China Mobile
Research Institute
China Mobile CMRI

White Paper on the Technical Architecture of Fully Scheduled Ethernet

The Technical Framework White Paper of

Global Scheduling Ethernet

(2023)

China Mobile Communications Research Institute

.

Preface

This white paper is aimed at the large-scale construction of future intelligent computing centers and the development and deployment requirements of AI large-scale models. It jointly proposes the Global Switching Ethernet (GSE) technology architecture with industrial partners, aiming to break through the network performance bottleneck of intelligent computing centers and create a new type of intelligent computing center network that is free of congestion, has high bandwidth and ultra-low latency, and facilitates the rapid development of high-performance services such as AIGC.

Contents

List of abbreviations

Abbreviation	Full name in English	Chinese explanation
AI	Artificial Intelligence	Artificial Intelligence
AIGC	AI-Generated Content	Artificial intelligence-generated content
CPU	Central Processing Unit	Central Processing Unit
DPU	Data Processing Unit	Data processing unit
ECMP	Equal Cost Multi Path	Equivalent multipath routing
ECN	Explicit Congestion Notification	Explicit congestion notification
FC	Fibre Channel	Optical Fiber Channel
GPU	Graphics Processing Unit	Graphics Processing Unit
GSF	Global Scheduling Fabirc	Full-switching exchange network
GSOS	Global Scheduling Operating System	Full-scheduling operating system
GSP	Global Scheduling Processor	Full scheduling network processing node
HoL	Head-of-line blocking	Team leader blockage
JCT	Job Completion Time	Task completion time
ML	Machine Learning	Machine learning
PFC	Priority-based Flow Control	Priority-based traffic control
PHY	Physical	Port physical layer
PKTC	Packet Container	Message container
RDMA	Remote Direct Memory Access	Remote Direct Memory Access
RoCE	RDMA over Converged Ethernet	Integrate Ethernet to carry RDMA
VOQ	Virtual Output Queue	Virtual output queue
DGSQ	Dynamic Global Scheduling Queue	Dynamic Global Dispatch Queue

1. Background and Requirements

At present, AIGC (AI-Generated Content) is developing rapidly, with an exponential growth in the speed of iteration. The economic value worldwide is expected to reach trillions of dollars. In the Chinese market, the application scale of AIGC is expected to exceed 200 billion yuan by 2025. This huge potential has attracted leading enterprises in the industry to competitively launch large models with parameters at the level of hundreds or thousands of billions. The deployment scale of underlying GPU computing power has also reached the level of tens of thousands of cards. Take GPT3.5 as an example. Its parameter scale reaches 175 billion, and the amount of Internet text used as the training dataset exceeds 45TB. Its training process relies on the AI supercomputing system specially built by Microsoft and the high-performance network cluster composed of 10,000 V100 GPUs, with a total computing power consumption of approximately 3640 PF-days (that is, 100 million billion calculations per second for 3640 days).

Distributed parallel computing is a key means to achieve large-scale AI model training, typically involving various parallel computing modes such as data parallelism, pipeline parallelism, and tensor parallelism. All parallel modes require multiple communication operations among computing devices. Additionally, synchronous mode is often adopted during the training process, and the next round of iteration or calculation can only be performed after the communication operations are completed among multiple machines and cards. As the underlying communication connection infrastructure, the intelligent computing center network needs to possess high-performance and low-latency communication capabilities. Once the network performance is poor, it will affect the quality and speed of distributed training.

In response to the future demands of large-scale construction of intelligent computing centers and the development and deployment of large AI models, China Mobile, in collaboration with multiple partners, has launched the full-scheduling Ethernet technology solution (GSE), creating a new type of intelligent computing center network with no congestion, high bandwidth and ultra-low latency, to facilitate the rapid development of high-performance services such as AIGC.

2. Introduction to the GSE Network Architecture

2.1 Overall Design Goals

The fully scheduled Ethernet is designed for high-performance computing scenarios such as AI and HPC. The architectural design follows the following three major principles:

The fully scheduled Ethernet builds an open, transparent and standardized technical system for all upstream industries involved in the high-performance computing ecosystem, such as chips (GPU, DPU, CPU, etc.), devices (servers, switches, network cards, etc.), meters, operating systems, etc., to use together.

Full-switching Ethernet can adapt to various high-performance computing scenarios. It can be universally applied to all business scenarios involving lossless, high bandwidth utilization rate, and ultra-low latency requirements.

Full scheduling Ethernet is not reinventing Ethernet; instead, it integrates high-performance computing requirements into Ethernet, maximizing the reuse of the Ethernet physical layer and being compatible with the Ethernet ecosystem, such as optical modules and PHY layer chips.

2.2 Overview of the overall architecture

To build a high-performance network with no congestion, high bandwidth and low latency, the GSE architecture emerged. This architecture mainly consists of three levels: the computing layer, the network layer and the control layer, and includes four types of devices: computing nodes, GSPs, GSFs and GSOSs.

2.2.1 Overall Architecture of GSE

Full-scheduling Ethernet is a new type of Ethernet architecture with non-blocking, high throughput, and low latency. It can better serve high-performance computing and meet the requirements of AI large model deployment, training, and inference. The full-scheduling Ethernet architecture is divided into three layers from top to bottom: the control layer, the network layer, and the computing layer. The key point lies in the innovation of introducing a brand-new dynamic global queue scheduling mechanism. The Dynamic Global Scheduling Queue (DGSQ) is different from the traditional VOQ. Instead of being statically assigned based on ports in advance, it is created on demand and dynamically based on the target device port of the data flow. To save the number of queues, it can even be created on demand based on congestion feedback from the target or intermediate devices. Based on DGSQ scheduling to achieve high throughput, low latency, and balanced scheduling at the entire network level.

- Control Layer: It contains the global centralized GSOS, as well as the distributed NOS at the GSP and GSF device ends. Among them, the centralized GSOS is used to provide global network information and implement addressing based on global information (for example, setting)

Such as standby node ID, etc.), daily operation and maintenance management and other functions. The distributed NOS at the device end has an independent control plane and management plane, and can run network functions such as load balancing of containers and DGSQ scheduling that belong to the device itself. Through the distributed management and control capabilities of the device, the reliability of the entire network is enhanced.

- Network Layer: Through the division of labor and collaboration of GSP and GSF, a switching network with the integration of technologies such as orderly scheduling of the entire network traffic, load balancing among various links, and fine reverse pressure on network anomalies is constructed. It is the main implementation layer of the fully scheduled Ethernet. Among them, the Fabric part can support the Layer 2 GSF extension to meet the networking requirements of larger scales.

