IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received April 12, 2020, accepted April 28, 2020, date of publication May 6, 2020, date of current version May 19, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.2991734*

# An Overview on Edge Computing Research

**KEYAN CAO[1,2], YEFAN LIU[1], GONGJIE MENG[1], AND QIMENG SUN[1]**
[1]College of Information and Control Engineering, Shenyang Jianzhu University, Shenyang 110168, China
[2]Liaoning Province Big data Management and Analysis Laboratory of Urban Construction, Shenyang 110168, China

Corresponding author: Keyan Cao (caokeyan@sjzu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61602323, in part by the National Postdoctoral Foundation of China under Grant 2016M591455, in part by the Youth Seedling Foundation of Liaoning Province under Grant lnqn201913, and in part by the Natural Science Funds of Liaoning Province under Grant 2019MS264 and Grant 20180550019.

**ABSTRACT** With the rapid development of the Internet of Everything (IoE), the number of smart devices connected to the Internet is increasing, resulting in large-scale data, which has caused problems such as bandwidth load, slow response speed, poor security, and poor privacy in traditional cloud computing models. Traditional cloud computing is no longer sufficient to support the diverse needs of today's intelligent society for data processing, so edge computing technologies have emerged. It is a new computing paradigm for performing calculations at the edge of the network. Unlike cloud computing, it emphasizes closer to the user and closer to the source of the data. At the edge of the network, it is lightweight for local, small-scale data storage and processing. This article mainly reviews the related research and results of edge computing. First, it summarizes the concept of edge computing and compares it with cloud computing. Then summarize the architecture of edge computing, keyword technology, security and privacy protection, and finally summarize the applications of edge computing.

**INDEX TERMS** Edge computing, cloud computing, Internet of Things.

## I. INTRODUCTION

With the development of intelligent society and the continuous improvement of people's needs, intelligence has involved various industries and people's daily lives in society. Edge devices have spread to all aspects of society, such as smart homes and autonomous vehicles in the field of transportation, camera, intelligent production robot in intelligent manufacturing, etc. As a result, the number of devices connected to the Internet has increased significantly. Cisco pointed out in the Global Cloud Index [1] that in 2016, there were 17.1 billion devices connected to the Internet,by 2019, the total number of data traffic in global data centers will reach 10.4 Zettabyte (ZB), 45% of the data will be stored, processed and analyzed on the edge of the network, and by 2020, the number of wireless devices connected to the network will exceed 50 billion. The amount of data generated by devices worldwide has also increased from 218ZB in 2016 to 847 ZB in 2021. International data company Internet Data Center (IDC) statistics show that by 2020, the number of terminals and devices connected to the network will exceed 50 billion, and the total global data in 2020 will also exceed 40 ZB [2]. Based on the continuous and massive growth of data volume and various data processing requirements, cloud-based big data processing has shown many shortcomings:

Real-time: If a large number of edge devices are added, a large amount of terminal data is still transmitted to the cloud for processing, the intermediate data transmission volume will be greatly increased, the data transmission performance will be reduced, resulting in a large load of network transmission bandwidth, resulting in data transmission delay. In some application scenarios that require real-time feedback, such as traffic, monitoring, etc., cloud computing will not be able to meet business real-time requirements.

Security and privacy: For example, when using various applications in smartphones, applications will require user data, including privacy data. There is a high risk of privacy leakage or attack on this data when uploaded to the cloud center.

Energy consumption: the number of smart devices continues to increase, and the power consumption of data centers in China has increased significantly. Improving the use efficiency of cloud computing energy consumption [3] cannot meet the increasing demand for data energy consumption. The rapidly developing intelligent society will have higher requirements for the energy consumption of cloud computing.

The associate editor coordinating the review of this manuscript and approving it for publication was Sabah Mohammed.

Due to the increasing amount of data and increasing requirements for data processing, edge computing has emerged at the historic moment. Edge computing technology provides artificial intelligence services for rapidly growing terminal devices and data, and makes services more stable. Edge computing is close to the source of the data, such as smart terminals. It stores and processes data at the edge of the network. It has proximity and location awareness, and provides users with near-end services. In terms of data processing, it is faster, real-time, and secure. It can also solve the problem of excessive energy consumption in cloud computing, reduce costs, and reduce the pressure of network bandwidth. Edge computing is applied in various fields such as production, energy, smart home, and transportation.

With the development of the Internet of Things (IoT), edge computing models are urgently needed and have become hot research issues. In this article, we introduce edge computing in detail from the aspects of edge computing introduction, architecture, key technologies, security and privacy, and applications to provide a reference for edge computing researchers.

## II. EDGE COMPUTING

### A. EDGE COMPUTING CONCEPTS

Edge computing is different from traditional cloud computing. It is a new computing paradigm that performs computing at the edge of the network. Its core idea is to make computing closer to the source of the data [4]. Researchers have different definitions of edge computing. Shi *et al.* [5]–[7] introduced the emergence of the concept of edge computing: "Edge computing is a new computing mode of network edge execution. The downlink data of edge computing represents cloud service, the uplink data represents the Internet of Everything, and the edge of edge computing refers to the arbitrary computing and network resources between the data source and the path of cloud computing center." Satyanarayanan, a professor at Carnegie Mellon university in the United States, describes edge computing as: "Edge computing is a new computing model that deploys computing and storage resources (such as cloudlets, micro data centers, or fog nodes, etc.) at the edge of the network closer to mobile devices or sensors" [4]. Zha *et al.* [8] proposed on the basis of the above two definitions: "Edge computing is a new computing model that unifies resources that are close to the user in geographical distance or network distance to provide computing, storage, and network for applications service." China's edge computing industry alliance defines edge computing as: "near the edge of the network or the source of the data, an open platform that integrates core capabilities such as networking, computing, storage, applications, and provides edge intelligent services nearby to meet the industry agility key requirements in connection, real-time business, data optimization, application intelligence, security and privacy" [9].

In other words, edge computing is to provide services and perform calculations at the edge of the network and data generation. Edge computing is to migrate the cloud's network, computing, storage capabilities and resources to the edge of the network, and provide intelligent services at the edge to meet the critical needs of the IT industry in agile linking, real-time business, data optimization, application intelligence, security and privacy, and meets the requirements of low latency and high bandwidth on the network. Edge computing has become a research hotspot nowadays [10]–[14].

### B. CLOUD COMPUTING AND EDGE COMPUTING

#### 1) CLOUD COMPUTING

Before the emergence of edge computing, traditional cloud computing transfers all data to the cloud computing center through the network, and solves the computing and storage problems in a centralized way. In literature [15], the development history of cloud computing is described. In the search engine conference (sessane jose 2006) in August 2006, the CEO of Google first proposed the concept of cloud computing. In the development history of cloud computing, this is the first time to formally put forward the concept of cloud computing. With the development of search engines represented by Google, cloud computing starts to show strong vitality. Nowadays, cloud computing has gradually developed. It is a very powerful network service platform including distributed computing, load balancing, parallel computing, network storage, virtualization and other technologies. However, nowadays, with the popularization and development of the Internet of Things in people's life, the number of devices connected to the Internet of Things is gradually increasing, and a large amount of data is generated. The network bandwidth of cloud computing has been unable to meet the needs of time-sensitive systems [16] and real-time performance. Therefore, cloud computing model has great defects in load, real-time [17]–[19], transmission bandwidth, energy consumption and data security and privacy protection [20].

#### 2) CONNECTION AND DIFFERENCE BETWEEN CLOUD COMPUTING AND EDGE COMPUTING

The emergence of edge computing will not replace cloud computing. In the aspects of network, business, application and intelligence, the two should exist together, complement each other and develop in a coordinated way, which will help the digital transformation of the industry to a greater extent. All data onto edge nodes still need to be summarized in the cloud to achieve in-depth analysis and obtain more meaningful analysis results. Therefore, cloud computing is still playing an important role in the development of Internet of Things devices that are gradually intelligent.

In the context of the Internet of Things, if all the large amount of data generated by the connected devices are transmitted to the cloud, cloud computing will cause a large load. At this time, edge computing is required to share the pressure of the cloud and take charge of tasks within its scope of the edge. When there is a problem in edge computing, the data in the cloud is not lost. In some Internet services, some data needs to be returned to the cloud for processing after being processed by edge computing, such as in-depth analysis of data mining and sharing, which requires the cooperation of cloud computing and edge computing. Both developments bring stability to connected devices in the Internet of Things network. The working method of the two can be that cloud computing is based on big data analysis and output, passed to the edge side, and then processed and executed by edge computing. Nowadays, the coordinated development of the two has been applied in many aspects of real life, such as intelligent manufacturing [21], energy [22], security [23] and privacy protection [24], and intelligent family. For example, in the industrial production of intelligent manufacturing, the role of the cloud is to control the whole. In the edge nodes, it is necessary to have the function of real-time detection and solve the problems in time. Edge computing takes advantage of the real-time characteristics, and in collaboration with cloud computing and synergy, not only improves production efficiency, but also can detect abnormalities of equipment in a timely manner. In the field of smart home, edge computing nodes mainly involve some intelligent terminals. Edge computing nodes calculate heterogeneous data from different devices and upload it to the cloud for processing, so as to realize the control of edge nodes from the cloud and the access of edge nodes to the cloud. In order to meet the needs of Internet of Things devices, cloud computing and edge computing play their respective advantages, and only the joint development of the two can continuously promote the progress of the Internet.

