databases [22].

Compared to edge computing models, distributed databases provide data storage in big data environments but pay less attention to the heterogeneous computing and storage capabilities of the devices they reside on, focusing mainly on achieving distributed data storage and sharing. Distributed database technology requires significant space and offers lower data privacy. For distributed transaction processing across multiple databases, data consistency technology is a major challenge for distributed databases [14]. In edge computing models, data resides on edge devices, offering higher privacy, reliability, and availability. In the era of the IoE, "heterogeneous edge architectures and the need to support multiple application services" will become the fundamental approach for edge computing models to handle big data processing.

1.2.1.2 Peer-to-Peer (P2P) Computing Models

P2P computing [27] is one of the early file transfer technologies that pushed computing to the edge of the network. The term *P2P* was first introduced in

2000 to implement file-sharing systems. Since then, it has gradually developed into an important subfield of distributed systems. The key research topics in P2P models include decentralization, maximizing scalability, tolerance of high-level node churn, and preventing malicious behavior. Major achievements in this field include:

- Distributed Hash Table (DHT), which later evolved into the general paradigm for key-value distributed storage in cloud computing models.
- Generalized gossip protocols, which have been widely used for complex task processing applications beyond simple information dissemination, such as data fusion and topology management.
- Multimedia streaming technology, in forms such as video on demand, real-time video, and personal communication.

However, widespread media coverage of P2P being used for illegal file sharing and related lawsuits has hindered the practical recognition of some commercial technologies based on the P2P model.

The edge computing model bears significant similarities to P2P technology, while it expands on the latter with new technologies and methods, extending the concept of P2P to network edge devices. This represents a fusion of P2P computing and cloud computing.

1.2.1.3 Content Delivery Network (CDN) Models

CDN [29] was proposed by Akamai in 1998. CDN is an Internet-based caching network, which relies on caching servers deployed in different places and points users' access to the nearest caching server through load balancing, content distribution, scheduling, and other functional modules of the central platform. Therefore, CDN can reduce network congestion and improve user access response speed and hit rate. It has gained significant attention from both academia and industry since it was proposed. Companies like Amazon [2] and Akamai [1] possess mature CDN technologies that provide users with the expected performance and experience while reducing the operational pressures on service providers.

Active content distribution networks (ACDNs), an improvement over traditional CDNs, help content providers avoid the hassle of predicting the preconfiguring resources and determining their locations [30]. ACDN allows applications to be deployed on any server and uses newly designed algorithms to replicate and migrate applications between servers as needed.

The concept of edge computing can be traced back to around the year 2000, when CDNs were deployed large scale. At that time, major companies like Akamai announced the distribution of web-based content through CDN edge servers. The primary goal of this method was to benefit from the short distances and available resources of CDNs to achieve large-scale scalability. In the early

days of edge computing, the “edge” was limited to CDN cache servers distributed around the world. However, today’s development of edge computing has far exceeded the scope of CDNs. The “edge” in the edge computing model is not confined to edge nodes; it includes any computational, storage, and networking resources along the path from data sources to cloud computing centers.

1.2.1.4 Function Cache and Cloudlet

To enable static content distribution, CDN emphasizes the backup and caching of data, while edge computing focuses more on function caching to improve computational capabilities. Function cache was proposed by Ravi et al. [31], where it is applied to personalized mailbox management services to save latency and bandwidth. Satyanarayanan et al. [33] introduced the concept of Cloudlet, which is a trusted, small-scale, and resource-rich host, located at the edge of the network, connected to the Internet, and can be accessed by mobile devices to provide services. Cloudlet is also known as “small cloud” as it can provide services for users, similar to the cloud server. At this point, edge computing focused on the downstream transfer of functions from cloud servers to edge servers, aiming to reduce bandwidth usage and minimize delays.