- Computing layer: It contains high-performance computing cards (GPU or CPU) and network cards, serving as the full-scheduling Ethernet boundary. Initially, the computing nodes will be used as the full-scheduling Ethernet boundary, and the training performance of the computing cluster will be improved only by optimizing the switching network capability. In the future, considering the deep integration of computing and network, the relevant GSP solutions will be extended to the network card layer or the direct output network card module of the GPU to form a full-scheduling Ethernet with algorithm-network synergy, further enhancing the high-performance computing performance.

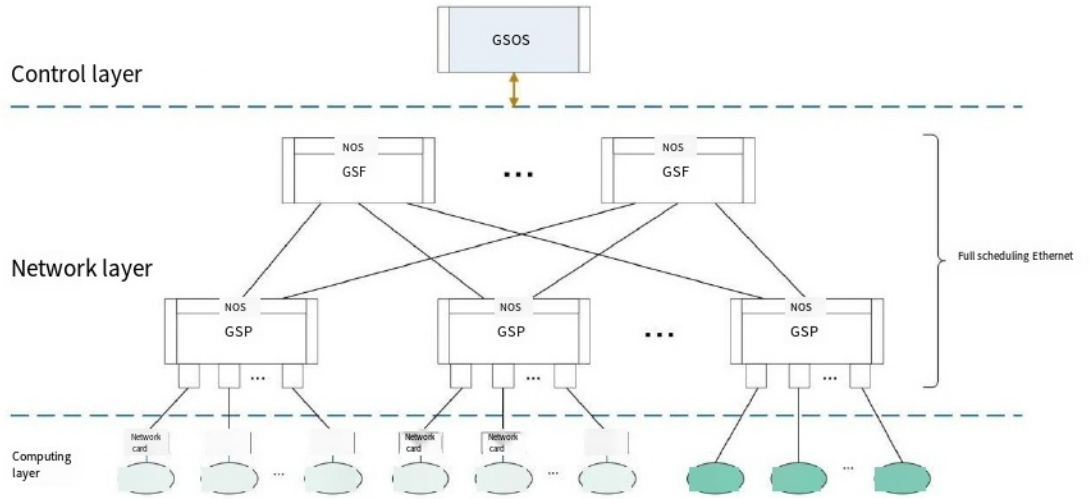

Figure 2-1 Hierarchical Architecture of GSE Technology

2.2.2 GSE architecture devices

The GSE architecture consists of four types of devices: computing nodes, GSPs, GSFs and GSOSs. These devices work collaboratively, and the division of labor is as follows:

- Computing nodes: namely, the computing cards and network cards on the server side, providing high-performance computing capabilities.

- GSP: Network edge processing node, used to access computing traffic and conduct global scheduling of the traffic; when the traffic is upstream, it has the ability of dynamic load balancing. When the traffic is downstream, it has the ability of traffic sequencing.

- GSF: Network core switching node, as the upper-level equipment of GSP, is used for flexible network expansion planning.

The mold has the ability of dynamic load balancing and the ability to release backpressure information.

- GSOS: Fully scheduled operating system, providing centralized network operating system capabilities for the entire network management and control.

2.2.3 GSE Architecture Features

Considering the rapid development of AI/ML applications such as AIGC and the current status of large-scale deployment of standard Ethernet, the GSE architecture should have flexible scalability and maximum compatibility with Ethernet features.

The specific characteristics of the GSE architecture are as follows:

- Flexible expansion: Support the deployment of high-performance computing clusters of Wanke, with the two-layer network of GSP + GSF as the common form, and support horizontal expansion. When the computing nodes expand further and the two-layer network architecture is insufficient to support, it can be flexibly expanded into the three-layer network architecture of GSP + GSF + GSF, retaining the ability to expand to more layers of GSF networking to meet the business deployment requirements.

- Ecological openness: Adhere to the principle of ecological openness, construct a standard and open technology protocol stack, facilitate the interconnection and communication among multi-vendor devices, jointly build the network layer of the fully scheduled Ethernet, and provide an efficient network foundation for large-scale distributed computing.

- General hardware: All network nodes support standard Ethernet, without the need for dedicated cell processing nodes, and can seamlessly switch with standard Ethernet devices. Among them, although GSP and GSF devices have different role divisions, they are all based on Ethernet message exchange. The forwarding hardware is universal, and the device roles can be controlled by software versions, thereby supporting more flexible deployment and maintenance.

2.3 Key technical features

2.3.1 Compatible with Ethernet technology

Ethernet standard is currently one of the most universal communication standards. China Mobile, with the aim of being universal and open, jointly builds the GSE network with the industry chain, maximizing compatibility with the existing Ethernet standards. The compatibility is mainly reflected in the following aspects:

- Follow the existing Ethernet PHY and MAC layer protocols: Follow the definitions of the physical layer and MAC layer of Ethernet in the existing IEEE 802.3 protocol to be compatible with existing Ethernet devices (containing optical modules, network cards, switches, etc.), integrate GSE in the form of functional increments into the existing Ethernet, and enhance the Ethernet.

- Complete Ethernet service message transmission: In the entire GSE network, in the form of a complete Ethernet message

Carry out the transmission while retaining the integrity of the payload content of the Ethernet packets to the greatest extent, in order to be compatible with more features in the GSE network subsequently, such as in-network computing.

☒ Follow the existing management and control system and operation and maintenance habits: The construction of the management and control system and the operation and maintenance system is as complex as the Ethernet forwarding technology, and the collaborative system with the transfer control plane has matured. The GSE network largely continues to use the existing management and control and operation and maintenance systems, maintaining the unchanged architecture and operation and maintenance habits, and ensuring the compatibility and inheritance of the existing management and operation and maintenance methods of the Ethernet.

2.3.2 Non-blocking network

With the continuous expansion of the network scale, message exchange has evolved from single-hop within a single network node to multi-hop between network nodes, and the relationship between each node has also changed from loose coupling to joint forwarding. The industry builds large-scale distributed forwarding structures through the CLOS architecture to meet the increasing forwarding scale requirements. Under this architecture, each node operates distributively and makes self-decisions on the forwarding path, which cannot achieve the optimal overall network performance. To make the forwarding effect of large-scale multi-node forwarding consistent with that of a single node, the blocking problem within the distributed forwarding structure needs to be solved.

The core reason for network congestion is that each node in the distributed forwarding structure cannot fully perceive the global information. When a network node sends to another network node, it cannot perceive the network conditions of the downstream nodes, resulting in congestion of traffic in the downstream. For example, in a network based on ECMP for load balancing, the network node only sends traffic through hash routing from its own perspective, eventually leading to problems such as link congestion, outport congestion, and low utilization of the switching network. The DGSQ technology is the key technology to solve this problem. This technology combines the mutually invisible network nodes through the mapping of the global queue of the switching network, ultimately achieving the optimal forwarding effect of the entire network.