Edge computing is an extension of cloud computing, which has its own characteristics with cloud computing. The main feature of cloud computing are that it can grasp the whole, can process a large amount of data, conduct in-depth analysis, and also plays an important role in non-real-time data processing, such as business decision-making and other fields. Edge computing focuses on the local, and can play a better role in small-scale, real-time intelligent analysis, such as meeting the real-time needs of local businesses. Therefore, in intelligent applications, cloud computing is more suitable for centralized processing of large-scale data, while edge computing can be used for small-scale intelligent analysis and local services. In terms of network resources, edge computing is responsible for data closer to the information source. Therefore, data can be stored and processed locally without uploading all the data to the cloud. The reduction of network burden greatly improves the utilization efficiency of network bandwidth. Cloud computing and edge computing play an important role in the future development of intelligent Internet of Things [25]. The main differences between cloud computing and edge computing are shown in Table 1.

### C. ADVANTAGES OF EDGE COMPUTING

Edge computing model stores and processes data on edge devices without uploading to cloud computing platform. Due to this feature, edge computing has obvious advantages in the following aspects:

**TABLE 1.** Main differences between cloud computing and edge computing.

| | Applicable situation | Network bandwidth pressure | Real-time | Calculation mode |
|---|---|---|---|---|
| Cloud computing | Global | More | High | Large scale centralized processing |
| Edge computing | Local | Less | Low | Small scale intelligent analysis |

Fast data processing and analysis, real-time: the rapid growth of data volume and the pressure of network bandwidth are disadvantages of cloud computing [26]. Compared with traditional cloud computing, edge computing has advantages in response speed and real-time. Edge computing is closer to the data source, data storage and computing tasks can be carried out in the edge computing node, which reduces the intermediate data transmission process. It emphasizes proximity to users and provides users with better intelligent services, thus improving data transmission performance, ensuring real-time processing and reducing delay time. Edge computing provides users with a variety of fast response services, especially in the field of automatic driving intelligent manufacturing, video monitoring and other location awareness, rapid feedback is especially important.

Security: traditional cloud computing requires all data to be uploaded to the cloud for unified processing, which is a centralized processing method. In this process, there will be risks such as data loss and data leakage, which cannot guarantee security and privacy. For example, account passwords, historical search records and even trade secrets can all be exposed. Since edge computing is only responsible for the tasks within its own scope, the processing of data is based on the local, there is no need to upload to the cloud, to avoid the risks brought by the network transmission process, so the security of data can be guaranteed. When data is attacked, it only affects local data, not all data.

Low cost, low energy consumption, low bandwidth cost: in edge computing, since the data to be processed does not need to be uploaded to the cloud computing center, it does not need to use too much network bandwidth, so the load of network bandwidth is reduced, and the energy consumption of intelligent devices at the edge of the network is greatly reduced. Edge computing is "small-scale," and in production, companies can reduce the cost of processing data in local equipment. Therefore, edge computing reduces the amount of data transmitted on the network, reduces the transmission cost and network bandwidth pressure, reduces the energy consumption of local equipment, and improves the computing efficiency.

## III. ARCHITECTURE OF EDGE COMPUTING

With the Internet of Everything era and the development of 5G, edge computing is considered as one of the key technologies in the next generation of communication network following the Internet of Things and artificial intelligence [25]. The reference architecture for edge computing is the focus of many organizations. This section begins with an overview of the general architecture for edge computing, followed by a detailed introduction of the reference architecture proposed by the edge computing industry alliance (ECC) and the Linux foundation in sections 3.2 and 3.3, respectively.

## A. GENERAL ARCHITECTURE OF EDGE COMPUTING

Edge computing architecture is a federated network structure that extends cloud services to the edge of the network by introducing edge devices between terminal devices and cloud computing [27], [28].

The structure of cloud-edge collaboration is generally divided into terminal layer, edge layer and cloud computing layer. The following is a brief introduction to the composition and functions of each layer in the edge computing architecture.

### 1) TERMINAL LAYER

The terminal layer consists of all types of devices connected to the edge network, including mobile terminals and many Internet of Things devices (such as sensors, smartphones, smart cars, cameras, etc.). In the terminal layer, the device is not only a data consumer, but also a data provider. In order to reduce the terminal service delay, only the perception of the various terminal devices is considered, not the computing power. As a result, hundreds of millions of devices in the terminal layer collect all kinds of raw data and upload it to the upper layer, where it is stored and calculated.

### 2) BOUNDARY LAYER

The edge layer is the core of the three-tier architecture. It is located at the edge of the network and consists of edge nodes widely distributed between terminal devices and clouds. It usually includes base stations, access points, routers, switches, gateways, etc. The edge layer supports the access of terminal devices downward, and stores and computes the data uploaded by terminal devices. Connect with the cloud and upload the processed data to the cloud [7]. Since the edge layer is close to the user, the data transmission to the edge layer is more suitable for real-time data analysis and intelligent processing, which is more efficient and secure than cloud computing.

### 3) CLOUD LAYER

Among the federated services of cloud-edge computing, cloud computing is still the most powerful data processing center. The cloud computing layer consists of a number of high-performance servers and storage devices, with powerful computing and storage capabilities, and can play a good role in areas requiring large amounts of data analysis such as regular maintenance and business decision support. The cloud computing center can permanently store the reported data of the edge computing layer, and it can also complete the analysis tasks that the edge computing layer cannot handle and the processing tasks that integrate the global information. In addition, the cloud module can also dynamically adjust the deployment strategy and algorithm of the edge computing layer according to the control policy.

## B. EDGE COMPUTING REFERENCE FRAME 3.0

ECC, jointly initiated by Huawei, Shenyang institute of automation, Chinese academy of sciences, China academy of information and communications, and other well-known enterprises, put forward the edge computing reference frame 3.0 in the edge computing white paper 3.0 released in December 2018 [29]. The frame of reference is based on model-driven engineering method. In order to model the knowledge of the physical and digital world, we need to achieve the following four goals:

1) Establish a real-time and systematic cognitive model of the physical world and achieve the cooperation between the physical world and the digital world;

2) Establish reusable knowledge model system in each vertical industry based on modeling method, and complete cross-industry ecological cooperation;

3) System to system, service to service and other model-based interface for interaction, to achieve decoupling of software interface and development language, reduce system heterogeneity;

4) Can effectively support the life cycle of development service, deployment operation, data processing and security.

As shown in Figure 1, the ECC edge computing reference architecture presents the architecture content from different perspectives in a multi-view manner, and the functionality of each layer is shown through the multi-layer functional perspective.

The edge reference framework has an underlying service layer that links the entire framework, including management services, data lifecycle services, and security services. Management services provide unified management, monitoring the operation of the architecture, and providing information to the management platform. The data lifecycle service provides integrated management for the preprocessing, analysis, distribution, and execution of machine data, as well as visualization and storage. The security service can define the business logic of the whole life cycle of data through the business orchestration layer, flexibly deploy and optimize the data service, and meet the real-time requirements of the business. Security services cover all levels of edge computing architecture, adapt to the specific architecture of edge computing, and make use of the unified security management and perception system to ensure the safe and reliable operation of the entire architecture.

From the perspective of vertical structure, the model-driven unified service framework is located at the top of the scale to realize the development and deployment of services. According to the general framework of edge calculation, it is divided into cloud, edge layer and field layer. The edge layer consists of two main parts: the edge node and the edge manager.