1.2.1.5 Mobile Edge Computing

The development of the IoE has enabled the interconnection of numerous types of devices, such as smartphones, tablets, wireless sensors, and wearable devices. However, the limited energy and computing resources of most network edge devices make the design of IoE particularly challenging. Mobile edge computing (MEC) [19] is a new network architecture that provides information technology services and cloud computing capabilities within the proximity of the mobile user’s wireless access network. It has become a standardized and regulated technology. In 2014, the European Telecommunications Standards Institute (ETSI) introduced the standardization of the term MEC, highlighting that MEC provides a new ecosystem and value chain. Utilizing MEC, intensive mobile computing tasks can be offloaded to nearby network edge servers. Because MEC is located within the wireless access network and close to mobile users, it can achieve lower latency and higher bandwidth, thereby improving service quality and user experience. MEC is also a key technology in the development of 5G, helping to meet the high standards of 5G in terms of latency, programmability, and scalability. By deploying services and caches at the network edge, MEC reduces congestion in the core network and efficiently responds to use requests.

Task migration is one of the challenges in mobile computing technology, particularly in environments where continuous service availability and seamless user experience are crucial. The process involves transferring ongoing tasks from one computational node to another, which can be triggered by various factors such

as device mobility, energy conservation needs, or load balancing requirements. Effective task migration must minimize latency, avoid data loss, and maintain application state continuity, which is challenging due to the heterogeneous and dynamic nature of mobile environments. Furthermore, ensuring security during data transfer, managing the energy consumption of mobile devices, and dealing with fluctuating network conditions are additional hurdles that need optimization solutions. As mobile computing continues to evolve, developing robust, efficient, and secure task migration mechanisms will be critical to fully leveraging the potential of mobile platforms. MEC has been applied in various scenarios, such as vehicular networks, IoT gateways, auxiliary computing, intelligent video acceleration, and mobile big data analysis.

MEC emphasized the establishment of edge servers between the cloud server and edge devices to process computing. However, mobile edge nodes are generally considered to lack computing capabilities. In contrast, the nodes in the edge computing model possess strong computing capabilities. Therefore, MEC resembles the architecture and hierarchy of an edge computing server, functioning as an important part of edge computing.

1.2.1.6 Fog Computing

Cisco introduced fog computing in 2012 and defined fog computing as a highly virtualized computing platform for migrating cloud computing center tasks to network edge devices [7]. Fog computing provides computing, storage, and network services between end devices and traditional cloud computing centers, complementing cloud computing. Vaquero and Roderio-Merino [39] have provided a comprehensive definition of fog computing, which extends cloud-based network architecture by introducing an intermediate layer between the cloud and mobile devices. This intermediate layer, known as the fog layer, consists of fog servers deployed at the network edge. Fog computing reduces the need for multiple communications between the cloud computing center and mobile users. It relieves the bandwidth load and energy consumption pressure of main links by reducing the number of communications between cloud computing centers and mobile users. When there is a large volume of mobile users, they can access cached content and request specific services from the fog computing servers. Additionally, fog computing servers can interconnect with cloud computing centers, leveraging their powerful computational capabilities and extensive applications and services.

The concepts of edge computing and fog computing have great similarities and often represent the same idea. If we are to distinguish between the two, this book posits that edge computing, in addition to focusing on infrastructure, also pays attention to edge devices and places more emphasis on the design and implementation of edge intelligence. In contrast, fog computing focuses more on the

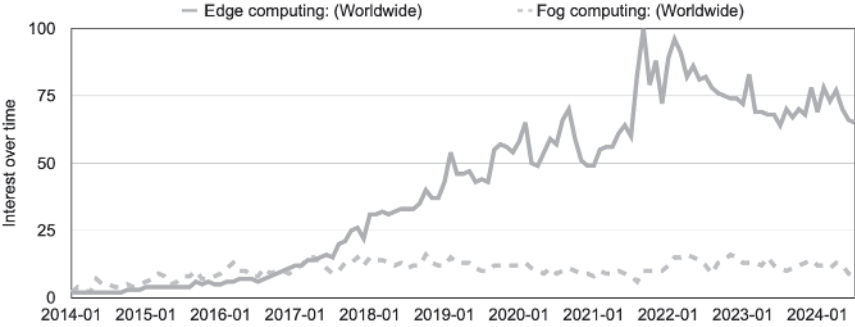


Figure 1.4 “Edge computing” and “fog computing” trends.

management of back-end distributed shared resources. As shown in Figure 1.4, since 2017, the level of attention to edge computing has gradually surpassed that of fog computing, and its attention continues to rise.