2.3.3 Improve effective bandwidth

Based on the DGSQ technology, it can be ensured that the traffic sent from the entry node of the distributed switching network to the switching network is optimal from the perspective of the exit node. However, when the traffic is switched in the network, the traditional ECMP load balancing can lead to uneven link load and hash polarization. Especially in the presence of giant flows, no matter how long the giant flow lasts, wherever it goes, congestion and packet loss may occur. Currently, the switching network lacks effective bandwidth control and priority management, and packet loss will be indiscriminate, which will have a direct negative impact on applications. Based on the Packet-based per-packet load sharing technology, any traffic is transformed into extremely short data units for transmission, completely eliminating the problem of hash polarization and thereby improving the bandwidth utilization of the switching network.

2.3.4 Optimize long-tail latency technology

There are a large number of Map-Reduce traffic models in the training of AI large-scale models. The end of any round of calculation depends on the return of the last result, reducing the long-tail delay of the network can effectively improve the training completion time. The overall forwarding delay of the switching network is positively correlated with the congestion of intermediate nodes on the forwarding path. Eliminating the congestion of intermediate nodes can eliminate the long-tail delay. The integration of DGSQ scheduling and high-precision load balancing technology is the key to solving this problem. On the one hand, through the PUSH + PULL combination mechanism of DGSQ, the amount of message data entering the switching network is controlled not to exceed the forwarding capacity of the entire network. On the other hand, with the support of high-precision load balancing, both can eliminate the congestion of any node in the switching network.

3. GSE Network Core Technology

Different from the mechanism of load sharing based on flow in traditional Ethernet, the GSE switching network uses fixed-length PKTC for packet forwarding and dynamic load balancing. By constructing a fully scheduled DGSQ mechanism based on PKTC, a fine backpressure mechanism, and a perception-free self-healing mechanism, it achieves precise control in microbursts and fault scenarios, comprehensively improving the effective bandwidth and forwarding delay stability of the network.

The specific traffic forwarding process is shown as follows:

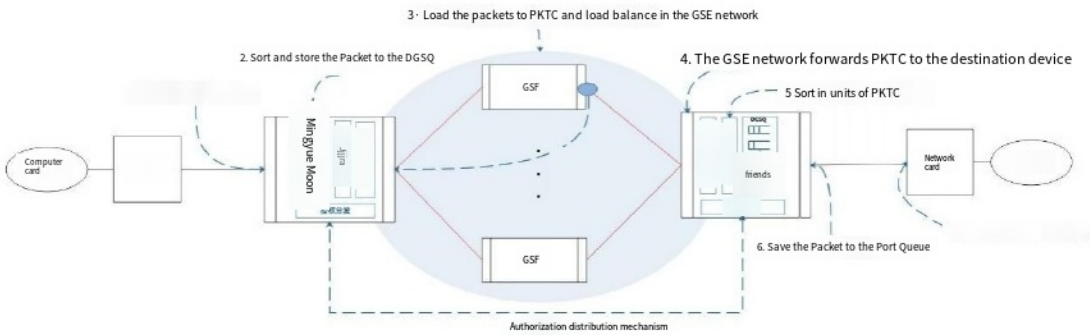

Figure 3-1 Schematic diagram of end-to-end traffic forwarding of the GSE network

(1) After the source-end GSP device receives the Packet from the computing side, it finds the final exit through the forwarding table and allocates the message to the corresponding DGSQ for authorization scheduling on demand based on the final exit.

(2) After the source GSP device obtains authorization, Packet will follow the load balancing requirements of PKTC and send the packets to the GSE network.

(3) When the message reaches the destination GSP device, it is first sorted at the PKTC level and then through the forwarding table

The packets are stored in the queues of the physical Port and eventually sent to the computing nodes through port scheduling.

3.1 PKTC Mechanism

PKTC is a core forwarding mechanism different from CELL forwarding. Under this mechanism, Ethernet packets logically form virtual containers and are transmitted in the switching network with this container as the minimum unit. This section will be elaborated from three aspects: the concept of PKTC, the overhead of PKTC, and the location of PKTC.

3.1.1 PKTC Concept

When implementing load balancing based on packet forwarding, it is necessary to overcome the influence of randomly generated packet lengths first. Therefore, the basic forwarding unit of load balancing needs to be normalized and a fixed-length packet container should be established. The setting of the number of packets that the packet container can accommodate can be adjusted based on the distribution of the lengths of service packets. It is required that it can accommodate at least one of the longest service packets, and the total length should be as short as possible within the limits of the chip's forwarding capacity and disordering resolution ability, in order to achieve the purpose of fine division of data streams and fully improve the degree of instant load balancing.

To solve the above problems, this scheme proposes the concept of message containers, and the design principle is shown in the following figure:

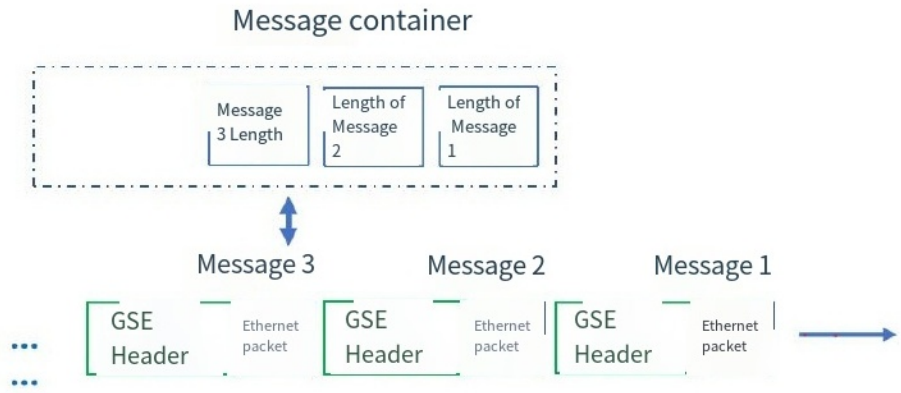

Figure 3-2 Schematic diagram of the PKTC forwarding mechanism

The implementation of the message container is logically virtual. When a message enters the GSP node, the GSP node will record information such as the number of the message container to which it belongs and the number of bytes it occupies in that container. When the number of bytes of the message exceeds the set length of the virtual message container, the message will be scheduled and recorded in the next message container.

Each node of the GSE network directly forwards packets without caching them to construct actual containers. For all packets belonging to the same packet container, they will be load-balanced to a unique path for forwarding in the switching network.

To ensure that the messages within the message container are no longer out of order, in order to reduce the disordering pressure of the exit GSP node.