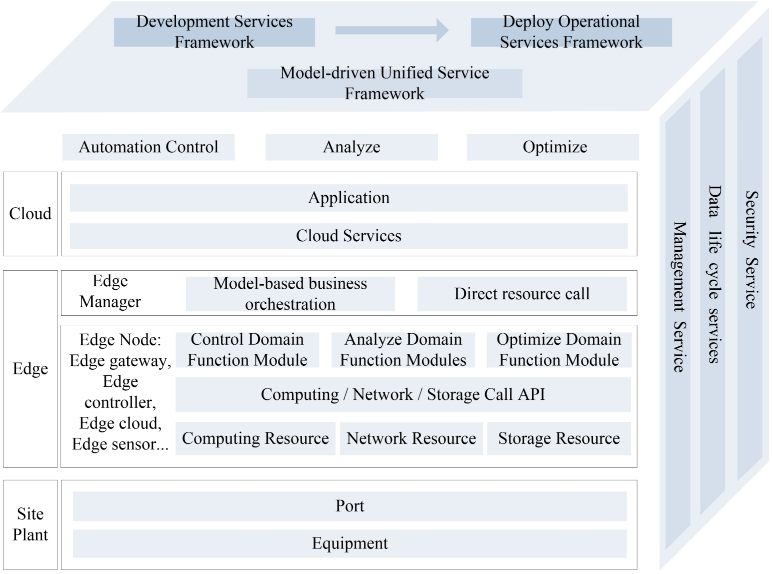
Development Services Framework
Deploy Operational Services Framework
Model-driven Unified Service Framework
Automation Control
Analyze
Optimize
Cloud
Application
Cloud Services
Edge
Edge Manager
Model-based business orchestration
Direct resource call
Edge Node: Edge gateway, Edge controller, Edge cloud, Edge sensor...
Control Domain Function Module
Analyze Domain Function Modules
Optimize Domain Function Module
Computing / Network / Storage Call API
Computing Resource
Network Resource
Storage Resource
Site Plant
Port
Equipment
Management Service
Data life cycle services
Security Service

**FIGURE 1.** Edge computing reference architecture 3.0.

Edge node is the entity hardware, which can carry the business of edge computing. The edge manager mainly uses software to uniformly manage edge nodes. According to hardware characteristics and service types, edge computing nodes are divided into edge gateways that process and convert network protocols, edge controllers that control real-time closed-loop services, edge clouds that process large-scale data, and edges that collect and process low-cost information Sensors, etc., can abstract the devices in the edge computing layer into computing, networking and storage. Next, implement generic capability calls using Application Programming Interfaces (API). The control, analysis and optimization domain function module is used to realize the information transmission of upper and lower layers and the planning of local edge resources. Edge computing reference architecture 3.0 provides four service development frameworks from the terminal to the cloud, including lightweight computing systems, real-time computing systems, intelligent distributed systems, and intelligent gateway systems.

### C. EdgeX FOUNDRY

EdgeX Foundry is a neutral open source project hosted by the Linux foundation and is a universal open framework for computing on the edge of the Internet of Things. Hosted on a reference software platform that is completely independent of hardware and operating systems, the framework enables a plug-and-play component ecosystem to unify the computing open platform at the edge of the Internet of Things and accelerate deployment of solutions.

Figure 2 shows the EdgeX Foundry architecture. As can be seen from the figure, the "Southbound" contains all IoT application devices that can directly communicate with the edge network. The "Northbound" contains the cloud computing center and the communication network with the cloud computing center. The "Southbound" is the source of the data, while the "Northbound" is used to collect the data from the south side and store, integrate and analyze the data. The EdgeX Foundry is located between the south and Northbounds and consists of a collection of microservices that are divided into four service layers and two underlying enhanced system services.

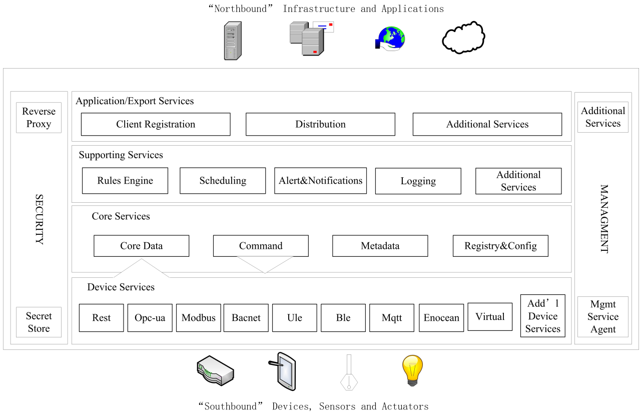

**FIGURE 2.** EdgeX foundry architecture.

From a vertical perspective, the software development kit (SDK) provided by the device service layer is used to establish the communication links with the "Southbound". The device service layer converts the data from the device and sends it to the core service layer. It can also receive commands from other microservices and pass them to devices. EdgeX Foundry includes multiple access methods such as Message Queue Telemetry Transmission Protocol (MQTT), Virtual Device (VIRTUAL), and Bluetooth Low Energy (BLE). As the core service layer of the center, it is the key to realize edge capabilities. The core service layer consists of four micro-service components: core data, commands, metadata, registry, and configuration. The core data service provides storage and management services for device data. The command service is responsible for defining the operation commands in the device file into a general API, caching and managing commands, and can pass the requirements of the cloud computing center to the device side. Metadata provides pairing for devices and services. Registration and configuration services provide configuration information for other micro-services. The support service layer is used to provide edge analysis and intelligent services, and to provide the framework itself with rules engine, scheduling, alerting and logging services. The application and export service layer can be connected with the cloud computing center to transfer data to the cloud computing center to ensure the independent operation of EdgeX Foundry. In the export service layer, the client registration service records the relevant information of the back-end registered system, and the distribution service exports the corresponding data to the specified client.

Similar to the Edge Computing Reference Architecture 3.0, the EdgeX Foundry framework also has a basic service layer that runs through the whole framework-management services and security services. The management service provides functions such as installing, upgrading, starting, stopping, and monitoring EdgeX Foundry operations. The components in the security service are used to protect the data from the device and the operation of the device.

The EdgeX Foundry framework was developed to simplify and standardize edge computing for the Industrial Internet of Things. It provides an operable open-source platform, in which all micro-services can run on various operating systems in the form of containers, and support dynamic addition or reduction of functions, with strong scalability. At present, the application fields of EdgeX Foundry have involved various industries such as retail, manufacturing [30], energy [31], urban parks, transportation [32] and other industries.

## IV. HOT RESEARCH CONTENT OF EDGE COMPUTING

The emergence of edge computing has promoted the rapid development of the Internet of Things and has made a significant contribution to the realization of an intelligent society. Therefore, edge computing has become a hot issue for scientists at home and abroad. This section mainly reviews the key technologies and data security and privacy protection.

### A. KEY TECHNOLOGIES

The key technologies of edge computing mainly include different levels of computing offloading, mobility management, traffic offloading technology, caching acceleration, network control and so on.

#### 1) COMPUTING OFFLOADING

Computing offloading refers to resource-constrained device that partially or fully migrates resource-intensive computing from mobile devices to resource-rich nearby infrastructure to address mobile device deficiencies in resource storage, computing performance, and energy efficiency [33]. The computing offloading technology not only reduces the pressure on the core network, but also reduces the delay caused by transmission. Mobile edge computing (MEC) can run new complex applications on user equipment (UE), and computing offloading is the key technology for edge computing. There have been many related research achievements, mainly including two main issues: offloading decision and resource allocation. Among them, the offloading decision is about how to offload computing tasks, how much to offload and what to offload for mobile devices. Resource allocation is to study where to offload resources.

##### a: OFFLOADING DECISION

Offloading decision refers to the problem that the UE decides how to offload the computing task, how much to offload, and what to offload. In the offloading system, the UE generally consists of a code parser, a system parser, and a decision engine. Its execution of the offloading decision is divided into three steps: (1) The code parser determines what can be offloaded, and the specific offloading content depends on the application type and code data partition; (2) The system parser is responsible for monitoring various parameters, such as the available bandwidth, the size of the data to be offloaded, or the energy consumed to execute a local application; (3) Finally, the decision engine determines whether to offload or not.

The UE offloading decision results are divided into three cases: local execution, full offloading, and partial offloading. The specific decision results are determined by the UE energy consumption and delay in completing computing tasks. According to the optimization goals of the offloading decision, the computing offloading can be divided into three types: reducing latency as the goal, reducing energy consumption as the goal, and balancing energy consumption and latency as the goal.

##### b: RESOURCE ALLOCATION

After completing the offloading decision, we must consider the issue of reasonable resource allocation, that is, where to offload. If the computing task of the UE is indivisible or can be divided but the divided parts are related, in this case, the offloading task needs to be offloaded to the same MEC server; and for the computing tasks that can be divided but not related to the divided part, it can be offloaded to multiple MEC servers. At present, resource allocation nodes are mainly divided into single-node allocation and multi-node allocation.

#### 2) MOBILITY MANAGEMENT

Edge computing relies on the geographical distribution of resources to support the mobility of applications. An edge computing node only serves users around it. The cloud computing mode supports the application mobility by fixing the location of the server and transmitting the data to the server through the network, so the mobile management of the application in the edge computing is a new mode. The main issues involved are resource discovery and resource switching.

Resource discovery, that is, users need to quickly discover the resources available around and choose the most suitable resources during the movement. The resource discovery of edge computing needs to adapt to the heterogeneous resource environment, and also needs to ensure the speed of resource discovery, so that applications can provide services to users without interruption; resource switching, that is, when users move, the computing resources used by mobile applications may be switched among multiple devices [34]. Resource switching will migrate the service program's operation site to ensure service continuity.