1.2.1.7 Cloud-Sea Computing

In the context of IoT, the amount of data to be processed will reach ZB levels. The sensing, transmission, storage, and processing capabilities of information systems need to be correspondingly enhanced. To address this challenge, in 2012, the Chinese Academy of Sciences launched a ten-year strategic priority research initiative called the Next Generation Information and Communication Technology (NICT) initiative. Its main purpose is to carry out research on the “Cloud-Sea Computing System Project” [40]. It aims to augment cloud computing by cooperation and integration of the “cloud computing” system and the “sea computing” system. “Sea” refers to an augmented client-side consisting of human-facing and physical world-facing devices and subsystems. The research focuses on proposing system-level solutions from perspectives such as overall system architecture, data center and server and storage system layers, and processor chip level.

Cloud-sea computing focuses on the two ends “sea” and “cloud” while edge computing focuses on the data path between “sea” and “cloud.” Cloud-sea computing is a great subset example of edge computing.

1.2.2 Rapid Growth Period

Since 2015, edge computing has been in a rapid growth period, attracting intensive close attention from academia and industry.

At the government level, in May 2016, the National Science Foundation (NSF) listed edge computing as one of the highlighted areas in the research of computer systems. In August 2016, NSF and Intel formed a partnership in information center

networks in wireless edge networks (ICN-WEN) [37]. In October 2016, the NSF held the NSF Workshop on Grand Challenges in edge computing [8]. The workshop focused on three topics: the vision of edge computing in the next five to ten years, the grand challenges to achieving the vision, and the best mechanisms for academia, industry, and the government to attack these challenges in a cooperative way. This indicates that the development of edge computing has attracted great attention at the government level.

In academia, a formal definition of edge computing is given in the paper *Edge computing: vision and challenges* [34]. Edge computing is defined as enabling technologies that allow computation to be performed at the edge of the network, processing downstream data on behalf of cloud services, and upstream data on behalf of IoT services. This paper pointed out the challenges of edge computing and is one of the most cited papers in the edge computing field. In October 2016, ACM and IEEE jointly organized the first ACM/IEEE Symposium on Edge Computing (SEC). Since then, the International Conference on Distributed Computing Systems (ICDCS), the International Conference on Computer Communications (INFOCOM), the International Middleware Conference, and other important international conferences have added an edge computing track and/or workshops to their main conferences.

At the same time, multiple industry sectors have actively promoted the development of edge computing. In September 2015, ETSI published a white paper on MEC. In November 2015, Cisco, ARM, Dell, Intel, Microsoft, and Princeton University jointly established the OpenFog Consortium, which is dedicated to the development of Fog Reference Architecture [28]. The OpenFog Consortium merged into the Industrial Internet-of-Things (IIoT) in January 2019. In November 2016, Huawei, Shenyang Institute of Automation of Chinese Academy of Sciences, China Academy of Information and Communications Technology (CAICT), Intel, ARM, and iSoftStone established the Edge Computing Consortium (ECC) in Beijing, China, which is dedicated to advancing cooperation among industry resources from government, vendor, academic, research, and customer sectors, and pushing forward the sustainable development of the edge computing industry [12]. In March 2017, the ETSI MEC Industry Specification Working Group was formally renamed to multiaccess edge computing, aiming to better meet the requirements of edge computing and related standards. Linux EdgeX Foundry was also built in 2017; it is a vendor-neutral open-source project hosted by The Linux Foundation. It aims to build a common open framework for IoT edge computing. In January 2018, Automotive ECC (AECC) was established to drive the network and computing infrastructure needs of automotive big data [4], which indicates that edge computing is valued in the vehicle domain. In the same year, the Cloud Native Computing Foundation (CNCF) Foundation and Eclipse Foundation cooperated to bring Kubernetes, which has been widely used