3.1.2 PKTC Overhead

Based on the per-packet forwarding mechanism, relevant information needs to be carried in the data packets to be correctly identified, processed and sent to the target node by the switching network. Therefore, when the packets enter the GSP, the DGSQ needs to be differentiated. The identification of the DGSQ is related to the establishment of the target of the system DGSQ. Generally, a unique DGSQ identification can be established based on the source device, the target port and the priority under that port. Of course, the granularity of the DGSQ can also be simplified according to the business requirements. For example, 4, 2 or 1 priorities can be set under a target port to reduce the demand for the DGSQ queue and the overhead of the switching chip.

The packets entering DGSQ need to go through downlink scheduling authorization before being sent to the switching network. At this time, the packets sent from the same entry Leaf node to the same exit Leaf node can be organized into a disorder queue. That is, the same sequence number (the sequence of the container) and the source GSP ID are added to all data packets within each packet container. After receiving these packets downlink, disorder queue processing can be performed based on the source GSP ID and sequence number.

The following figure describes the construction and forwarding principle of the message container under the construction of other internal Ethernet messages by taking the addition of the standard Ethernet message header as an example.

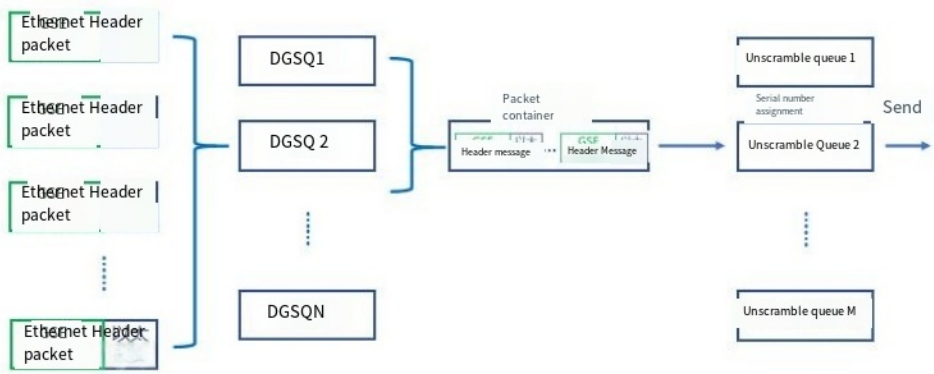

Figure 3-3 Schematic diagram of the construction method of PKTC header

3.1.3 Location of GSE Header

The GSE network requires adding additional information to the business packets for global load balancing forwarding and sequencing. There are three ways to carry this information, including:

☒Adding a standard extension header outside the standard Ethernet frame: The greatest advantage of this carrying method is that it does not damage the original

Business messages, but there will be certain losses in terms of compatibility and transmission efficiency. If the external Ethernet Tunnel mode is chosen to improve the compatibility of Ethernet, the transmission efficiency will be further reduced.

Figure 3-4 Standard Extended Head Mode

☒ Redefine the standard Ethernet frame: Redefine the MAC header of the message. The greatest advantage of this transmission mode is its high transmission efficiency, but its compatibility with Ethernet is poor and it can only be used in specific scenarios.

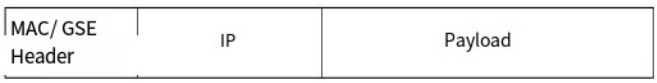

Figure 3-5 Redefining the Ethernet Frame Mode

☒Expanding the protocol header after Ethernet MAC or IP has the greatest advantage of balancing the compatibility and transmission efficiency of Ethernet. However, the processing of additional information of GSE in the network requires delving into the internal information of the packets, which will affect the forwarding delay.

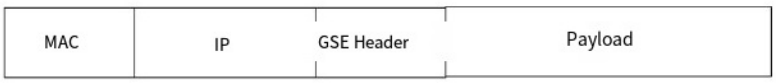

Figure 3-6 Protocol Header Expansion Mode

3.2 Load balancing technology based on PKTC

To reduce and eliminate the long-tail latency or packet loss caused by problems such as hash polarization and uneven load in the traditional ECMP forwarding model, the technology based on Packet Container can be divided into three parts: load information construction, dynamic path switching, and traffic sequencing mechanism.

3.2.1 Construction of Dynamic Load Information

After evaluating and quantifying the load information of the egress port, one of the links with a lighter load can be randomly selected to provide a basis for the subsequent PKTC path selection of traffic. As shown in the forwarding model in the following figure, GSP1 serves as the access switch. When a certain PKTC passes through the GSP1 switch to the A2 port of GSP2, the upstream link needs to be evaluated for load to determine the transmission port of this PKTC.

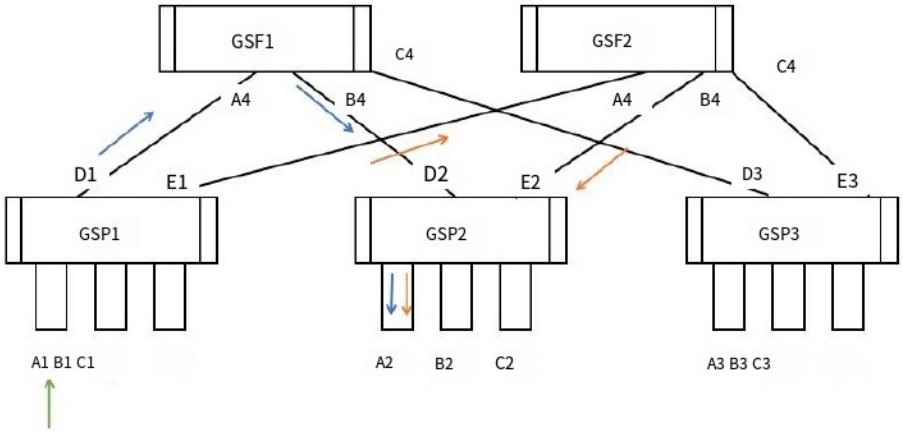

Figure 3-7 Schematic diagram of the traffic forwarding model

The decision-making process can be referred to the following figure: Regarding the path selection of PKTC, the congestion Level is selected first, and the set of exits with the lowest Level layer is chosen. Then, an exit is randomly selected from these exit sets to prevent the synchronization effect under multi-path selection.