In MEC, one of the key issues to be considered is how to ensure the continuity of users' access to services during the movement. Some applications expect to continue serving users after their location changes. The heterogeneity of edge computing resources and the diversity of networks require adaptive device computing capabilities and changes in network bandwidth during the migration process. Reference [35] further optimized the virtual machine migration strategy by predicting the user's movement, and proposed a mobility-based service migration prediction scheme (MSMP), which adopted a compromise between cost and service quality.

#### 3) OTHER KEY TECHNOLOGIES

In addition to the above two key technologies, the key technologies of edge computing also include traffic offloading technology, cache acceleration, and network control.

##### a: TRAFFIC OFFLOADING TECHNOLOGY

In order to realize the localization, short-distance deployment, and low-latency, high-bandwidth transmission capabilities of business applications in wireless networks, wireless networks have the capability of traffic offloading. Offloading (traffic offloading) of edge network traffic is very important in mobile edge computing. Traffic offloading is used to offload traffic that meets specific offloading rules to mobile edge networks (that is, a local specific network, which can be an intranet or the Internet) to save backhaul bandwidth, reduce latency, and facilitate the expansion of other MEC services [36]. Reference [37] proposed a method of energy-efficient traffic offloading for mobile users in a two-layer heterogeneous wireless network. Experimental results show that the method can save up to 34 percents of energy under typical network settings.

##### b: CACHING ACCELERATION

Mobile edge caching technologies include base station caching, mobile content distribution networks, and transparent caching. Caching acceleration technology can improve the efficiency of content distribution and improve the user experience. After the content is cached to the edge of the mobile network, users can obtain the content nearby, thereby avoiding repeated transmission of content, and alleviating the pressure on the backhaul network and core network. At the same time, the edge caching can reduce the network delay requested by the user, thereby improving the user's network experience. In addition, the edge caching can also open up the mobile network resource environment and provide more abundant services for tenants and users [38]. Reference [39] proposed a cognitive agent (CA) to help users cache and perform tasks on the MEC in advance, and to coordinate communication and cache to ease the pressure on the MEC.

##### c: NETWORK CONTROL

An edge network is a given non-technically described network (the edge of a public telecommunication network). The edge network includes part or all of the aggregation layer network and the access layer network, and is the last segment of the network for accessing users. In terms of value, edge networks are commercial networks between existing core networks and large users. In terms of network control, reference [40] proposed an effective workload slicing scheme, in which users use software-defined networks to process data-intensive applications in multi-edge cloud environments.

### B. DATA SECURITY AND PRIVACY PROTECTION

The security of edge computing is one of the hot research issues. Network edge data involves personal privacy. Although the concept of data processing nearby also provides better structured support for data security and privacy protection, the distributed architecture of edge computing increases the dimension of attack vectors. The smarter the edge computing client, the more vulnerable they are to malware infections and security breaches. Existing data security protection methods are not fully applicable to edge computing architectures. Moreover, the highly dynamic environment at the edge of the network also makes the network more vulnerable and difficult to protect. Data security and privacy protection in edge computing faces four new challenges:

(1) New requirements for lightweight data encryption and fine-grained data sharing based on multiple authorized parties in edge computing. Because edge computing is a computing mode that integrates multiple trust domains with authorized entities as trust centers, traditional data encryption and sharing strategies are no longer applicable. Therefore, it is particularly important to design a data encryption method for multiple authorization centers. At the same time, the complexity of the algorithm should be considered.

(2) Multi-source heterogeneous data propagation control and security management issues in a distributed computing environment. Users or data owners want to be able to use effective information dissemination control and access control mechanisms to achieve data distribution, search, access, and control of the scope of data authorization. In addition, due to the outsourcing nature of data, its ownership and control are separated from each other, so an effective audit verification scheme can ensure the integrity of the data.

(3) Security challenges between large-scale interconnected services for edge computing and resource-constrained terminals. Due to the multi-source data fusion characteristics of edge computing, the superposition of mobile and Internet networks, and the resource limitations of storage, computing, and battery capacity of edge terminals, traditional and more complex encryption algorithms, access control measures, identity authentication protocols and privacy protection methods cannot be applied in edge computing.

(4) Diversified services for the Internet of Things and the new requirements of edge computing mode for efficient privacy protection. In addition to the need to design effective data, location and identity privacy protection schemes, how to combine traditional privacy protection schemes with edge data processing characteristics in edge computing environments to enable user privacy protection in a diverse service environment is the future research trend.

At present, the research on edge computing security and privacy protection is still in its infancy, and there are relatively few existing research results. Among them, a really feasible research idea is to port existing security technologies in other related fields to the edge computing environment. Scholars at home and abroad have carried out in-depth research on mobile cloud computing and its security. Roman *et al.* [41] conducted security analysis on several common mobile edge paradigms, elaborated a general cooperative security protection system, and gave research opinions. These works provide theoretical reference for the security research of edge computing. This paper divides the research system of data security and privacy protection in edge computing into four parts: data security, identity authentication, privacy protection and access control.

#### 1) DATA SECURITY

Data security is the foundation of creating a secure edge computing environment, whose fundamental purpose is to ensure the confidentiality and integrity of data. It is mainly aimed at the characteristics of separation of ownership and control of outsourced data and randomization of storage, and is used to solve problems such as data loss, data leakage, and illegal data operations. At the same time, on this basis, users are allowed to perform secure data operations. So far, most of the research results of scholars at home and abroad have focused on cloud computing [42], mobile cloud computing [43] and fog computing [44]. Therefore, a main research idea of data security in edge computing is to migrate data security solutions in other computing paradigms to the edge computing paradigm, and to parallelize the distributed computing architecture in edge computing, limited terminal resources, edge big data processing, highly dynamic environment and other characteristics are organically combined to finally achieve a lightweight and distributed data security protection system.

##### *a: DATA CONFIDENTIALITY AND SECURE DATA SHARING*

Existing data confidentiality and secure data sharing solutions are usually implemented by encryption technology. The conventional process is that the data owner encrypts and uploads the outsourced data in advance, and the data user decrypts the data when necessary. Traditional encryption algorithms include symmetric encryption algorithms (such as DES, 3DES, ADES, etc.) and asymmetric encryption algorithms (such as RSA, Diffe-Hellman, ECC, etc.), but the operability of data encrypted by traditional encryption algorithms is low, which causes great obstacles to the subsequent data processing. At present, the commonly used data encryption algorithms include attribute based encryption (ABE) [45], proxy re encryption (PRE) [46], and all homomorphic encryption (FHE) [47], etc.

##### *b: INTEGRITY AUDIT*

After the user's data is stored in the edge or cloud data center, an important issue is how to determine the integrity and availability of outsourced storage data. The current research on data integrity audits has focused on the following four functional requirements [48]:

1) Dynamic audit. User data in the data storage server is often dynamically updated. The common dynamic data operations include modification, copying, inserting, and deleting. Therefore, the data integrity audit scheme cannot be limited to static data, but should have dynamic audit capabilities.

2) Batch audit. When a large number of users issue audit requests at the same time or data is stored in multiple data centers in blocks, in order to improve audit efficiency, the integrity audit solution should have the ability to perform batch audits.

3) Privacy protection. Since neither the data storage server nor the data owner is suitable for performing integrity audit schemes, they often need to be built with a third-party auditing platform (TPA). In this case, when the TPA is semi-trusted or untrusted, security threats such as data leakage and tampering are very likely, and data privacy cannot be guaranteed. Therefore, protecting user data privacy during the integrity audit process is essential.

4) Low complexity. For data storage servers (edge data centers) and data owners (edge devices) have limitations in terms of computing power, storage capacity, network bandwidth, etc., in addition to ensuring data integrity when designing an integrity audit scheme, the scheme is complex The issue of degree is also an important factor.

##### *c: SEARCHABLE ENCRYPTION*

In the traditional cloud computing paradigm, in order to achieve data security and reduce terminal resource consumption, users often use some encryption method to outsource file encryption to a third-party cloud server. However, when users need to find related files that contain a certain keyword, they will encounter the difficulty of how to perform a search operation on the cipher text of the cloud server. Therefore, searchable encryption (SE) came into being. SE can guarantee the privacy and availability of data, and support the query and retrieval of ciphertext data. Similarly, in the edge computing paradigm, user's file data will be encrypted and outsourced to the edge computing center or cloud server. Searchable encryption is also an important method to protect user privacy in edge computing.

#### d: SUMMARY

1) In terms of data confidentiality and secure data sharing, combining with application encryption theory such as attribute encryption, proxy re-encryption, and homomorphic encryption, how to design a low-latency, distributed secure storage system that supports dynamic operations and correctly handle network edge devices The synergy between cloud centers is an important research idea.