in the ultra large-scale cloud computing environment, into the edge computing scene of the IoT. Subsequently, KubeEdge, a Kubernetes native edge computing framework, was accepted into the CNCF sandbox in March 2019 [24]. In April 2019, the Bio-IT World Conference and Expos added the edge track [6], which means that edge computing is important to the health domain as well.

In the rapid growth period, the industry has seen significant advancements, evidenced by the availability of multiple edge environment solutions from major service providers. Today, options like AWS Greengrass [5], Microsoft Azure [25, 26], Google Cloud Platform Edge Zones [15] have made it easier for businesses and developers to deploy and manage edge computing infrastructures effectively. These developments underscore the maturity and widespread adoption of edge computing across various sectors.

1.2.3 Intelligence Integration Period

Edge computing has seen substantial growth and transformation in recent years, driven by the increasing demand for low-latency data processing and efficient resource utilization. The development of edge computing is characterized by bringing together IoT, big data, and mobile computing into an integrated and ubiquitous computing platform. The capability of delivering on-demand computing power at the edge and processing a vast amount of data from various devices/sensors enables real-time analytics and decision-making. A significant advancement within this domain is the integration of edge intelligence, where artificial intelligence (AI) and machine learning (ML) algorithms are deployed at the edge. This symbiotic relationship enhances the capability of edge computing, allowing for sophisticated data analysis and autonomous decision-making directly at the data source. Edge intelligence empowers devices to process and act on data locally, leading to smarter, faster, and more efficient systems across various industries, from autonomous vehicles to smart cities and beyond.

Building on this foundation, the intelligence integration period marks a crucial phase in the evolution of edge computing. Technological strategies such as pruning, quantization, and knowledge distillation are employed to optimize AI models for efficient operation on edge devices. Simultaneously, AI algorithms find wide application across systems such as smart surveillance, autonomous vehicles, health monitoring systems, industrial IoT, smart agriculture, and retail enhancements, further demonstrating the pervasive impact of this integration. These advancements not only improve responsiveness but also deliver substantial societal benefits. While the potential of these integrations is immense, the associated privacy and security concerns are non-negligible and will be discussed in Chapter 8, with deeper technological and application-based discussions slated for Chapters 3, 4, and 7.

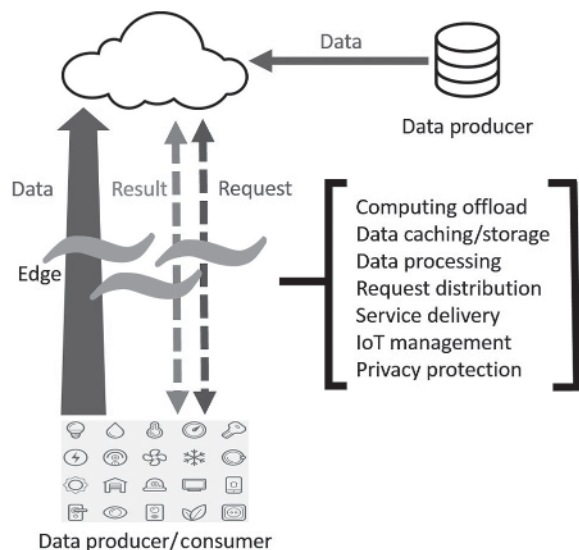
1.3 What Is Edge Computing?

After exploring the background of edge computing's emergence and examining its three distinct phases of development, it's time to address a fundamental question: what exactly is edge computing?

