# An overview of compute first networking

Liang Tian

Institute of Computer and Information Engineering,
Xinxiang University,
Henan, China
Email: tianliang@xxu.edu.cn

Mingzhe Yang*

Institute of Network Technology,
Beijing University of Posts and Telecommunications,
Beijing, China
Email: yangmingzhe@bupt.edu.cn
*Corresponding author

Shangguang Wang

Institute of Network Technology,
Beijing University of Posts and Telecommunications,
Beijing, China
and
Peng Cheng Laboratory,
Shenzhen, China
Email: sgwang@bupt.edu.cn

**Abstract:** Edge computing has become an important innovative business model in the 5G era, especially its low latency feature, which is considered to be not available in traditional solutions. Therefore, edge computing can provide more service capabilities and has a wider range of application scenarios. However, the collaboration of computing power between edge computing and cloud computing in the central position has become a new technical problem. That is, it is necessary to achieve cloud-to-network collaboration, cloud-to-edge collaboration, or even edge-to-edge collaboration among edge computing, cloud computing, and network to achieve the optimisation of resource utilisation. On the basis of studying the compute distribution and scheduling requirements of edge computing, this paper introduces a compute network scheme based on the deep fusion of cloud, edge and network, compute first networking. Firstly, we introduce the basic concept and related work of compute first networking. Then, we mainly discuss the framework and key technology of compute first networking. After that, we present some applications with respect to compute first networking. Finally, we discuss the challenges and opportunities in the area of compute first networking.

**Keywords:** compute first networking; CFN; mobile edge computing; cloud computing; 5G.

**Reference** to this paper should be made as follows: Tian, L., Yang, M. and Wang, S. (xxxx) 'An overview of compute first networking', *Int. J. Web and Grid Services*, Vol. x, No. x, pp.xxx–xxx.

**Biographical notes:** Liang Tian received his Master's at the Huazhong University of Science and Technology in 2009. He is an Associate Professor at the Xinxiang University Computer and Information Engineering College. His research interests include service computing, cloud computing, and mobile edge computing.

Mingzhe Yang received his Bachelor in Telecommunication and Management from the Beijing University of Posts and Telecommunications in 2015. He is currently a PhD candidate at the Beijing University of Posts and Telecommunications. His research interests include mobile augmented reality and mobile edge computing.

Shangguang Wang received his PhD at the Beijing University of Posts and Telecommunications in 2011. He is a Professor and the Vice Director at the State Key Laboratory of Networking and Switching Technology (BUPT). He has published more than 100 papers, and played a key role at many international conferences, such as general chair and PC chair. His research interests include service computing, cloud computing, and mobile edge computing. He is a senior member of the IEEE, and the Editor-in-Chief of the *International Journal of Web Science*.

This paper is a revised and expanded version of a paper entitled [title] presented at [name, location and date of conference].

## 1 Introduction

Nowadays, with the development of 5G technology, the world has set off a wave of digital transformation in the industry. A typical feature of an intelligent society is the deep integration of the physical world and the digital world. In the future, the digital world will interact with the real world through sensors and actuators provided by technologies such as internet of things (Li et al., 2018) and augmented reality (Sukhmani et al., 2018). The network realises data flow as a bridge connecting the physical world and the digital world.

According to the forecast of the Cisco Annual Internet Report (2018–2023), the number of terminal devices connected to the network will be greater than 29 billion, where more than 50% of the network data needs to be analysed, processed, and stored at the network edge. The transmission, analysis, and storage of massive data pose great challenges to traditional network and cloud computing, making cloud computing and network face a situation of 'unstable transmission, immobility, and unsustainability'. This situation drives computing from the cloud to the edge of the network which is closer to data sources. This forms a distributed computing resource in the network. Under such a trend, a single decentralised site has limited resource and it is difficult to guarantee service quality. Therefore, while the network realises the interconnection of decentralised nodes, it also needs to have the ability to coordinately schedule

the network and computing power, and dynamically dispatch services to the optimal compute node for processing through the optimal path (Mao et al., 2017).

However, the current deployment of edge computing faces various challenges (Roman et al., 2018). First, from a network perspective, edge computing and even ubiquitous computing scenarios have limited computing power for a single node. Edge nodes cannot perceive each other and cannot work together. Computing tasks cannot be scheduled to the optimal edge node for computation. In the prior art, computing tasks are generally managed through a centralised orchestration layer, but centralised architectures have scalability and scheduling performance. Secondly, from the perspective of business requirements, the existing business application layer is decoupled from the network. The application layer cannot accurately grasp the status of the network in real-time. The overall performance of the addressing results dominated by the application layer may not be optimal or even worse, resulting in an unbalanced network load. Service cannot be scheduled to the optimal edge node, resulting in poor business experience. At the same time, the current Internet assumption is the static server plus mobile client mode. This mode cannot guarantee the consistency of user experience in edge computing scenarios, and it is difficult to take advantage of dynamic microservices and ubiquitous computing. Also, it cannot guarantee the maximum computing efficiency. Under the general trend of network and computing convergence, a high degree of collaboration between network and computing is required. Based on ubiquitous connections, we interconnect ubiquitous computing to achieve efficient collaboration among the cloud, edge, and network. It can improve the efficiency of computing resource utilisation, and thus achieves consistency in user experience (Mach and Becvar, 2017).