2) In the field of data integrity auditing, one of the main research goals is to improve auditing efficiency and reduce verification overhead while realizing various auditing functions. Secondly, designing an integrity audit scheme that supports multi-source heterogeneous data and dynamic data updates is expected to become the focus of future research.

3) In terms of searchable encryption, first, how to construct a keyword-based search scheme under the distributed storage service model and further expand it into the edge computing environment is a feasible research idea; second, how to implement it in a secure multi-party sharing mode Fine-grained search permission control makes it suitable for multi-user search environments with different trust domains, while ensuring search speed and accuracy. Finally, for the distributed ciphertext data storage model in edge computing, how to efficiently construct a secure index suitable for resource-constrained network edge devices and design a distributed searchable encryption algorithm is an urgent problem.

### 2) IDENTITY AUTHENTICATION

To use the computing services provided by edge computing, IoT users must first perform identity authentication. Because edge computing is a distributed interactive computing environment where multiple trust domains coexist, it is necessary not only to assign an identity to each entity, but also to consider mutual authentication between different trust domains. The main research content of identity authentication includes identity authentication within a single domain, cross-domain authentication, and handover authentication.

#### a: IDENTITY AUTHENTICATION IN A SINGLE DOMAIN

The identity authentication in a single trust domain is mainly used to solve the identity allocation problem of each entity. Each entity must first pass the security authentication of the authorization center to obtain storage and computing services. With the deepening of research, designing identity authentication protocols with privacy protection features is the focus of current research.

#### b: CROSS-DOMAIN AUTHENTICATION

At present, the research on the authentication mechanism applicable to entities in different trust domains is still in its infancy, and a relatively complete research context and theoretical methods have not yet been formed. In the research of identity authentication in cloud computing, identity management between multiple cloud service providers can be regarded as a form of cross-domain authentication, which makes some authentication standards applicable to multi-cloud (such as SAML, OpenID, etc.) And the single sign-on (SSO) authentication mechanism is expected to be applied to identity authentication between multiple trust domains [49]. Reference [50] designed an attribute-based authentication and authorization framework for structured P2P networks. The framework uses attribute certificates and distributed certificate revocation systems to replace the traditional P2P network's public key certificates and access control list authentication mechanisms. A server or third-party trusted authority that achieves flexible, efficient, and privacy-protected rights assignments without requiring any external intervention. Literature [51] proposed a cross-domain dynamic anonymous group key management authentication system (CD-AGKMS). This system achieves cross-domain group density by establishing a tree-level hierarchy with the key generation center (KGC) as the top layer. Key negotiation. At the same time, in terms of group key management, the scheme provides a time-controlled key revocation mechanism, and the user's key is revoked when the validity period expires. In addition, CD-AGKMS does not require the calculation of bilinear pairs, which improves the feasibility and efficiency of the system.

#### c: SWITCHING AUTHENTICATION

Due to the high mobility of terminal equipment in edge computing, the geographic location of mobile users often changes, making the traditional centralized authentication protocol no longer applicable to such situations. Handover authentication is a kind of authentication handover technology to solve the problem of high mobility user identity authentication. Therefore, the research on handover authentication technology can provide strong guarantee for real-time and accurate authentication of edge devices in edge computing. At the same time, the privacy of user identity in the process of authentication handover is also a research focus.

#### d: SUMMARY

At present, domestic and foreign researchers have mostly improved and optimized the authentication protocols based on the existing security protocols, including the flexibility, high efficiency, energy saving and privacy protection of the protocols. In edge computing, the research of identity authentication protocol should draw on the advantages of existing schemes, and at the same time combine the characteristics of distribution and mobility in edge computing to strengthen the research of unified authentication, cross-domain authentication and handover authentication technologies to ensure the data and privacy security of users in different trust domains and heterogeneous network environments.

### 3) PRIVACY PROTECTION

Not all authorized entities in edge computing are trusted, but the user's identity information, location information and private data are stored in these semi-trusted entities, which easily leads to privacy problems. At present, the research on privacy protection is mainly concentrated in the environment of mobile cloud and fog computing. Therefore, privacy protection is a research system that has attracted much attention in edge computing based on open interconnection. Its main contents include data privacy protection, location privacy protection and identity privacy protection.

#### *a: DATA PRIVACY PROTECTION*

Because of the user's privacy data will be stored and processed by entities that are not under the user's control. Therefore, it is the current research focus to allow users to perform various operations (such as auditing, searching and updating) on data while ensuring the privacy of users is not leaked.

#### *b: LOCATION PRIVACY PROTECTION*

With the popularity of location based service, the issue of location privacy has also become a research focus. At present, the research focus in this field mainly focuses on the use of K-anonymity technology to achieve privacy protection in location services. However, the location privacy protection scheme based on K-Anonymity consumes a large amount of network bandwidth and computational overhead in practical applications, and is not suitable for edge devices with limited resources.

#### *c: IDENTITY PRIVACY PROTECTION*

At present, the protection of user identity privacy in the edge computing paradigm has not attracted widespread attention, only some exploratory research results in the mobile cloud environment. Khalil *et al.* [52] pointed out that the current third-party identity management system (IDM) is vulnerable to three kinds of attacks: IDM server compromise, mobile device compromise and network traffic interception. Aiming at these attacks, this paper proposes a comprehensive third-party identity management system (CIDM), which manages mobile users' digital identities on behalf of service providers by introducing IDM servers. Firstly, the authorization certificate, IDM server and service provider are separated to resist illegal access IDM and traffic interception attacks. At the same time, an additional authentication layer is added to prevent mobile devices from compromising.

In addition, in order to improve the performance and security of the scheme, the credential information of mobile devices will be updated in real time according to the mobile cloud packet switching mechanism to prevent credential theft attacks.

#### *d: SUMMARY*

The privacy problems of users in edge computing can be summarized into the following three contradictions: 1) Contradiction between Outsourcing Data and Data Privacy; 2) The contradiction between location based service and location privacy; 3) The contradiction between data sharing and identity privacy protection. Scholars at home and abroad have carried out in-depth research to solve these three contradictions, but the proposed scheme still has many defects, and some possible research directions are as following.

1) Support users to perform various operations (such as auditing, searching and updating) on data while ensuring the privacy of users is not leaked.

In addition, the privacy issue in the cooperative interoperability among users deserves extensive attention.

2) In view of the shortcomings of TTP-based privacy protection scheme in computing energy consumption, it is particularly important to design a lightweight and efficient privacy protection scheme.

3) A large amount of real-time dynamic data will be generated by edge devices in the actual network, which provides the possibility of data correlation, integration analysis and privacy mining for attackers. Therefore, it is an important research content to construct dynamic and fine-grained data security and privacy protection schemes from the perspective of user's identity, behavior, interest and location.

### 4) ACCESS CONTROL

Access control is a key technology and method to ensure system security and protect user privacy. Currently, popular access control schemes include attribute-based and role-based access control. Among them, attribute-based access control can be well applied to distributed architecture and realize fine-grained data sharing.

#### *a: ATTRIBUTE-BASED ACCESS CONTROL*

Because edge computing is a data-oriented computing mode, access control of edge computing is usually implemented by cryptographic technology. Traditional cryptographic technology is not suitable for distributed parallel computing environment, while Attribute Encryption (ABE) can be well applied to distributed architecture to realize fine-grained data sharing and access control.

#### *b: ROLE-BASED ACCESS CONTROL*

Role-based access control provides flexible control and management through a dual privilege mapping mechanism, that is, the privilege mapping from users to roles and roles to data objects.

#### *c: SUMMARY*

To sum up, access control technology is the key technology and important method to ensure system security and protect user privacy. In principle, the access control system in edge computing should be applicable to multi-entity access control between different trust domains, and various factors such as geographic location and resource ownership should also be considered. Therefore, the design of a fine-grained, dynamic, lightweight and multi-domain access control mechanism is the focus of the next research, and an efficient access control method based on attributes and roles should be a more suitable technical means for edge computing environment.

## V. EDGE COMPUTING APPLICATION SCENARIOS

The birth of each emerging technology is followed by its corresponding application in different scenarios. The important criterion to test the feasibility of the new technology is whether it is efficient to solve the existing problems in the actual environment. The various challenges and opportunities that edge computing will face in the application process are presented. With the improvement of edge computation in theory, more and more applications based on edge computation are called reality. This section will deeply understand edge calculation through the application of edge calculation in five scenes: edge calculation video cache scene, 5G communication scene, edge calculation network video live broadcast scene, predictive maintenance scene and security monitoring scene.