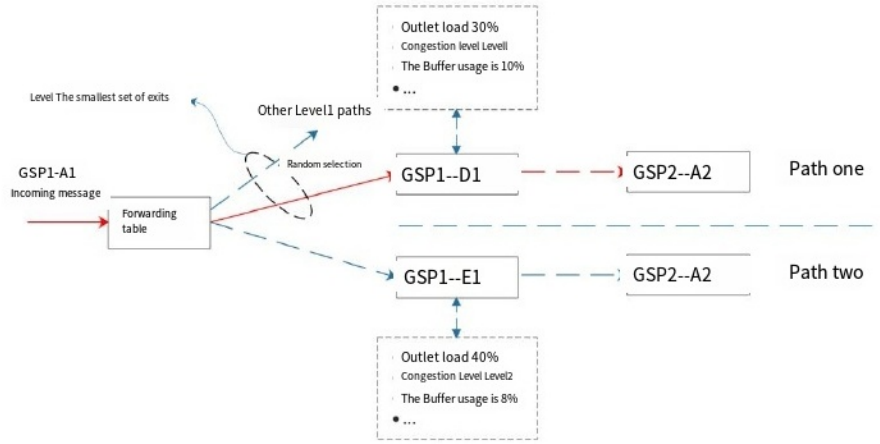

Figure 3-8 Dynamic Load Balancing Decision Process

3.2.2 Dynamic path switching technology

When the load of the export undergoes dynamic changes, each PKTC can re-select the path according to the routing algorithm to ensure the global load balancing effect. During the switching process, it is necessary to ensure the consistency of each PKTC in path selection; otherwise, the degree of disorder will increase and the sorting pressure will increase. The path selection is still carried out by first selecting the Level layer and then randomly choosing the exit.

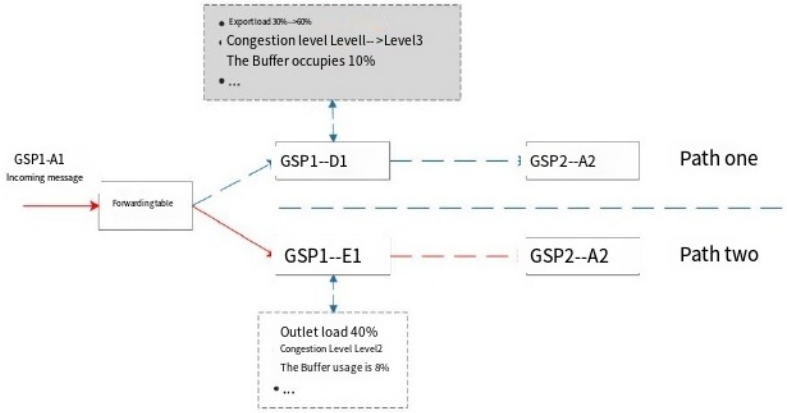

Figure 3-9 Dynamic Path Switching Mechanism

3.2.3 Traffic Sorting Mechanism

After load balancing and dynamic path switching of the traffic, multiple transmission paths are formed. Due to the certain differences in transmission delay of different paths, reordering processing is required when the traffic of different paths reaches the node where the final exit is located.

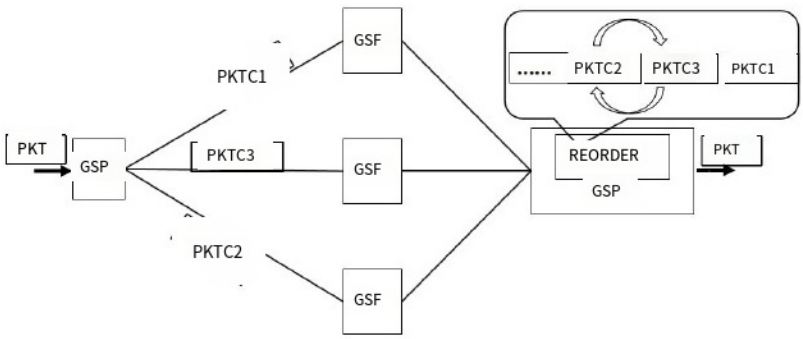

Figure 3-10 Traffic Sorting Mechanism at the Destination End

3.3 DGSQ scheduling technology based on PKTC

In network transmission, the phenomenon of multiple ports hitting one port at certain moments often occurs. If this phenomenon is short-lived, it can be absorbed through a certain Buffer at the exit. If the duration is too long and the total traffic of multiple entrances is much greater than the line speed bandwidth of the exit, in order to avoid packet loss, the exit device needs to enable the backpressure mechanism to protect the traffic. Once the backpressure occurs, the forwarding performance of the network will drop significantly.

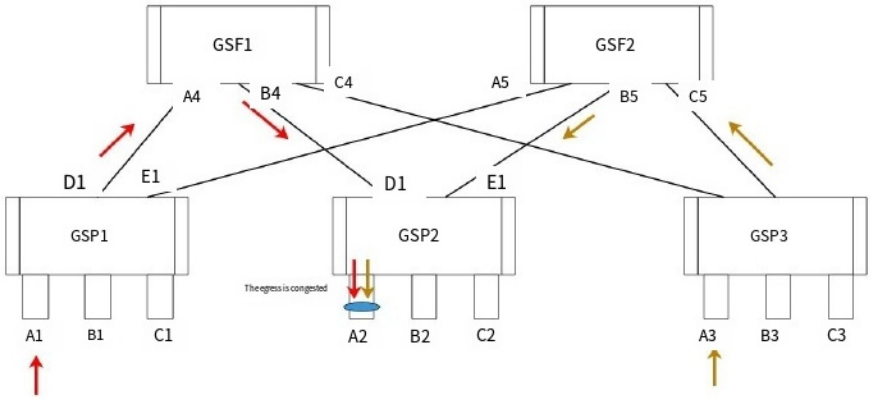

Figure 3-11 Network Incast Traffic Generation Scenario

As shown in the above figure, the A1 port of GSP1 and the A3 port of GSP3 simultaneously send traffic to the A2 port of GSP2, and the total traffic exceeds the exit bandwidth of the A2 port, causing congestion in the exit queue of the A2 port. In this case, it cannot be avoided merely through load balancing. Global control is required to ensure that the traffic sent to the A2 port does not exceed its exit bandwidth. Therefore, the introduction of global-based forwarding technology and DGSQ-based scheduling technology is necessary to achieve global traffic scheduling control.

3.3.1 Forwarding technology based on the global view

In the traditional data center Ethernet forwarding model, the forwarding table takes the information carried by the message as the main body and edits the message header information according to the exit of the next-hop connection, as shown in the following figure:

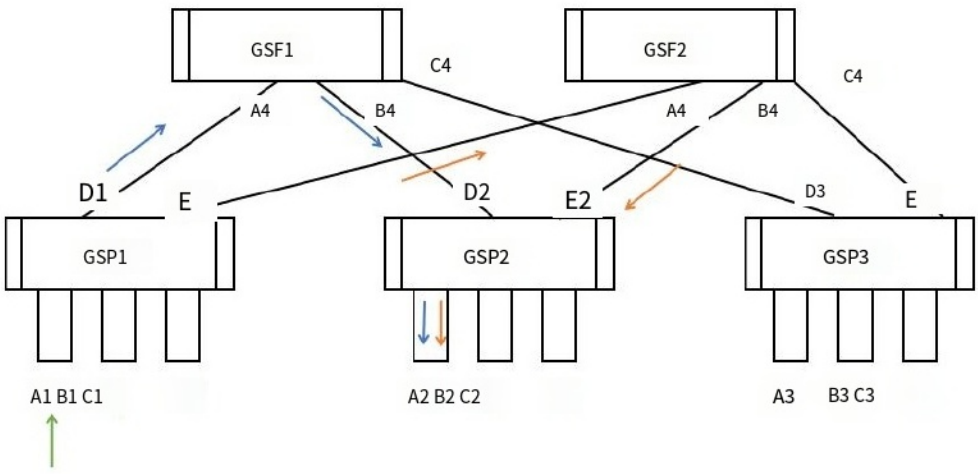

Figure 3-12 Schematic diagram of the traffic forwarding model

The packets coming from any port of GSP1 are forwarded to the GSP2-A2 port. It is necessary to form a forwarding table and the corresponding exit information on GSP1. These information are formed by this device based on its own and the status of adjacent devices, but