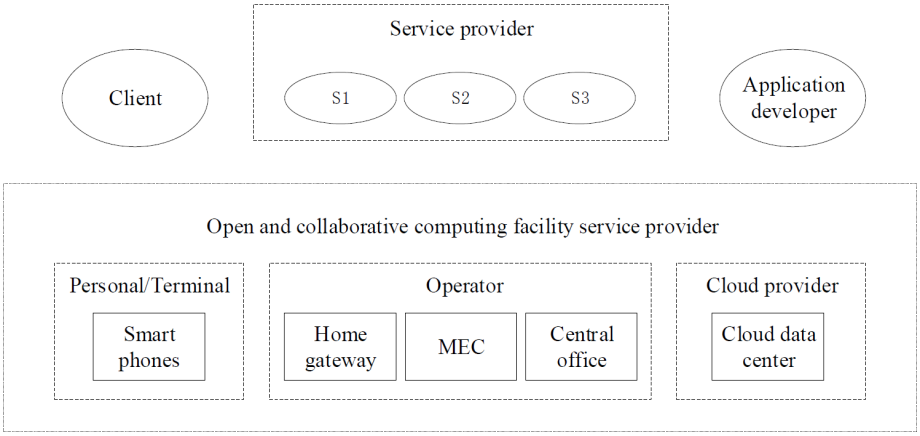
In order to solve the above problem, this paper introduces a new architecture based on distributed system for computing network fusion, compute first networking (CFN) (Li et al., 2019; Geng and Willis, 2019; Gu et al., 2019). CFN achieves the optimisation of user experience, resource utilisation and network efficiency. For example, the edge nodes in residential area receive very few requests during working hours, and the number of requests received during non-working hour is high. The number of requests received by the edge nodes of the industrial park is exactly the opposite of that of residential areas. This phenomenon can cause large differences in computing load on different edge nodes. At peak times, the computing resource attached to the nearest edge node may not be sufficient to handle all incoming requests. Users may experience longer response time and even requests are dropped. It is neither feasible nor economical to increase the computing resource hosted on each edge node to the potential maximum capacity. Traditional static or hash-based service dispatch cannot adapt to the unbalanced nature of computing load and the rapid changes in computing load on different edge nodes. One edge node (such as the edge closest to the client) may be overloaded, while the other edge node may still have a large amount of computing resource to satisfy the requests. In order to effectively utilise the computing resource hosted on all edges, service requests should be dynamically allocated and processed to balance the consumption of computing and network resource.

CFN assumes that there are multiple equivalent edge nodes providing resource for a single service (Geng and Willis, 2019). A single edge node has limited computing resource, and different edge nodes may have different resource for serving a particular service at a particular time. CFN handles a large number of requests by sharing the computing resource of multiple edge nodes in a cooperative manner. That is, a service request can be processed by different service nodes located at different edges. CFN must

determine which node is best suited to serve the request in real-time. By intelligently distributing the workload of multiple edge nodes, CFN can improve system utilisation to serve more end users. CFN considers network conditions and available computing resource at the same time, so that edge nodes can interact with each other to provide network-based service allocation to achieve better load balancing. In general, CFN is a network-based approach, so requests can be dispatched to the optimal edge based on available computing resource and running network status (Król et al., 2019).

CFN improves the current end-to-end model of computing at the edge of the network, supporting a sunflower-like network model. Computing is embedded in the middle of the network like sunflower seeds. CFN is a network model of dynamic, distributed computing and deep integration of the network. At the same time, CFN changes the services that users apply on the internet. Currently, users obtain original information through the network. The network only serves as a carrier for information transmission and does not have computing process capabilities. In the future, CFN provides network-as-service capability by providing connectivity and computing. The user directly obtains the computation result through the network.

**Figure 1** CFN-based industrial ecological chain




As an open infrastructure, CFN enables massive applications, services, and computing resource, which is beneficial to break through the traditional closed computing industry and build a more open and win-win industrial ecological environment (Lei et al., 2019). As shown in Figure 1, users, service providers and application developers work together to build an open industry ecosystem. Users obtain services through the terminal. Operators perform CFN operations with the MEC through the cloud. Developers can release applications more freely based on CFN. Any computing resource and service model can be published to CFN. Each application and developer can use it as needed. The above-mentioned computing is based on the edge computing provided by the operator's site and equipment. In the long run, it also supports the sharing of compute in future smart terminals to achieve crowdfunding computing. At the same time, in the future, it is also possible to introduce decentralised transaction models, such as blockchain (Xiong et al., 2018) and smart contracts (Pan et al., 2018). Users make purchase based on the service conditions of computing and services, enable new business models, and achieve the monetisation of network and computing resource.

The main object of this article is to review and discuss existing and emerging CFN technologies. Firstly, we introduce the basic concepts and conclude the characteristics of CFN. Then, we discuss the framework of CFN and introduce each key function module in detail. After that, we introduce several related technologies of CFN. Finally, we present some CFN applications and discuss the challenges and opportunities in the area of CFN.