### A. EDGE COMPUTING VIDEO CACHE

With the rapid development of the emerging Internet industry, the booming application programs have led to the continuous growth of Internet traffic. According to the forecast [53], the total global IP traffic will triple between 2016 and 2021, from an annual average of 1.2ZB(96EB/ month) in 2016 to 3.3ZB(278EB/ month) in 2021. Figure 3 shows the statistics and forecast of the annual monthly average of global internet traffic. The proportion of video in Internet traffic will increase from 73 percent in 2016 to 82 percent, and it can be predicted that this proportion will increase year by year.

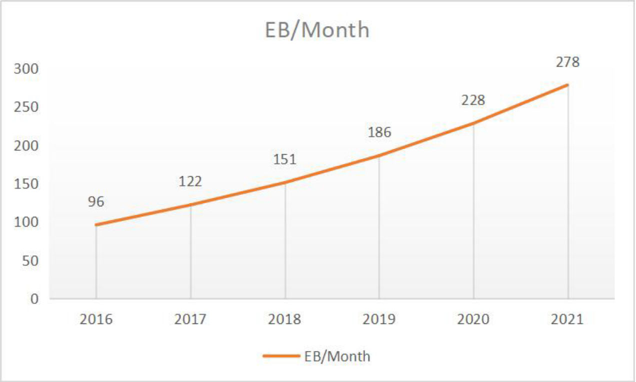

**FIGURE 3. Annual monthly average statistics/forecast of global internet traffic.**

The increasing video data traffic year by year will occupy more Internet bandwidth resources. In front of the limited bandwidth resources, it is undoubtedly an effective scheme to reasonably use the edge computing platform to cache local video. MEC, a platform based on edge computing with video analysis and caching functions, is deployed in areas with high traffic such as university towns, residential areas, commercial streets and others and frequent requests for video playback. The edge computing intelligent analysis function (based on search heat) [54] is used to cache popular TV plays, movies and other video resources with high download frequency on the nearby MEC server. When a user sends out a video playing request, the video resource can achieve the effect of loading from the local, thus not only saving bandwidth, but also greatly reducing the waiting time of the user. In addition, content in MEC platform is optimized based on RAN-side perception, so that content can be dynamically optimized according to network real-time information (network load, link quality, data throughput rate, etc.), and the effect of improving Quality of Experience (QoE) and network efficiency is achieved.

### B. EDGE CALCULATION AND 5G

The upcoming commercialization of the fifth generation mobile communication network (5G network) provides new opportunities for the development of edge computing. 5G has the advantages of small delay, large bandwidth and large capacity, which solves many problems encountered in the traditional communication field, but also leads to the rapid growth of data volume. At this time, it is urgent to provide a reliable, useful and executable business model. The characteristics of 5G, such as fast processing and low latency, can provide a new way for rapid response, and can jointly optimize the end, edge and cloud. This capability of edge computing can intelligently allocate resources among Internet of Things devices, edge devices and cloud devices from aspects of user experience, power consumption, computing load, performance, cost, etc., providing a new approach for joint optimization. Therefore, the development of edge computing technology is closely related to 5G: on the one hand, edge computing can support 5G, and the important component of 5G is edge computing. On the other hand, because 5G is expressed in the form of software, edge calculation can be flexibly applied [4].

In the European market, the edge computing industry has formed an industrial alliance, with large technology enterprises represented by Vodafone, Deutsche Telekom, Siemens and other companies already joining in. The European Telecommunications Standards Institute (ETSI) has initiated the formulation of standardized Mobile Edge Computing (MEC). Operators can open their wireless network edges to authorized third parties so that they can flexibly and quickly deploy innovative applications and services for mobile users, enterprises and vertical network segments [8], [9]. Mobile edge computing is the result of the natural development of mobile base station iteration and the integration of IT and telecommunication networks. It will provide new vertical service for consumers and enterprise customers, including video analysis, location service, Internet of Things, augmented reality, optimization of local content distribution and data caching, etc. In February 2018, ETSI released two white papers, namely ‘‘Cloud RAN and Mobile Edge Computing: Perfect Pairing’’ and ‘‘Deployment of Mobile Edge Computing in 4G and Evolution to 5G’’, in order to keep mobile edge computing synchronized with 5G.

In addition, people are also applying 5G to the Internet of Things, such as the Internet of Things network management system [55], the intelligent resource allocation management of the vehicle network [56] and the selection of reliable mobile vehicles for code distribution based on machine learning [57], etc.

### C. EDGE COMPUTING NETWORK VIDEO LIVE BROADCAST

The network video live broadcast system is a multimedia network platform, which aims to transmit live audio and video live events such as ongoing competitions, conferences, performances and teaching to remote audiences in real time through the network. Although the server port of the traditional live video broadcasting system generally adopts 100 megabytes or gigabytes of network, due to the large audio and video files, the delay problem in the whole process cannot be ignored.

To solve the above problems, Shanghai Mercedes-Benz Cultural Center has introduced multi-access edge computing technology. Edge nodes are deployed by adopting the technical scheme of China Unicom edge video scheduling network. Video shot in the venue is stored in a dedicated edge cloud, and audience in the venue can access the video information stored in the edge cloud through mobile devices, thus avoiding the time delay caused by connecting to the central cloud. The EVO solution can control the delay of live network video broadcasting to millisecond level, which is more than 60 times lower than the average delay of ordinary live network video broadcasting, and can support hundreds of millions of off-site Internet viewers to watch high-definition live broadcast.

### D. PREDICTIVE MAINTENANCE

A reliable and efficient maintenance plan is of vital importance to manufacturers, because passive shutdown caused by equipment failure may have a great impact on production efficiency and safety. Predictive Maintenance [58] is one of the key innovations proposed by Industry 4.0. It forecasts and optimizes the maintenance of equipment operation based on continuous measurement and analysis of the system.

At present, most manufacturers use preventive maintenance to improve the stability of production lines. Preventive maintenance is usually carried out on a time basis. Through regular maintenance, the probability of equipment failure or shutdown within a period of time is reduced. Preventive maintenance can reduce the total downtime of the production line and the number of failures caused by equipment deterioration, and has the advantages of convenient implementation and strong operability. However, as the maintenance time is determined by experience, there may be insufficient or excessive maintenance. With the maturity of Internet of Things, big data and other technologies, predictive maintenance makes maintenance more intelligent by analyzing real-time monitoring data of equipment, predicting possible failures of equipment, and proposing the causes and solutions of failures. The key to predictive maintenance is to troubleshoot hidden troubles in advance and solve them, which can reduce the total maintenance cost, failure rate and total downtime and improve the reliability of the equipment. However, due to the large number and variety of terminals, the practical application of predictive maintenance faces the problems of handling the connection and management of mass terminals, ensuring the real-time analysis and protecting the privacy of industrial data.

According to the American Journal of Efficient Plant, the emergence of edge computation is of great significance to the implementation of predictive maintenance schemes. The strong perception ability of edge nodes and their close proximity to equipment can meet the real-time and privacy protection requirements of predictive maintenance. Matt Boujonnier, an analysis application architecture engineer at Schneider Electric, pointed out that machine learning algorithms can usually only be run in cloud computing centers, but in Internet of Things applications, people hope that algorithms can be run wherever necessary. The Realift Rod Pump controller jointly developed by Schneider Electric and Microsoft Azure has realized real-time analysis and prediction of the operating state of the equipment at the edge of the network and has been applied in the oilfield industry on a pilot basis. Huawei also pointed out that Internet of Things based on edge computing (e.g., Edge Computing-IoT) can effectively construct predictive maintenance solutions, and has launched services for designing and deploying predictive maintenance solutions. Huawei uses intelligent gateways to provide intelligent services, real-time monitoring and analysis of key indicators of maintenance objects, prediction of possible failures of maintenance objects, and information reporting. The cloud computing center performs a comprehensive state evaluation based on the comprehensive information of multiple objects, and can continuously iterate and optimize the prediction algorithm run by edge nodes to realize dynamic deployment. This maintenance scheme comprehensively utilizes the advantages of edge nodes and cloud computing centers, and can meet the requirements of predictive maintenance for real-time and privacy protection, while ensuring the accuracy of fault prediction.

### E. SECURITY MONITORING

Vision is one of the important ways for human beings to know the world and obtain information. The ‘‘Skynet’’ monitoring system deployed by our public security organs maintains a stable and safe social order through a large number of cameras deployed in public areas. Many families also take the initiative to use household cameras and pet monitors to protect the safety of their houses and family members. At first, people can only process image information manually, but this method often has a long time delay and fluctuation accuracy. With the continuous development of artificial intelligence technology, image data computer already has very strong learning and processing capabilities. However, the traditional cloud computing model is difficult to support the application of image processing in some scenes. This paper takes railway track foreign body detection as an example to elaborate in more detail. First of all, since the railway tracks are laid outdoors and include tunnels, mountains and other areas where network quality cannot be guaranteed, the use of cloud computing models is likely to result in image loss or serious quality loss. Secondly, rail foreign matter detection requires very high real-time performance. Danger must be found in time and alarm must be made. Cloud computing processing link is long, and there is no guarantee of network bandwidth, so the real-time performance of detection and alarm may not be guaranteed. At the same time, rail monitoring images have certain confidentiality, once uploaded to the cloud, there is a risk of theft and tampering for criminals. Finally, the length of China's railway is 121,000 kilometers. If all the monitoring image data are uploaded to the cloud server for analysis, it is bound to require the cloud server to have extremely strong processing and computing capabilities. Obviously, image processing based on edge computation can provide better services in scenes with high real-time requirements, network quality cannot be guaranteed and privacy is involved.