The status of network devices on the subsequent path is neither sensed nor controlled. This approach cannot construct a congestion-free fully scheduled Ethernet. It is necessary to construct a forwarding technology based on a global vision, which supports indicating the final destination in the forwarding table of the access switch, and through end-to-end path scheduling and comprehensive authorization mechanisms, dynamically form load sharing information and form the next hop exit information.

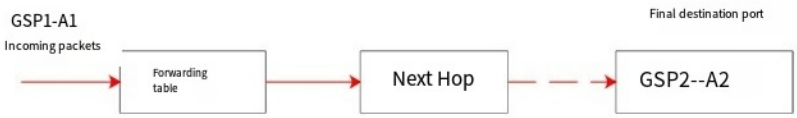

Figure 3-13 Routing Mechanism Based on Global View

3.3.2 Scheduling Technology Based on DGSQ

The global scheduling technology based on DGSQ is shown as follows. Virtual queues of all device exits in the network are established on the GSP to simulate the traffic scheduling from this device to the corresponding port. The scheduling bandwidth of the DGSQ of this device depends on the authorization request and response mechanism, and the end-to-end authorization of the entire network is uniformly carried out by the final device exit and the passing devices. Due to the difference in traffic pressure of intermediate nodes, the GSP no longer selects the path through the ECMP path authorization weight to the final destination port, but needs to perform traffic scheduling on different paths based on the granted weights. In this way, it can be ensured that the traffic to any port of the entire network will not only not exceed the load capacity of the port, but also will not exceed the forwarding capacity of any intermediate network node. This can reduce the probability of Incast traffic generation in the network and the occurrence of internal backpressure mechanisms of the entire network.

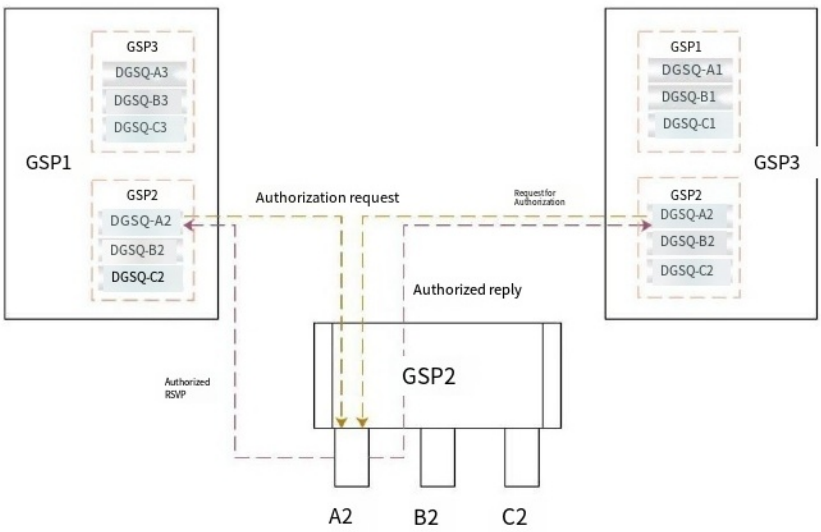

Figure 3-14 Scheduling Technology Based on DGSQ

3.4 Fine backpressure mechanism

The load balancing technology based on PKTC and the global scheduling technology of DGSQ can well regulate and distribute traffic in a stable state. However, in abnormal scenarios such as microbursts and link failures, the network will still experience congestion in a short period of time. At this time, the backpressure mechanism still needs to be relied on to suppress the traffic transmission at the source end. Traditional PFC or FC are point-to-point local backpressure technologies. Once triggered, they spread throughout the network, causing problems such as HoL and network storms. In the fully scheduled Ethernet technology, a sophisticated backpressure mechanism is required to safeguard the network's defense line and stabilize the network load at the minimum backpressure cost.

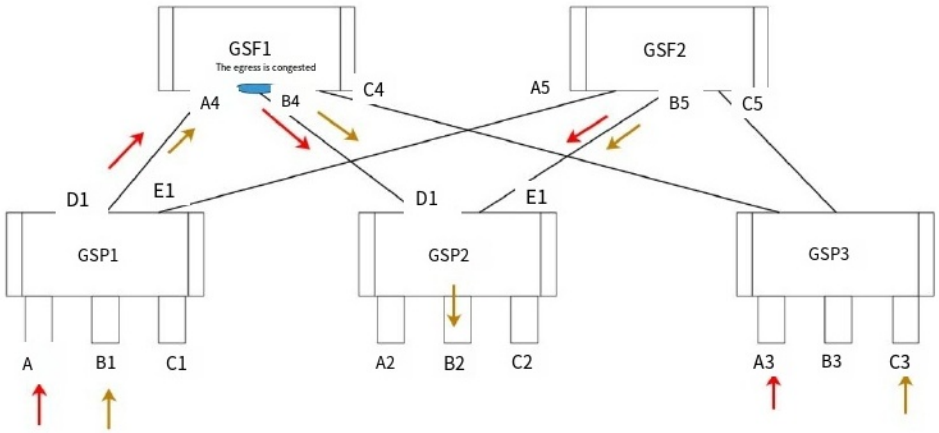

Figure 3-15 Schematic diagram of network congestion scenarios

For example, as shown in the above figure, if congestion occurs at the B4 outlet of GSF1, it will reduce or even suspend the DGSQ scheduling authorization for this port. If there are other path options, it will trigger a switch to other links in a dynamic load balancing manner. If there is only this link in the current network, or other links are also about to be congested, it does not constitute a switching condition. At this time, the backpressure mechanism needs to be activated. To sacrifice the minimum traffic to ensure the stability of the entire network traffic, the scope of backpressure needs to be controlled precisely enough. For example, only the traffic to GSP2 is suppressed, while the traffic to other devices is not affected. A more refined control strategy is that the traffic to GSP2 through GSF1-B4 is suppressed, while the traffic to other devices is not affected. The final degree of refinement will be stipulated in the subsequent GSE standards.