The remainder of this article is organised as follows. Section 2 presents the basic concept of CFN. Section 3 reviews the CFN framework in detail. In Section 4, we introduce some key technologies regarding to CFN. Then we present some CFN applications in Section 5. Section 6 discusses the challenges and opportunities. Finally, concluding remarks are given in Section 7.

## 2 Concept and related work

### *2.1 Concept*

CFN is a new network architecture in response to the development trend of computing network convergence (Lei et al., 2019). It interconnects dynamically distributed computing resource based on ubiquitous network connections. Through the unified and coordinated scheduling of multi-dimensional resource such as network, storage, and computing power, massive applications can call computing resource in different places on demand and in real-time, realise the global optimisation of connections and computing power, and provide a consistent user experience.

At present, the industry does not have a standard definition of CFN, but this paper believes that CFN needs to meet the following four characteristic requirements:

- *Resource abstraction:* CFN needs to abstract computing resource, storage resource, network resource (especially connection resource in a wide area), and algorithm resource, and offers them to customers as part of the product.
- *Business guarantee:* CFN divides service levels by business requirement, rather than simply by region. It promises customers' service-level agreement (SLA) such as network performance and compute, and shields the underlying differences such as heterogeneous computing, different types of network connections.
- *Unified management and control:* CFN uniformly controls cloud computing nodes, edge computing nodes, network resource and wide area networks. It performs unified scheduling of compute resource, corresponding network resource, and storage resource according to business requirements.
- *Flexible scheduling:* CFN monitors business traffic in real-time, dynamically adjusts compute resource, and completes various tasks to efficiently process and integrate the output. On the premise of meeting business requirements, it achieves elastic scaling of resource and optimises the allocation of compute.

In summary, the compute network is a new type of information infrastructure that flexibly allocates and schedules computing resource, storage resource, and network resource among the cloud, network, and edge according to business reqiurements.

### 2.2 Related work

Li et al. (2019) proposed a CFN framework that enables service requests to be sent to the best edge to improve overall system load balancing. This paper defined routing protocols for distributing computing resource information and dynamic anycast based on late binding on the control plane and data plane, respectively. Geng and Willis (2019) introduced some CFN application scenarios and proposed some CFN related requirements. Gu et al. (2019) proposed a field test for the CFN system to demonstrate the effects that CFN can achieve. Field trials have shown that CFN can greatly increase the overall query speed of services hosted on multiple edges per second in a more balanced manner. It is a feasible and effective approach to provide multi-edge service balance in edge computing. Król et al. (2019) used CFN to design a computing graph representation for distributed programs, which has the advantages of simplicity, performance and fault recovery capabilities. Lei et al. (2019) proposed a CFN solution based on the deep integration of cloud, network, and edge. And he gave a typical implementation system for AI applications. This solution can effectively deal with the multi-level deployment of computing, storage, network, and even algorithm resources in the future business.

## 3 Framework

In order to achieve computational awareness, interconnection, and collaborative scheduling for ubiquitous computing and service, the CFN architecture system can be divided into five functional modules: compute service layer, compute platform layer, compute resource layer, compute routing layer, and network resource layer, as shown in Figure 2 (Li et al., 2019). Based on the ubiquitous compute resource on the network, the compute platform layer completes the abstraction, modelling, control and management of compute resource. It informs compute routing layer through compute announcement module. The compute routing layer comprehensively considers user requirements, network resource conditions, and computing resource conditions. It schedules service applications to the appropriate node to achieve optimal resource utilisation and ensures the ultimate user experience, details as follows:

- *Compute service layer:* Based on the fractional microservice architecture, the compute service layer supports the application to decompose atomic functional components and forms an algorithm library. It is uniformly dispatched by application programming interface (API) gateway to realise atomic algorithms on demand in ubiquitous computing resource. Through the I1 interface, the compute service layer passes the SLA and other information of the business application to the compute platform layer.
- *Compute platform layer:* Perception, measurement and operations, administration and maintenance (OAM) management of compute resource need to be completed to support the network's perception, measurement, management and control of compute resource. In the face of heterogeneous computing resource, the compute platform layer first abstracts and expresses the compute resource through compute modeling, forms a compute template, and passes it to the corresponding network nodes. In addition, it is also necessary to monitor the performance of the compute

resource and notify the corresponding network nodes of the performance and failure information of the compute resource. Specifically, the compute platform layer includes:

a *Compute modeling:* In the face of heterogeneous computing resource, we first need to study the measurement dimension and the system of compute resource. This sub-module uses the information such as general algorithm or customary requirement to form the corresponding compute capacity template. Several compute capacity templates are combined into compute contracts to meet the compute requirements of the business.

b *Compute OAM:* This sub-module includes compute performance monitoring, compute billing management, and compute fault management.