Hikvision is a provider of Internet of Things solutions with video as the core, and has been deeply engaged in security monitoring for many years. In October 2017, Hikvision first publicly released the Cloud Edge Collaboration Architecture of the AI Cloud, which consists of a Cloud Computing Center, an Edge Domain and an Edge Node, enabling AI reasoning capability to edge, thus realizing fast and efficient perception, while the Cloud Computing Center focuses on global cognition and analysis. "Deep Eyes" binocular behavior analysis camera is a representative edge intelligent product of Hikvision. It has built-in high-performance GPU and artificial intelligence algorithm. It can also analyze and detect 9 behaviors such as crossing the warning line, wandering, running and leaving the post in an offline state. It can be applied to bank vaults, guard rooms, hospitals and other scenes. Academia is also optimistic about the application of edge computing in [59]. The system architecture consists of a monitoring application layer, a fog computing layer and a cloud computing layer. The fog computing layer includes cameras, smart tablets, smart phones and other devices. Fog nodes can track suspicious targets and calculate the driving speed in real time. The processing results will be sent to the cloud computing center. This mode can greatly reduce network traffic and improve the real-time performance of the system. Because the resources of edge intelligence equipment are usually limited, and artificial intelligence algorithms based on deep learning often need a lot of computing and storage resources to run, so the lightweight algorithm is also one of the research hotspots in the application of edge computing in security monitoring. In 2016 DeepScale Company proposed a lightweight target detection neural network-Squeeze Net [60]. After compression, the network volume is as low as 0.5 MB, but this algorithm cannot meet the real-time detection. In 2018, scholars from Binghamton University used a lightweight real-time detection and tracking algorithm in reference [61] to realize pedestrian recognition, tracking and abnormal behavior detection in public areas. Kerman algorithm [62] was proposed in June 2018. It uses a hybrid filter based on decision tree to construct a lightweight convolution neural network for human target tracking.

## VI. CONCLUSION

This paper systematically introduces the edge computing model from the aspects of basic concepts, architecture, key technologies, security and privacy protection. Edge computing provides data storage and computing at the edge of the network, and provides Internet intelligent services nearby, providing support for the digital transformation of various industries, and meeting the requirements of different industries for data diversification. Edge computing has become a hot research issue. In the future, with the continuous development of the Internet and human society, edge computing will play a more important role and effectively promote the development of various industries. It plays an important application role in Content Delivery Network(CDN), industrial Internet, energy, smart home, smart transportation, games and other fields.

## REFERENCES

[1] D. Evans. *The Internet of Things How The Next Evolution of the Internet is Changing Everything*. Accessed: Dec. 3, 2016. [Online]. Available: https://www.researchgate.net/publication/30612290

[2] V. Turner, J. F. Gantz, and D. Reinsel. (Nov. 26, 2018). *The digital universe of opportunities: Rich Data and the Increasing Value of the Internet of Things*. [Online]. Available: https://www.emc.com/leadership/digitaluniverse/2014iview/index.htm

[3] Y. Q. Gao, H. Bguan, and Z. W. Qi, "Service level agreement based energy-Effificient resource man agreement in cloud data centers," *Comput. Elect. Eng.*, vol. 40, no. 5, pp. 1621–1633, 2014, doi: 10.1016/j.compeleceng.2013.11.001.

[4] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, Jan. 2017.

[5] W. S. Shi, X. Z. Zhang, and Y. F. Wang, "Edge computing: State-of-the-art and future directions," *J. Comput. Res. Develop.*, vol. 56, no. 1, pp. 1–21, 2019.

[6] W. Shi, H. Sun, J. Cao, Q. Zhang, and W. Liu, "Edge computing-an emerging computing model for the Internet of everything era," *J. Comput. Res. Develop.*, vol. 54, no. 5, pp. 907–924, May 2017.

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

[8] Z. M. Zha, F. Liu, and Z. P. Cai, "Edge computing: Platforms; Applications and challenges," *J. Comput. Res. Develop.*, vol. 55, no. 2, pp. 327–337, 2018.

[9] X. Hong and Y. Wang, "Edge computing technology: Development and countermeasures," *Chin. J. Eng. Sci.*, vol. 20, no. 2, p. 20, 2018.

[10] "The Internet of Things reference model," in *Internet of Things World Forum*, Oct. 14. Chicago, IL, USA: C. S. Inc, 2014, pp. 1–12.

[11] X. Sun and N. Ansari, "Edge IoT: Mobile edge computing for the Internet of Things," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 22–29, Dec. 2016.

[12] A. Alrawais, A. Alhothaily, C. Hu, and X. Cheng, "Fog computing for the Internet of Things: Security and privacy issues," *IEEE Internet Comput.*, vol. 21, no. 2, pp. 34–42, Mar. 2017.

[13] J. Kang, R. Yu, X. Huang, and Y. Zhang, "Privacy-preserved pseudonym scheme for fog computing supported Internet of vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 8, pp. 2627–2637, Aug. 2018.

[14] C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow, and P. A. Polakos, "A comprehensive survey on fog computing: State-of-the-Art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 416–464, 1st Quart., 2018.

[15] Z. M. Xu and Y. F. Tian, "The history and application of cloud computing," *Inf. Recording Mater.*, vol. 19, no. 8, pp. 66–67, 2018.

[16] L. Jiao, R. Friedman, X. M. Fu, S. Secci, Z. Smoreda, and H. Tschofenig, "Cloud based computation off loading for mobile devices: State of the art, challenges and opportunities," *Future Netw. Mobile Summit*, Lisbon, Portugal, Jul. 2013, pp. 1–11.

[17] I. Stojmenovic, "Fog computing: A cloud to the ground support for smart things and machine-to-machine networks," in *Proc. Australas. Telecommun. Netw. Appl. Conf. (ATNAC)*, Nov. 2014, pp. 117–122.

[18] S. Yangui, P. Ravindran, O. Bibani, R. H. Glitho, N. Ben Hadj-Alouane, M. J. Morrow, and P. A. Polakos, "A platform as-a-service for hybrid cloud/fog environments," in *Proc. IEEE Int. Symp. Local Metrop. Area Netw. (LANMAN)*, Jun. 2016, pp. 1–7.

[19] X. Zhu, D. S. Chan, and M. S. Prabhu, "Improving video performance with edge servers in the fog computing architecture," *Intel Technol. J.*, vol. 19, no. 1, pp. 202–224, 2015.

[20] J. L. Zhang, Y. C. Zhao, and B. Chen, "Survey on data security and privacy-preserving for the research of edge computing," *J. Commun.*, vol. 39, no. 3, pp. 1–21, 2018.

[21] Q. Qi and F. Tao, "A smart manufacturing service system based on edge computing, fog computing, and cloud computing," *IEEE Access*, vol. 7, pp. 86769–86777, 2019.

[22] L. Ruan, Y. Yan, S. Guo, F. Wen, and X. Qiu, "Priority-based residential energy management with collaborative edge and cloud computing," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1848–1857, Mar. 2020.

[23] T. Wang, G. Zhang, A. Liu, M. Z. A. Bhuiyan, and Q. Jin, "A secure IoT service architecture with an efficient balance dynamics based on cloud and edge computing," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4831–4843, Jun. 2019.

[24] M. S. Hossain and G. Muhammad, "Emotion recognition using secure edge and cloud computing," *Inf. Sci.*, vol. 504, pp. 589–601, Dec. 2019.

[25] Y. Jararweh, A. Doulat, O. AlQudah, E. Ahmed, M. Al-Ayyoub, and E. Benkhelifa, "The future of mobile cloud computing: Integrating cloudlets and mobile edge computing," in *Proc. 23rd Int. Conf. Telecommun. (ICT)*, May 2016, pp. 1–5.

[26] M. Armbrust, A. Fox, and R. Griffith, *Above the Clouds: A Berkeley View of Cloud Computing*. Berkeley, CA, USA: EECS Dept., 2009.

[27] J. Ren, H. Guo, C. Xu, and Y. Zhang, "Serving at the edge: A scalable IoT architecture based on transparent computing," *IEEE Netw.*, vol. 31, no. 5, pp. 96–105, Aug. 2017.