3.5 No perception self-healing mechanism

In the fully scheduled Ethernet architecture, a virtual queue path from the ingress port to the egress port is constructed through the fully scheduling technology.

For the forwarding services of the inbound port, there is no need to perceive the hop path of the outbound port. Only the outbound port needs to be identified. It is insensitive to the Fabric network composed of GSF. The reachability and switching of the path are guaranteed by the load balancing technology of the Fabric network.

GSF adopts the hierarchical load balancing technology based on PKTC. When a link or a GSF in the Fabric network fails, the connected device nodes can perceive the link state change in real time and automatically remove the corresponding link from the load balancing alternative list, and reclaim the scheduling authorization of DGSQ involving this path, thereby allowing PKTC to be distributed to other available links. When the device or link failure is restored, the connected device nodes can also perceive the link state change in real time and complete self-healing. The load balancing technology based on PKTC can maintain stable balance during the above link switching process and will not be affected by the hash result or the small number of links like flow-based load balancing, avoiding the situation of sudden load superposition on a certain link.

3.6 Low-latency forwarding technology

The forwarding aspect mainly reduces the time-delay of the forwarding path within the device by means such as simplification, parallelization, and bypassing the forwarding process. With the continuous increase of port rate, the challenge of high-speed signal integrity becomes greater and greater, and more powerful FEC (forward error correction) algorithms need to be continuously introduced. The stronger the FEC is, the higher the encoding and decoding complexity is, and the greater the time-delay it adds. The time-delay occupied by FEC for rates above 100G has reached about 20% of the overall forwarding time-delay.

The process of FEC can be divided into error-checking logic and error-correction logic. In low-speed FEC processing, the above processes are often not differentiated. However, as the rate increases and the detection and error-correction logic become more complex, the differentiated processing of subdivision becomes increasingly meaningful. The separation technology of error-checking and error-correction can verify in advance whether there are error codes within the data block. In the case of no errors, the FEC decoding process can be bypassed to eliminate the FEC frame receiving and decoding delay in the error-free scenario, reduce the interface delay in the error-free situation, and eliminate the delay drawback of high-gain FEC codewords. Error correction processing is carried out only when there are errors. Because the probability of error codes is, after all, much smaller than that of no error codes, this method can optimize the average forwarding delay of the port. The FlexFEC technology can automatically select the appropriate FEC error-correction algorithm based on the error code rate status of the link to provide low latency while maintaining reliability.

3.7 Full-scheduling Ethernet operating system

The GSOS of the fully scheduled Ethernet takes into account the advantages of both distributed NOS and centralized SDN controllers, and is divided into two major parts: the fully scheduled controller and the NOS on the device side. Meanwhile, an in-band management path within the band is adopted.

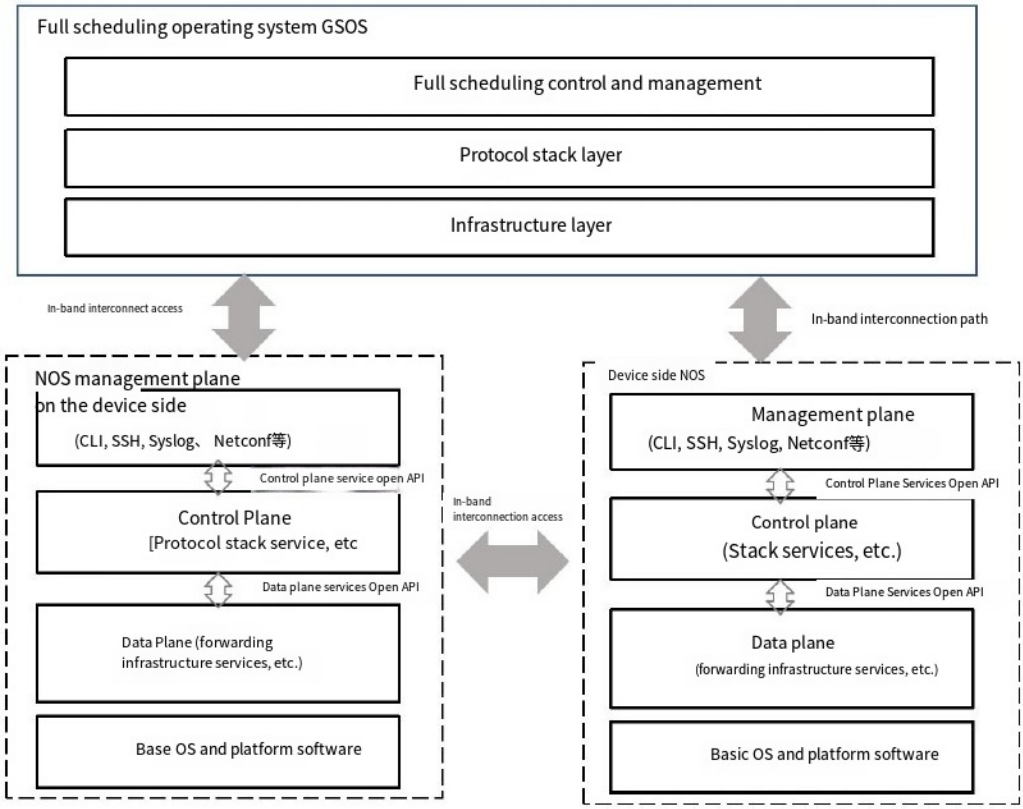

Figure 3-16 The architecture of the fully scheduled Ethernet operating system

- NOS on the device side: Box-type devices of GSP and GSF support the independent deployment of NOS and construct a distributed network operating system. Each GSP and GSF has an independent control plane and management plane, which can run network functions belonging to the device itself, improving system reliability and reducing deployment difficulty. The distributed NOS can limit single-point device failures to the local scope and avoid affecting the entire network. To provide open services and support full-scheduling Ethernet features, NOS also decouples the traditional integrated network function services into control plane services and data plane services, and opens service interfaces. The openness of data plane services provides greater flexibility for the deployment of full-scheduling Ethernet. For example, it can establish a full-scheduling DGSQ system in coordination with the controller, and select the appropriate distributed or centralized discovery and synchronization protocol based on the network scale or software implementation to establish a Fabric interconnection network, etc.

- Full-scheduling GSOS: The centralized GSOS provides better global network information, simplifying the establishment and maintenance of the DGSQ system based on global port information. At the same time, GSOS is also the brain of the entire network operation and maintenance monitoring, which can collaborate with devices to achieve the recording and presentation of real-time paths and history to support network operation and maintenance.