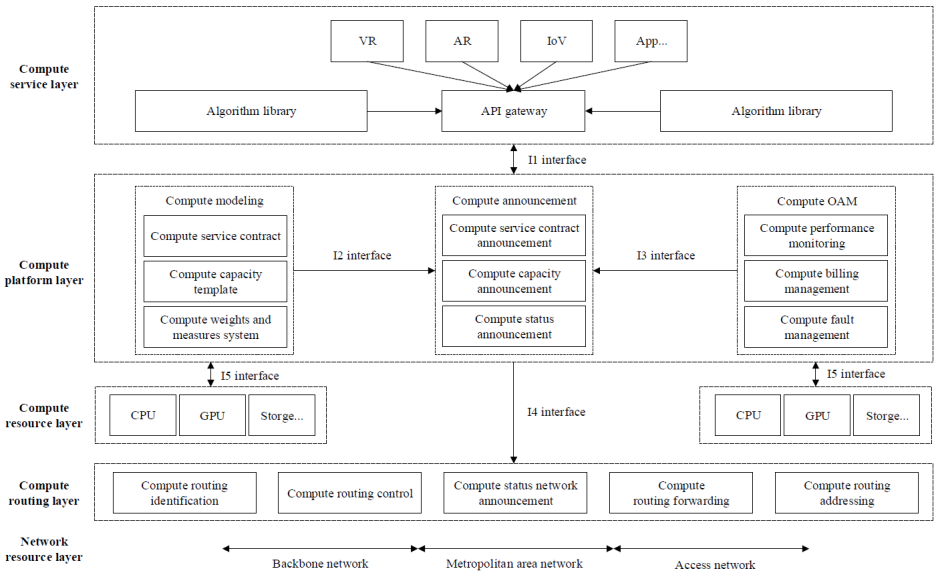
c *Compute announcement:* This submodule is responsible for abstractly representing the actual deployed compute resource through the compute template. Together with information such as the compute service contract, it informs the corresponding network node. This sub-module includes compute service contract announcement, compute capability announcement, and compute status announcement. Compute service contract announcement refers to generating compute service requirements based on compute service layer's SLA requirements and notifying the corresponding network nodes. Compute capability announcement means that the actual deployed compute resource are notified to the corresponding network nodes after being abstractly represented by the compute template. Compute status announcement refers to notifying the real-time status of compute resource to the corresponding network node through the I4 interface.

- *Compute resource layer:* In order to meet the diverse computing requirements in the field of edge computing, the industry has proposed a concept of diverse computing, which is oriented to different applications. From single-core CPU to multi-core CPU to a variety of compute combinations, computing innovation is promoted under the system-level restoration of Moore's law. Faced with a variety of heterogeneous computing resource distributed in the network, the compute resource layer needs to implement a measurement system of compute resource and the approach of abstract representation.
- *Compute routing layer:* Based on the abstracted computing resource discovery, it comprehensively considers the network status and computing resource status, and flexibly schedules services to different computing resource nodes on demand. The specific functions include compute routing identification, compute routing control, compute status network announcement, compute routing forwarding and compute routing addressing.
- *Network resource layer:* It provides network infrastructure for information transmission, including backbone network, metropolitan area network and access network.

The CFN architecture not only defines the functional modules such as compute routing layer, compute resource layer and compute platform layer, but also defines the interfaces between some functional modules:

- *I1 interface:* It defines the interface between compute service layer and compute platform layer, which is used to transfer SLA requirements and configuration information of compute service deployment.
- *I2 interface:* It is used to transfer compute service contracts, compute capability templates and other information from the compute modeling module to the compute announcement module.
- *I3 interface:* It is used to transfer compute resource performance monitoring, compute billing management, compute resource failure and other information from the compute OAM module to the compute announcement module.
- *I4 interface:* It is used to transfer compute service contract information and compute resource status announcement from the compute platform layer to the compute routing layer.
- *I5 interface:* It refers to the interface between the compute resource layer and the compute platform layer, which is mainly used for compute resource registration management, performance status and failure information transmission.

**Figure 2** CFN framework

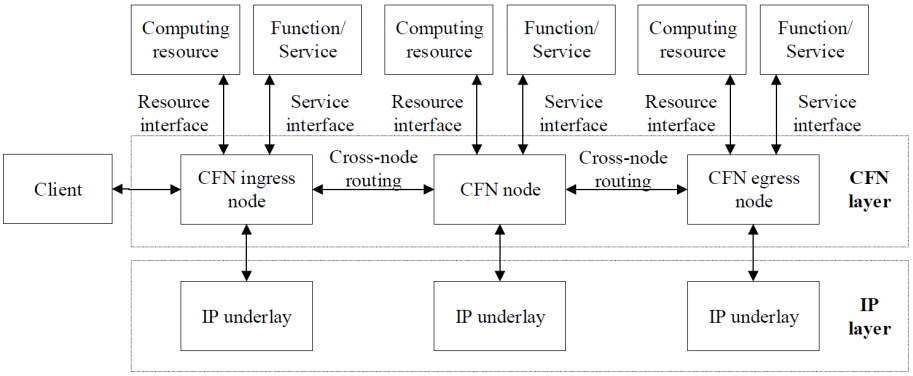



In summary, CFN is a new network architecture for deep integration of computing networks. Based on ubiquitous network connections and highly distributed computing nodes, CFN truly achieves omnipresent networks, compute and intelligence through automatic deployment of services, optimal routing and load balancing. Mass applications, functions, and computing resource constitute an open ecosystem. Among them, massive applications can call computing resource in different places on demand and in real-time. It improves the utilisation efficiency of computing resource, and finally realises the optimisation of user experience, computing resource utilisation and network efficiency.