[28] H. Bangui, S. Rakrak, S. Raghay, and B. Buhnova, "Moving to the Edge-Cloud-of-things: Recent advances and future research directions," *Electronics*, vol. 7, no. 11, p. 309, 2018.

[29] S. Carlini, "The drivers and benefits of edge computing," in *Schneider Electric–Data Center Science Center*, 2016.

[30] C. K. M. Lee, Y. Z. Huo, S. Z. Zhang, and K. K. H. Ng, "Design of a smart manufacturing system with the application of multi-access edge computing and blockchain technology," *IEEE Access*, vol. 8, pp. 28659–28667, 2020.

[31] W. Huo, F. Liu, L. Wang, Y. Jin, and L. Wang, "Research on distributed power distribution fault detection based on edge computing," *IEEE Access*, vol. 8, pp. 24643–24652, 2020.

[32] J. Barthélemy, N. Verstaevel, H. Forehead, and P. Perez, "Edge-computing video analytics for real-time traffic monitoring in a smart city," *Sensors*, vol. 19, no. 9, p. 2048, 2048.

[33] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, 3rd Quart., 2017.

[34] C. T. Ding, J. N. Cao, L. Yang, and S. G. Wang, "Edge computing: Applications, state-of-the-art and challenges," *Zte Technol.*, vol. 25, no. 3, pp. 2–7, 2019.

[35] A. Nadembega, A. S. Hafid, and R. Brisebois, "Mobility prediction model-based service migration procedure for follow me cloud to support QoS and QoE," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[36] L. I. N. Kaiqun and C. T. C. F. Branch, "Pilot study on application of MEC local shunting service," *Modern Inf. Technol.*, vol. 1, no. 3, pp. 65–67, 2017.

[37] F. Lu, J. Hu, L. T. Yang, Z. Tang, P. Li, Z. Shi, and H. Jin, "Energy-efficient traffic offloading for mobile users in two-tier heterogeneous wireless networks," *Future Gener. Comput. Syst.*, vol. 105, pp. 855–863, Apr. 2020.

[38] D. Zhu and X. D. Wang, "Mobile network edge computing and caching technology," *Railway Comput. Appl.*, vol. 26, no. 8, pp. 51–54, 2017.

[39] R. Wang, M. Li, L. Peng, Y. Hu, M. M. Hassan, and A. Alelaiwi, "Cognitive multi-agent empowering mobile edge computing for resource caching and collaboration," *Future Gener. Comput. Syst.*, vol. 102, pp. 66–74, Jan. 2020.

[40] G. S. Aujla, N. Kumar, A. Y. Zomaya, and R. Ranjan, "Optimal decision making for big data processing at edge-cloud environment: An SDN perspective," *IEEE Trans. Ind. Informat.*, vol. 14, no. 2, pp. 778–789, Feb. 2018.

[41] R. Roman, J. Lopez, and M. Mambo, "Mobile edge computing: A survey and analysis of security threats and challenges," *Future Gener. Comput. Syst.*, vol. 78, pp. 680–698, 2018.

[42] D. Zissis and D. Lekkas, "Addressing cloud computing security issues," *Future Gener. Comput. Syst.*, vol. 28, no. 3, pp. 583–592, Mar. 2012.

[43] H. Liang, D. Huang, L. X. Cai, X. Shen, and D. Peng, "Resource allocation for security services in mobile cloud computing," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2011, pp. 191–195.

[44] I. Stojmenovic and S. Wen, "The fog computing paradigm: Scenarios and security issues," in *Proc. Federated Conf. Comput. Sci. Inf. Syst.*, Sep. 2014, pp. 1–8.

[45] A. Sahai and B. Waters, "Fuzzy identity-based encryption," in *Proc. 24th Annu. Int. Conf. Theory Appl. Cryptograph. Techn. (EUROCRYPT)*, 2005, pp. 457–473.

[46] M. Blaze, G. Bleumer, and M. Strauss, "Divertible protocols and atomic proxy cryptography," in *Proc. 17th Annu. Int. Conf. Theory Appl. Cryptograph. Techn. (EUROCRYPT)*, 1998, pp. 127–144.

[47] R. L. Rivest, L. Adleman, and M. L. Dertouzos, "On data banks and privacy homomorphisms," *Found. Secure Comput.*, vol. 4, pp. 169–179, Oct. 1978.

[48] K. Yang and X. Jia, "Data storage auditing service in cloud computing: Challenges, methods and opportunities," *World Wide Web*, vol. 15, no. 4, pp. 409–428, Jul. 2012.

[49] A. N. Toosi, R. N. Calheiros, and R. Buyya, "Interconnected cloud computing environments: Challenges, taxonomy, and survey," *ACM Comput. Surv.*, vol. 47, no. 1, pp. 1–47, Jul. 2014.

[50] D. S. Touceda, J. M. S. Cámara, S. Zeadally, and M. Soriano, "Attribute-based authorization for structured Peer-to-Peer (P2P) networks," *Comput. Standards Interfaces*, vol. 42, pp. 71–83, Nov. 2015.

[51] Y. Yang, X. Zheng, X. Liu, S. Zhong, and V. Chang, "Cross-domain dynamic anonymous authenticated group key management with symptom-matching for e-health social system," *Future Gener. Comput. Syst.*, vol. 84, pp. 160–176, Jul. 2018.

[52] I. Khalil, A. Khreishah, and M. Azeem, "Consolidated identity management system for secure mobile cloud computing," *Comput. Netw.*, vol. 65, pp. 99–110, Jun. 2014.

[53] M. Zhang and Y. Sun, "Analysis of the development trend of China's Internet industry in 2017," *China Telecom*, vol. 5, pp. 67–72, 2018.

[54] O. Makinen, "Streaming at the edge: Local service concepts utilizing mobile edge computing," in *Proc. 9th Int. Conf. Next Gener. Mobile Appl., Services Technol.*, Sep. 2015, pp. 1–6.

[55] M. Huang, A. Liu, N. N. Xiong, T. Wang, and A. V. Vasilakos, "An effective service-oriented networking management architecture for 5G-enabled Internet of Things," *Comput. Netw.*, vol. 173, May 2020, Art. no. 107208.

[56] T. Li, M. Zhao, and K. K. L. Wong, "Machine learning based code dissemination by selection of reliability mobile vehicles in 5G networks," *Comput. Commun.*, vol. 152, pp. 109–118, Feb. 2020.

[57] M. Chen, T. Wang, K. Ota, M. Dong, M. Zhao, and A. Liu, "Intelligent resource allocation management for vehicles network: An A3C learning approach," *Comput. Commun.*, vol. 151, pp. 485–494, Feb. 2020.

[58] Mobley, R. Keith. *An Introduction to Predictive Maintenance*. Amsterdam, The Netherlands: Elsevier, 2002.

[59] N. Chen, Y. Chen, S. Song, C.-T. Huang, and X. Ye, "Smart urban surveillance using fog computing," in *Proc. IEEE/ACM Symp. Edge Comput. (SEC)*, Oct. 2016, pp. 95–96.

[60] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and & 0.5MB model size," 2016, *arXiv:1602.07360*. [Online]. Available: http://arxiv.org/abs/1602.07360

[61] S. Y. Nikouei, Y. Chen, and T. R. Faughnan, "Smart surveillance as an edge service for real-time human detection and tracking," in *Proc. IEEE/ACM Symp. Edge Comput. (SEC)*, Oct. 2018, pp. 336–337.

[62] S. Y. Nikouei, Y. Chen, S. Song, and T. R. Faughnan, "Kerman: A hybrid lightweight tracking algorithm to enable smart surveillance as an edge service," in *Proc. 16th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2019, pp. 1–6.

**KEYAN CAO** received the M.S. and Ph.D. degrees in computer science and technology from Northeastern University, China, in 2009 and 2014, respectively. She is currently an Associate Professor with the College of Information and Control Engineering, Shenyang Jianzhu University. Her current research interests include data management, cloud computing, and query process and optimization.

**YEFAN LIU** received the B.S. degree from the Shenyang University of Technology, China. He is currently pursuing the master's degree with the School of Computer Science and Technology, Shenyang Jianzhu University. His current research interest is big data processing.

**GONGJIE MENG** received the bachelor's degree in software engineering from the Jincheng College, Nanjing University of Aeronautics and Astronautics, in 2019. He is currently pursuing the master's degree in software engineering with Shenyang Jianzhu University. He is currently conducting research on pulmonary.

**QIMENG SUN** was born in Tieling, China, in 1994. She received the B.Eng. degree from Shenyang Urban Construction University, China, in 2017. She is currently pursuing the master's degree with the School of Computer Technology, Shenyang Jianzhu University. Her research interest is big data analysis.