There is no standard definition for edge computing yet. In the field of edge computing, industry experts and other researchers have provided their own definitions. For example, IBM views edge computing as a distributed computing framework that brings enterprise applications closer to data sources such as IoT devices or local edge servers [20]. CISCO interprets edge computing as a model that shifts computing resources from central data centers or public clouds closer to devices, that is, embedded at the edge of service provider networks [9]. Satyanarayanan defines edge computing as a computing paradigm in which substantial computing and storage resources—variously referred to as cloudlets, micro data centers, or fog nodes—are placed at the Internet's edge in close proximity to mobile devices or sensors [32]. Yousefpour et al. believe edge computing is located at the edge of the network close to IoT devices, and edge can be more than one hop away from IoT devices in the local IoT network [41].

Here, we give our own definition of edge computing. In the vision paper published in 2016, we highlighted that **edge computing refers to the enabling technologies allowing computation to be performed at the edge of the network, on downstream data on behalf of cloud services and upstream data on behalf of IoT services.** [34]. As shown in Figure 1.5, in our definition,

Figure 1.5 Edge computing paradigm.



“edge” can be any computing and network resources along the path between data sources and cloud data centers. The rationale of edge computing is that **computing should happen at the proximity of data source, close to the users**. We can interpret the word “close” in two ways. First, the edge computing resource and the end users may be close in the communication network. Therefore, the small network size makes it more feasible to deal with network instability (e.g., bandwidth, delay, and jitter). Second, the resource and users may be close in spatial distance, which means they share similar environmental information. The computing resources may leverage the shared information to provide personalized services and improve the user experience. Network distance and spatial distance are not correlated to each other, and it may depend on the concrete scenarios to decide which type of close (or both) is appropriate.

If we view resources on the path between IoT services and cloud services as a continuum, *edge* can be any computing, storage, and network resources on this path. Depending on the specific requirements and concrete scenarios, the edge can be one or multiple resources (nodes), as shown in Figure 1.6. It can be a smartphone or a desktop serving as the edge between body things and the cloud, a gateway in a smart home as the edge between home things and the cloud. It can also be as small as embedded devices such as wearable sensors and security cameras, or as big as a micro data center. There are a huge amount of edge resources; they are scarcely distributed around end users, independent of each other. Edge computing is dedicated to unifying these resources that are close to end users in either the communication network or spatial distance and provides computing, storage, and network services for applications.

If we understand edge computing from a biological perspective, we can make the following analogy: cloud computing is akin to the human brain, while edge

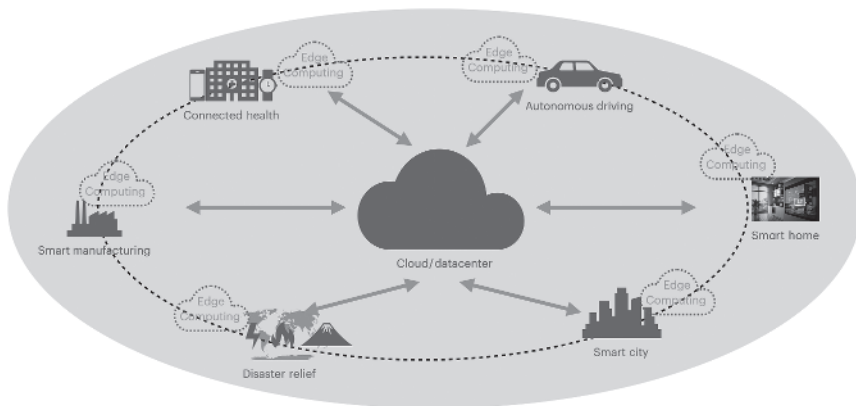


Figure 1.6 Edge computing is a continuum.

computing is akin to nerve endings. When a needle pricks the hand, a person instinctively pulls their hand back before the brain even realizes the prick because the reflex action is processed by the nerve endings. This reflex action speeds up the response, preventing further harm, while allowing the brain to focus on more complex tasks. In the future era of the IoE, it is impractical for cloud computing to act as the “brain” for every device. Instead, edge computing allows edge devices to have their own “brains.”