## 4 Key technology

CFN realises the global optimisation of connection and compute on the network through the unified management and collaborative scheduling of multi-dimensional resource such as network, storage, and compute. It provides the ultimate user experience. CFN comprehensively considers the network and computing resource. It dispatches different applications to appropriate computing nodes for processing to achieve edge-to-edge collaboration.

**Figure 3** CFN technology framework




### *4.1 CFN technology framework*

The compute routing layer, as a key technology of the CFN technology architecture, is located between the IP layer and the application layer. It realises the optimal scheduling of services through the collaborative network and compute status information (Li et al., 2019; Geng and Willis, 2019). As shown in Figure 3, the CFN technology architecture includes:

- *CFN internal network element:* The CFN node connects to distributed service endpoints. Its basic function is based on the routing of service ID and data ID in the request. Advanced feature also includes data prefetching that can improve efficiency and user experience. The computing routing node includes CFN ingress node and CFN egress node. The ingress and egress nodes can be the same node. The ingress node faces the client and is responsible for real-time addressing and traffic scheduling of services. The egress node faces the server, and is responsible for querying, aggregating and publishing the service status. Service is a unit in the CFN node service registry and represents an application with a unique service ID. A service is composed of multiple service endpoints. These endpoints are implemented by workload instances running on containers or virtual machines. Each network endpoint is distinguished by IP address.
- *CFN protocol:* It contains abstracted computing resource discovery, topology, routing generation and healing at the computing resource level. Topology and

routing are not only IP reachability, but also contains dynamic changes and serviceability of computing resource. Healing is not only the healing of routing, but also includes the rerouting and scheduling of computing tasks, so that it can complete the computation on the appropriate computing service node.

### 4.2 CFN routing protocol

The routing table contains computing performance and network performance data. By adding computing performance's evaluation parameters such as computing remaining capacity, computing delay, and other extensible parameters to the routing table, the system weights the sum of network performance and computing performance, selects the optimal execution node, and performs route forwarding for computing services. Because the selected route is based on the principle of optimal computation, the delay is greatly reduced, and the edge compute can meet the requirements of low-latency applications.

When the local routing node receives the data packet of the computing task, the system first determines the type of the computing task of the data packet, such as service ID and flow viscosity's demand attribute (Hossain et al., 2018). Based on the correspondence relationship among the type of computing task, the other computing nodes, and the computing performance acquired in advance, it determines at least one other node corresponding to the computing task type and its corresponding computing performance. Based on the computing performance of other nodes, as well as the network performance between the local node and other nodes such as link status, the system comprehensively considers and determines the target node for execution. The address of the target node is the routing destination address of the data packet, and then forwards the data packet based on the target address. Computing and network performance information can be diffused and synchronised in the CFN by extending existing border gateway protocol (Rekhter et al., 2006), interior gateway protocols (Sidhu et al., 1993; Hedrick, 1988), etc.

### 4.3 Flow viscosity retention

For some business flows that need to be kept in the same service node, if the flow viscosity cannot be guaranteed, there will be flow interruption, packet loss, and traffic chaos (Yu et al., 2017). Therefore, CFN needs to maintain the flow stickiness of this type of business. The service submits the flow stickiness requirements to the CFN, including the flow stickiness type and timeout time. The system generates and writes the flow viscosity demand data table and spreads it in the CFN network. When a new flow arrives, the flow sticky data table is queried based on the service IP, port, etc. And a forwarding table entry for the flow is generated based on the flow sticky data. The usual approach is to establish a path information table in the CFN node based on the attributes of the service. The path information table records path information of the session, such as source address, service ID, and service node IP. Therefore, the system can direct service packets belonging to the same session to the same service node according to the path information table. With the huge amount of access brought by the rise of edge computing and the internet of things (Yu et al., 2017), CFN node needs to establish a huge amount of session flow table to maintain the flow stickiness of each session. The cost burden of CFN node is large or even impossible. Therefore, in the implementation,

it is necessary to record the forwarding flow table based on the service ID on the client, and carry the service ID and the service node IP in subsequent messages to keep the network device lightly loaded.

### *4.4 Trusted telecommunications blockchain*

As a disruptive technology, blockchain is leading a new round of technological and industrial changes in the world (Zhang et al., 2019), and promoting the transition from 'internet of information' to 'internet of value'. In order to apply this technology in the communication field, the concept of 'token' has been introduced in the relevant ITU standards for the sharing and transaction of data or resource (Anjum et al., 2017). For CFN, in order to facilitate the quantification, sharing and trading of compute, the concept of 'compute token' can be used.

Blockchain is no longer just a technology, a tool, but also a way of thinking. Blockchain as a new type of technology combination, its decentralisation, difficult tampering, and non-repudiation not only bring a new credit model to the telecommunications industry, but also make its digital services more competitive. In turn, it helps the telecommunications industry to reduce costs and brings a new perspective to this field.