- NOS Control and Management Pathway: Thanks to the compatibility principle of the fully scheduled Ethernet architecture, the GSF nodes of the network can also support the Ethernet message exchange feature. This enables the unification of the management and control plane into the data forwarding plane, forming an In-band interconnection path, and in the Fabric interconnected data forwarding plane

An internal high-priority channel is reserved in the middle to ensure the priority of the control and management passage. The fully scheduled Ethernet no longer adopts the out-of-band control and management passage but is unified into the in-band passage, which is convenient for operation and maintenance management and avoids maintaining two sets of physical networks.

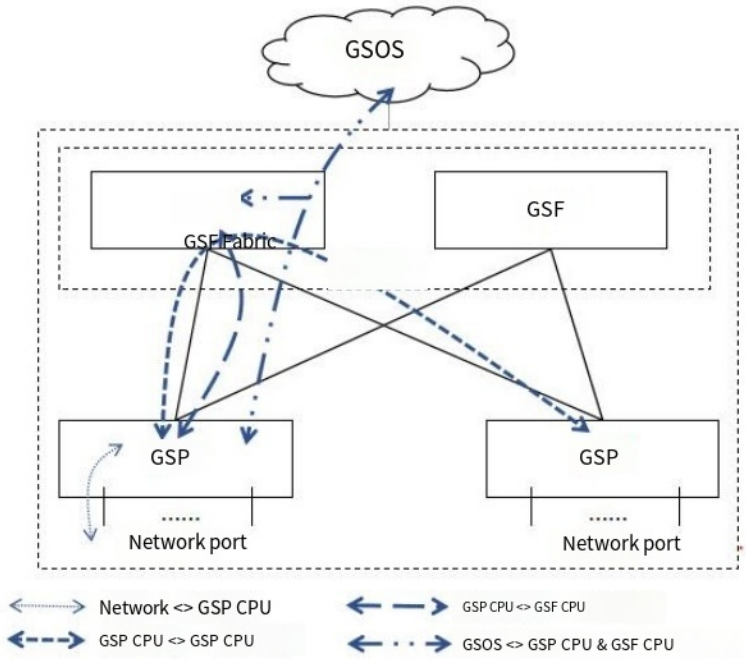

Figure 3-17: Control and Management Pathways of Intra-Band Modes

4. Networking Application Outlook

GSE is targeted at high-performance network demand business scenarios such as lossless, high bandwidth, and ultra-low latency. It is compatible with the Ethernet ecosystem. By adopting technologies such as the full scheduling and forwarding mechanism, load balancing technology based on PKTC, full scheduling technology based on DGSQ, fine backpressure mechanism, perception-free self-healing mechanism, centralized management, and distributed control, it realizes a new type of intelligent computing center network with low latency, no blocking, and high bandwidth. When this technical architecture is implemented, there are two ways. One is to run this architecture only on the network side, and the other is to run it end-to-end.

Run this architecture only on the network side

GSE itself can support the network solution without awareness on the network card side. If the network card side is capable of participating in collaboration, it can provide the end-to-end full scheduling feature more precisely. The DGSQ queue of the GSP device can feed back its status to the network card side for the network card or the service to perceive the network status, thereby achieving better end-network collaboration. For example, when a certain DGSQ queue of the GSP reaches a certain waterline, it indicates that there is congestion in the network traffic from the corresponding computing node to the peer computing node. At this time, the GSP can feed back this information through the backpressure mechanism.

The corresponding network card, network or business side can appropriately adjust the packet sending rate to this peer computing node based on this information, avoiding possible congestion aggravation or packet loss from the source.

The architecture is run end-to-end.

The functions of GSE are redistributed in network establishment. The original mechanism remains unchanged. The network card or the network card module of the GPU implements authorization distribution and backpressure response. The switch still integrates load balancing routing based on PKTC, traffic sequencing, fine backpressure information generation, and the most basic forwarding control based on the global scope. In this way, while the original GSE networking model and functions remain unchanged, taking advantage of the network card being closest to the service side, traffic can be scheduled from the source of the service.

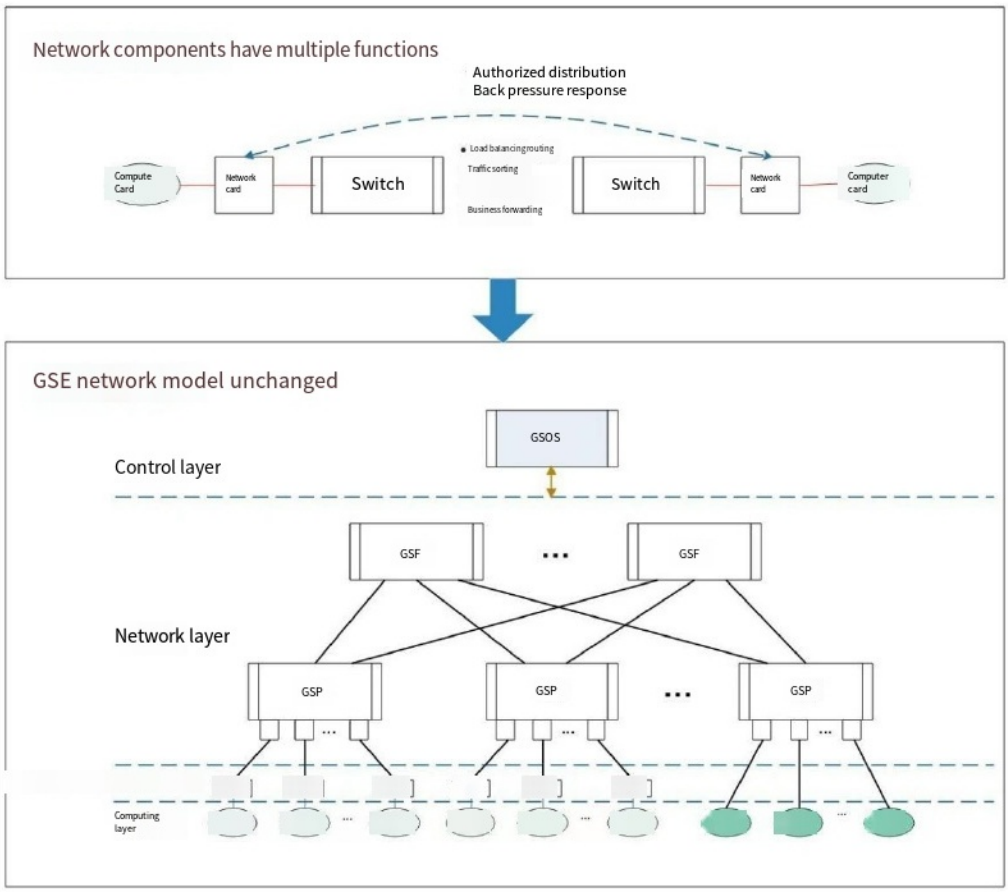

Figure 4-1 The subsequent evolution direction of GSE technology