To have an overview of the development of edge computing, we investigate the time period when it got attention and became popular. If you search for “edge computing” in Google Trends, set the *time range* to be 01/01/2010–now, you will see a trend like Figure 1.7. “Interest over time” shows search interest relative to the highest point on the chart for the given region (in this example, we choose *Worldwide*) and time (from 2010 till 2024). A value of 100 means the term is in the peak popularity. A value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term. We can see that the attention to edge computing has been continuously increasing since 2016, reflecting its importance in the development of technology. In the years 2022 to 2023, there is a drop in “interest over time” for edge computing. At the same, another term, **Edge AI**, has been raising attention and has been getting more and more interest since 2023, when the large language model (LLM) started to dominate the artificial intelligence (AI) and machine learning (ML) market. The trend of these two words perfectly shows the focus of the research area in edge computing. Before LLM showed up, the research focus in this field was on the computing paradigm itself, such as the architectures and components of edge computing (Chapter 3), edge computing hardware and software (Sections 4.1 and 4.2), and challenges and solutions in edge computing (Chapter 5). With the development of AI and ML, especially LLMs, the world has witnessed the power of AI models. However, to make these powerful models accessible to the masses, the computing and storage resource constraints became a significant bottleneck. Therefore, the research



Figure 1.7 “Edge computing” and “edge AI” trends.

focus in this field has shifted to enabling edge-based AI by tackling the problem of resource constraints (Sections 4.3 and 4.4).

1.4 Summary and Practice

1.4.1 Summary

This chapter provides the definition and core concept of edge computing, emphasizing that edge computing is a continuum. The “edge” in edge computing refers to any computing, storage, and network resources along the path from the data source to the cloud computing center. The discussion of the development and challenges of big data processing and the Internet of Everything helps to understand the background of the emergence of this computing paradigm. This chapter also reviews the historical development of data-oriented computing models such as distributed databases, P2P, CDN, MEC, fog computing, and Cloud-Sea computing. Additionally, it introduces the current status of edge computing and its close connection with edge intelligence.

1.4.2 Practice Questions

- 1) What is the “edge”?
- 2) What are the main characteristics that distinguish edge computing from traditional cloud computing?
- 3) Identify and explain the challenges that traditional cloud computing faces in handling the large volumes of data generated by IoE devices.
- 4) Why is edge computing necessary in the era of the Internet of Everything?
- 5) Explain examples of real-world applications for each use case presented in Figure 1.6.
- 6) Based on the background and evolutionary history of edge computing discussed in this chapter, what do you think are the key challenges that could arise during the development of edge computing?

1.4.3 Course Projects

- 1) Analyze real-world case studies where edge computing is used to solve specific problems and understand why edge computing was necessary in each case. Case studies can be found from sources like <https://lfedge.org/>.
- 2) Conduct a comprehensive study of various open-source edge platforms to understand the capabilities, strengths, weaknesses, and potential use cases of each platform. The platforms to be researched could include, but are not

limited to: LF Edge Projects (EdgeX Foundry, Akraino, and Open Horizon), Kubernetes for edge (KubeEdge, K3s, and MicroK8s), OpenFaaS.

- 3) Explore a basic edge computing use case and present a simple prototype. For example, smart home, healthcare monitoring. The prototype will involve the design of data processing system architecture, and operate on an edge device.
- 4) Build a smart home environment that leverages edge computing to manage and control devices such as lights, temperature sensors, security cameras locally, instead of relying solely on cloud services.