With the development of communication network technology, many services have put forward new requirements for broadband and delay. In order to improve the user experience, a lot of massive data needs to be localised or processed at the network edge in the future, which can reduce the network load and obtain lower latency. Operators can open up the compute of edge computing. In terms of transaction mode, they can consider the combination of blockchain. The use of alliance chains in technology can be more efficient and better meet the requirements of supervision and auditing. In actual deployment, blockchain platforms or applications can be installed and deployed on edge computing servers to provide blockchain technology and capability support for different application scenarios. The combination of MEC and blockchain technology can provide rich compute and other shared resource for video live broadcast, local cache and other services (Rahman et al., 2018). GPU resource can also be used for AI training. In addition, MEC operators can use on-chain points or off-chain payment approaches to return. After the transaction, the MEC can operate and use the corresponding resource on the access blockchain.

## 5 Application

CFN is the basic network architecture that carries the ubiquitous compute in the future, and the distributed MEC nodes are interconnected through compute routing nodes. Facing the wireless access network scenario, 5G sinks further along with the core network. Edge computing nodes are further distributed. CFN contributes to the highly distributed MEC scale deployment. Facing the fixed access network, the CFN connects the edge computing nodes in different locations and the central cloud to form a converged business network, thereby achieving plug-and-play computing resource and solving the problem of multiple copies of services and dynamic services.

The existing network architecture is mainly based on the application layer and applies domain name system for addressing. Since the network state and the change of

the destination node's compute are not considered, its comprehensive performance is poor in some cases. For example, when the computing load of the destination node is too high, or the network is congested, the user experience may decline or even become unacceptable.

For computing services, CFN considers the real-time network status in the current network and the computing status of serviceable computing resource according to business requirements. Through flexible matching and dynamic scheduling, the CFN routes the computing tasks of the terminal to the appropriate target computing node to support the computing requirements of the business and ensure the user experience of the business.

Through intelligent collaborative scheduling of edge-to-edge and edge-to-cloud resource, CFN achieves cloud, edge, and network load balancing, improving network efficiency, resource utilisation, and user experience. Typical application scenarios include mobile augmented reality (MAR) (Qiao et al., 2018), cloud virtual reality (VR) (Erol-Kantarci and Sukhmani, 2018) and internet of vehicles (Wan et al., 2019), as shown in Figure 4.

**Figure 4** CFN applications (see online version for colours)

Augmented Reality Virtual Reality Internet of Vehicles

### *5.1 Cloud-based recognition in MAR*

In the MAR environment, terminal devices capture images through cameras and issue computationally intensive service requirements. Generally, the edge service node is responsible for tasks with low latency requirements and medium computational complexity, such as image preprocessing, object recognition, and feature extraction. The service nodes in the cloud are responsible for the most intensive computing tasks, such as object recognition, template matching. The terminal device only handles tasks such as tracking and registration and image display, which reduces the computing burden on the client (Ren et al., 2019).

Computing resource for specific services at edge nodes can be instantiated on demand. After the task is completed, the edge node can release this resource. This 'function as a service' life cycle can be in the order of milliseconds. Therefore, the computing resource of the edge nodes are dynamic and distributed. The service request must be sent to the edge with sufficient computing resource and a good network path for processing. CFN can improve the overall performance of the MAR system through the reasonable allocation of edge node resource.

### 5.2 Cloud VR video service

Cloud VR introduces the concepts and technologies of cloud computing and cloud rendering into VR business applications. With the help of a high-speed and stable network, the system encodes and compresses the cloud's display output and sound output. Then it transmits them to the user's terminal device to realise the cloud rendering of VR business content. The cloud VR service has extremely high requirements for network and computation (Mangiante et al., 2017). For example, the entry-level cloud VR (full-view 8K 2D video) uses a 110-degree field of view transmission. Typical network requirements are 40 Mbps bandwidth, 20 ms RTT, and 2.4E-5 packet loss rate. Typical computing requirements are 8 K H.265 real-time hard decoding, 2 K H.264 real-time hard coding, and multi-channel parallel computing capabilities (Huawei iLab, 2017). Based on the above requirements, CFN distributes computing tasks to the central cloud and the edge cloud through collaborative optimisation of distributed computing and network resource. CFN deploys tasks such as multi-channel parallel computing and content generation in cloud VR that are under heavy computing load on the central cloud. Tasks such as video coding and decoding, content rendering which require less computation, are dynamically offloaded to the edge node to complete. Through hierarchical offloading of computing tasks, CFN improves cloud, edge, and network resource utilisation and business experience, which has a positive significance for cloud VR from pilot to scale deployment.

### 5.3 Internet of vehicles

Internet of vehicles' business scenarios include assisted driving business and on-board entertainment business (Zhang and Letaief, 2019). Through resource allocation and scheduling, CFN can efficiently and reliably complete a variety of services. For road traffic conditions due to occlusion and blind spots, CFN obtains comprehensive traffic information around the vehicle location under the edge node and performs unified data processing, so as to issue warning signals to vehicles with potential safety hazards and assist vehicle safety drive.

In the internet of vehicles' scenario, CFN schedules traffic according to business priorities. High-priority business traffic is scheduled to the nearest node for computation, such as early warning information. Non-real-time traffic is dispatched to remote nodes or the cloud for processing, such as on-board entertainment services.

When the local edge node is overloaded, the auxiliary safe driving announcement will be delayed, which may cause a traffic accident (Contreras-Castillo et al., 2017). CFN dispatches delay-insensitive services from the local node to other nodes for computation, reducing the load on the local node, so that low-latency services are preferentially processed locally, ensuring its user experience and availability.

## 6 Challenges and opportunities

### 6.1 Service upgrade for operators

Operators have control of the two most important physical resource, a large number of computer room sites and fiber optic and cable network with good coverage (National

Research Council, 2000). With the full exploration of the potential of optical fiber and cable, the room of the computer room is idle. CFN provides operators with a feasible path for service upgrade. CFN utilises a large number of computer room resource to avoid extensive, low-level rental space. CFN can provide credible and guaranteed integrated computing and network integrated services for high-value customers who require bandwidth and delay in various industries.

### *6.2 Control plane intelligence*

With the separation of 5G transfer control, cloudification of control functions and three-level distributed deployment of the network, it not only brings network flexibility, but also brings a rapid increase in the complexity of integration, operation and maintenance (Kliks et al., 2018). The traditional manual operation and maintenance has been difficult to support. In the future, big data analysis and AI will gradually be introduced to alternative labour. However, excessively complex network protocols also increase the difficulty of implementing AI algorithms. In addition to the intelligent control plane, it is also necessary to simplify the forwarding plane protocol and network topology. The intelligent control plane can promote the CFN to operate more efficiently and intelligently, thereby improving the user experience.

### *6.3 Compute aware technology*

In the traditional core network, the existing service application layer is decoupled from the network (Kreutz et al., 2014). It is difficult for the application layer to accurately control the network performance in real-time. The addressing results based on the application layer are difficult to ensure optimal or even poor, resulting in poor service experience. Therefore, for the problem that traditional core networks are difficult to accurately sense compute in real-time, we can study compute aware approach to support the rapid discovery of compute and its attribute information. For CFN, we can study the unified modeling of heterogeneous compute and determine the compute measurement index. By designing the compute mapping and conversion model in CFN, we can build the compute index and express the compute to construct compute perception and measurement system. In addition, we can also perform clustering and anomaly detection on compute state information data based on machine learning theory, and study data mining technology for compute state information.

### *6.4 Compute opening and sharing mechanism*

The compute opening and sharing of multiple edge core networks to third parties makes CFN more efficient (Shahzadi et al., 2017). We can use the technical characteristics of blockchain decentralisation, smart contracts, collective maintenance, and reliable data to study the 5G core network's compute opening and sharing mechanism, while also ensuring the security, verifiability, and non-tampering of compute transactions. Since the compute of the 5G core network is based on service encapsulation, in order to better support and promote the compute opening and sharing, we can use the distributed consensus mechanism in the blockchain to design a compute opening oriented distributed ledger to achieve on-demand compute transaction.

## 7 Conclusions

As the most important innovation scenario in the 5G era, edge computing can provide customers with various service guarantees such as low latency and large bandwidth. However, with the deepening of research and deployment, the collaboration among edge computing, cloud computing, and networks (especially wide area networks) has become a new research point. This paper introduces a solution based on cloud, network, and edge deep fusion, CFN, to meet the requirements of on-demand deployment and flexible compute scheduling between multi-level computing nodes. We first discussed in detail the concept and related work of CFN. Then we discussed the framework and key technology of CFN. After that, we presented some CFN applications. Finally, we discussed the challenges and opportunities in CFN area. In the future, CFN will become a universal service presence for operators. This is an important direction for the future development of network and computing integration. Large bandwidth, low latency, simple intelligence, MEC, and trusted blockchain will become the key technologies and indicators of CFN.

## Acknowledgements

This work is supported by the national natural science foundation of China (61922017) and the verification platform of multi-tier coverage communication network for oceans (LZC0020).

## References

Anjum, A., Sporny, M. and Sill, A. (2017) 'Blockchain standards for compliance and trust', *IEEE Cloud Computing*, Vol. 4, No. 4, pp.84–90.

Cisco Annual Internet Report (2018–2023) *White Paper*, pp.1–35.

Contreras-Castillo, J., Zeadally, S. and Guerrero-Ibañez, J.A. (2017) 'Internet of vehicles: architecture, protocols, and security', *IEEE Internet of Things Journal*, Vol. 5, No. 5, pp.3701–3709.

Erol-Kantarci, M. and Sukhmani, S. (2018) 'Caching and computing at the edge for mobile augmented reality and virtual reality (AR/VR) in 5G', in *Ad Hoc Networks*, pp.169–177, Springer.

Geng, L. and Willis, P. (2019) 'Compute first networking (CFN) scenarios and requirements', pp.1–7.

Gu, S., Zhuang, G., Yao, H. and Li, X. (2019) 'A report on compute first networking (CFN) field trial', pp.1–14.

Hedrick, C.L. (1988) 'Routing information protocol', *RFC*, Vol. 1058, pp.1–33.

Hossain, M.S., Muhammad, G. and Amin, S.U. (2018) 'Improving consumer satisfaction in smart cities using edge computing and caching: a case study of date fruits classification', *Future Generation Computer Systems*, Vol. 88, pp.333–341.

Huawei iLab (2017) *Cloud VR Bearer Networks*.

Kliks, A., Musznicki, B., Kowalik, K. and Kryszkiewicz, P. (2018) 'Perspectives for resource sharing in 5G networks', *Telecommunication Systems*, Vol. 68, No. 4, pp.605–619.