Chapter 1 Suggested Papers

- 1 Mahadev Satyanarayanan. “The emergence of edge computing”. In: *Computer* 50. 1 (2017), pp. 30–39.
- 2 Weisong Shi et al. “Edge computing: Vision and challenges”. In: *IEEE Internet of Things Journal* 3. 5 (2016), pp. 637–646.
- 3 Weisong Shi, George Pallis, and Zhiwei Xu. “Edge computing [scanning the issue]”. In: *Proceedings of the IEEE* 107. 8 (2019), pp. 1474–1481.

References

- 1 Akamai. *Amazon CloudFront*. <https://www.akamai.com/solutions/content-delivery-network>. Accessed: 2024-07-24.
- 2 Amazon AWS. *Amazon CloudFront*. <https://aws.amazon.com/cloudfront/>. Accessed: 2024-07-24.
- 3 Kevin Ashton. “That ‘Internet of Things’ thing”. In: *RFID Journal* 22. 7 (2009), pp. 97–114.
- 4 Automotive Edge Computing Consortium. *Automotive Edge Computing Consortium*. <https://aecc.org/>. Accessed: 2024-07-31.
- 5 AWS. *IoT Edge*. <https://aws.amazon.com/greengrass/>. Accessed: 2024-08-26.
- 6 Bio-IT World. *bio-IT World Conference Edge Track*. <https://kubeedge.io/>. Accessed: 2024-07-31.
- 7 Flavio Bonomi et al. “Fog computing and its role in the Internet of Things”. In: *Proceedings of the 1st Edition of the MCC Workshop on Mobile Cloud Computing*. 2012, pp. 13–16.
- 8 Weisong Shi Mung Chiang. *NSF Workshop Report on Grand Challenges in Edge Computing*. <https://www.weisongshi.org/papers/shi16-nsfreport.pdf>. Accessed: 2024-07-30.
- 9 Cisco. *Edge Computing Solutions*. <https://www.cisco.com/c/en/us/solutions/service-provider/edge-computing.html>. Accessed: 2024-05-11.

- 10 Cisco. *Redefine Connectivity by Building a Network to Support the Internet of Things*. <https://www.cisco.com/c/en/us/solutions/service-provider/a-network-to-support-iot.html>. Accessed: 2024-05-25.
- 11 Laura DeNardis. *The Internet in everything*. Yale University Press, 2020.
- 12 Edge Computing Consortium. *Introduction of Edge Computing Consortium*. <http://en.eccconsortium.net/Uploads/file/20180328/1522232376480704.pdf>. Accessed: 2024-07-31.
- 13 Exploding Topics. *Amount of Data Created Daily (2024)*. <https://explodingtopics.com/blog/data-generated-per-day>. Accessed: 2024-05-14.
- 14 Iggy Fernandez. *No! to SQL and No! to NoSQL*. <https://iggyfernandez.wordpress.com/2013/07/28/no-to-sql-and-no-to-nosql/>. Accessed: 2024-07-24.
- 15 Google Cloud. *Google Distributed Cloud*. <https://cloud.google.com/distributed-cloud/edge/latest/docs>. Accessed: 2024-08-26.
- 16 Katarina Grolinger et al. “Data management in cloud environments: NoSQL and NewSQL data stores”. In: *Journal of Cloud Computing: Advances, Systems and Applications* 2 (2013), pp. 1–24.
- 17 Adam Grzywaczewski. *Training AI for Self-Driving Vehicles: the Challenge of Scale*. <https://developer.nvidia.com/blog/training-self-driving-vehicles-challenge-scale/>. Accessed: 2024-08-26.
- 18 Jayavardhana Gubbi et al. “Internet of Things (IoT): A vision, architectural elements, and future directions”. In: *Future Generation Computer Systems* 29. 7 (2013), pp. 1645–1660.
- 19 Yun Chao Hu et al. “Mobile edge computing—A key technology towards 5G”. In: *ETSI White Paper* 11. 11 (2015), pp. 1–16.
- 20 IBM. *What is edge computing?* <https://www.ibm.com/topics/edge-computing>. Accessed: 2024-05-11.
- 21 Insurance Information Institute. *Background on: Self-driving cars and insurance*. <https://www.iii.org/article/background-on-self-driving-cars-and-insurance>. Accessed: 2024-08-26.
- 22 Hosagrahar V Jagadish et al. “Timber: A native XML database”. In: *The VLDB Journal* 11 (2002), pp. 274–291.
- 23 KPMG. *Corporate data responsibility: Bridging the consumer trust gap*. <https://kpmg.com/us/en/articles/2023/bridging-the-trust-chasm.html>. Accessed: 2024-08-26.
- 24 KubeEdge Project Authors. *KubeEdge*. <https://kubeeedge.io/>. Accessed: 2024-07-31.
- 25 Microsoft Azure. *Azure Stack*. <https://azure.microsoft.com/en-us/products/azure-stack/>. Accessed: 2024-08-26.
- 26 Microsoft Azure. *IoT Edge*. <https://azure.microsoft.com/en-us/products/iot-edge/>. Accessed: 2024-08-26.
- 27 Dejan S Milojicic. *Peer-to-peer computing*. 2002.

- 28 OpenFog Consortium. *OpenFog Reference Architecture for Fog Computing*. https://www.iiconsortium.org/pdf/OpenFog_Reference:Architecture_2_09_17.pdf. Accessed: 2024-07-31.
- 29 Gang Peng. “CDN: Content distribution network”. In: *arXiv preprint cs/0411069* (2004).
- 30 Michael Rabinovich, Zhen Xiao, and Amit Aggarwal. “Computing on the edge: A platform for replicating internet applications”. In: *Web Content Caching and Distribution: Proceedings of the 8th International Workshop*. Springer. 2004, pp. 57–77.
- 31 Jayashree Ravi, Weisong Shi, and Cheng-Zhong Xu. “Personalized email management at network edges”. In: *IEEE Internet Computing* 9. 2 (2005), pp. 54–60.
- 32 Mahadev Satyanarayanan. “The emergence of edge computing”. In: *Computer* 50. 1 (2017), pp. 30–39.
- 33 Mahadev Satyanarayanan et al. “The case for VM-based cloudlets in mobile computing”. In: *IEEE Pervasive Computing* 8. 4 (2009), pp. 14–23.
- 34 Weisong Shi et al. “Edge computing: Vision and challenges”. In: *IEEE Internet of Things Journal* 3. 5 (2016), pp. 637–646.
- 35 Statista. *Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2023, with forecasts from 2022 to 2030*. <https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/>. Accessed: 2024-05-25.
- 36 Harald Sundmaeker et al. “Vision and challenges for realising the Internet of Things”. In: *Cluster of European Research Projects on the Internet of Things, European Commission* 3. 3 (2010), pp. 34–36.
- 37 U.S. National Science Foundation. *NSF 16-586: NSF/Intel Partnership on Information-Centric Networking in Wireless Edge Networks (ICN-WEN)*. <https://new.nsf.gov/funding/opportunities/nsfintel-partnership-information-centric/505310/nsf16-586/solicitation>. Accessed: 2024-07-30.
- 38 UN Trade and Development. *Digital economy report 2024*. <https://unctad.org/publication/digital-economy-report-2024>. Accessed: 2024-08-26.
- 39 Luis M Vaquero and Luis Rodero-Merino. “Finding your way in the fog: Towards a comprehensive definition of fog computing”. In: *ACM SIGCOMM Computer Communication Review* 44. 5 (2014), pp. 27–32.
- 40 Zhi-Wei Xu. “Cloud-sea computing systems: Towards thousand-fold improvement in performance per watt for the coming zettabyte era”. In: *Journal of Computer Science and Technology* 29. 2 (2014), pp. 177–181.
- 41 Ashkan Yousefpour et al. “All one needs to know about fog computing and related edge computing paradigms: A complete survey”. In: *Journal of Systems Architecture* 98 (2019), pp. 289–330.