Król, M., Mastorakis, S., Oran, D. and Kutscher, D. (2019) 'Compute first networking: distributed computing meets ICN', in *Proceedings of the 6th ACM Conference on Information-Centric Networking*, pp.67–77.

Kreutz, D., Ramos, F.M.V., Verissimo, P.E., Rothenberg, C.E., Azodolmolky, S. and Uhlig, S. (2014) 'Software-defined networking: a comprehensive survey', *Proceedings of the IEEE*, Vol. 103, No. 1, pp.14–76.

Lei, B., Liu, Z., Wang, X., Yang, M. and Chen, Y. (2019) 'Computing network: a new multi-access edge computing', *Telecommunications Science*, Vol. 9, pp.44–51.

Li, S., Xu, L.D. and Zhao, S. (2018) '5G internet of things: a survey', *Journal of Industrial Information Integration*, Vol. 10, pp.1–9.

Li, Y., He, J., Geng, L., Liu, P. and Cui, Y. (2019) 'Framework of compute first networking (CFN)', pp.1–14.

Mach, P. and Becvar, Z. (2017) 'Mobile edge computing: a survey on architecture and computation offloading', *IEEE Communications Surveys & Tutorials*, Vol. 19, No. 3, pp.1628–1656.

Mangiante, S., Klas, G., Navon, A., GuanHua, Z., Ran, J. and Silva, M.D. (2017) 'VR is on the edge: how to deliver 360 videos in mobile networks', in *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, pp.30–35.

Mao, Y., You, C., Zhang, J., Huang, K. and Letaief, K.B. (2017) 'A survey on mobile edge computing: the communication perspective', *IEEE Communications Surveys & Tutorials*, Vol. 19, No. 4, pp.2322–2358.

National Research Council (2000) *Networking Health: Prescriptions for the Internet*, National Academies Press.

Pan, J., Wang, J., Hester, A., AlQerm, I., Liu, Y. and Zhao, Y. (2018) 'Edgechain: an edge-IoT framework and prototype based on blockchain and smart contracts', *IEEE Internet of Things Journal*, Vol. 6, No. 3, pp.4719–4732.

Qiao, X., Ren, P., Dustdar, S. and Chen, J. (2018) 'A new era for web ar with mobile edge computing', *IEEE Internet Computing*, Vol. 22, No. 4, pp.46–55.

Rahman, M.D.A., Hossain, M.S., Loukas, G., Hassanain, E., Rahman, S.S., Alhamid, M.F. and Guizani, M. (2018) 'Blockchain-based mobile edge computing framework for secure therapy applications', *IEEE Access*, Vol. 6, pp.72469–72478.

Rekhter, Y., Li, T. and Hares, S. (2006) 'A border gateway protocol 4 (BGP-4)', *RFC*, Vol. 4271, pp.1–104.

Ren, J., He, Y., Huang, G., Yu, G., Cai, Y. and Zhang, Z. (2019) 'An edge-computing based architecture for mobile augmented reality', *IEEE Network*, Vol. 33, No. 4, pp.162–169.

Roman, R., Lopez, J. and Mambo, M. (2018) 'Mobile edge computing, Fog et al.: a survey and analysis of security threats and challenges', *Future Generation Computer Systems*, Vol. 78, pp.680–698.

Shahzadi, S., Iqbal, M., Dagiuklas, T. and Ul Qayyum, Z. (2017) 'Multi-access edge computing: open issues, challenges and future perspectives', *Journal of Cloud Computing*, Vol. 6, No. 30, pp.1–13.

Sidhu, D., Fu, T., Abdallah, S., Nair, R. and Coltun, R. (1993) 'Open shortest path first (OSPF) routing protocol simulation', *ACM SIGCOMM Computer Communication Review*, Vol. 23, No. 4, pp.53–62.

Sukhmani, S., Sadeghi, M., Erol-Kantarci, M. and El Saddik, A. (2018) 'Edge caching and computing in 5G for mobile AR/VR and tactile internet', *IEEE Multimedia*, Vol. 26, No. 1, pp.21–30.

Wan, S., Li, X., Xue, Y., Lin, W. and Xu, X. (2019) 'Efficient computation offloading for internet of vehicles in edge computing-assisted 5G networks', *The Journal of Supercomputing*, Vol. 75, pp.1–30.

Xiong, Z., Zhang, Y., Niyato, D., Wang, P. and Han, Z. (2018) 'When mobile blockchain meets edge computing', *IEEE Communications Magazine*, Vol. 56, No. 8, pp.33–39.

Yu, W., Liang, F., He, X., Hatcher, W.G., Lu, C., Lin, J. and Yang, X. (2017) 'A survey on the edge computing for the internet of things', *IEEE Access*, Vol. 6, pp.6900–6919.

Zhang, J. and Letaief, K.B. (2019) 'Mobile edge intelligence and computing for the internet of vehicles', *Proceedings of the IEEE*, Vol. 108, No. 2, pp.246–261.

Zhang, K., Zhu, Y., Maharjan, S. and Zhang, Y. (2019) 'Edge intelligence and blockchain empowered 5G beyond for the industrial internet of things', *IEEE Network*, Vol. 33, No. 5, pp.12–19.