Fusion Gateway QoS

White Paper

Copyright © 2026 TP-Link Systems Inc. All rights reserved.

No part of this document may be reproduced, copied, or transmitted in any form or by any means without the prior written permission of the company.

Contents

1 Overview 4

1.1 Introduction 4

1.2 Background 4

1.3 Fusion Gateway QoS 4

1.4 Evolution of Omada Gateway QoS 5

1.4.1 Historical Background 5

1.4.2 Key Drivers of the Evolution 5

1.4.3 QoS Technology Evolution Path 6

1.4.4 Key Improvements of Fusion Gateway QoS 7

1.4.5 Future Trends 8

1.5 Purpose of This Document 8

2 Key Technologies 9

2.1 Traffic Identification 9

2.1.1 Basic Concepts 9

2.1.2 Technical Advantages 9

2.1.3 Market Demand 10

2.2 Bandwidth Control 10

2.2.1 Basic Concepts 10

2.2.2 Technical Advantages 10

2.2.3 Market Demand 10

2.3 Priority Scheduling 11

2.3.1 Basic Concepts 11

2.3.2 Technical Advantages 11

2.3.3 Market Demand 11

3 Key Technical Principles 11

3.1 Traffic Identification 11

3.1.1 Protocol 12

3.1.2 Principle 15

3.1.3 Framework 15

3.1.4 Application Scenario 15

3.2 Bandwidth Control 16

3.2.1 Protocol 16
3.2.2 Principle 16
3.2.3 Framework 18
3.2.4 Application Scenario 18
3.3 Priority Scheduling 19
3.3.1 Protocol 19
3.3.2 Principle 20
3.3.3 Framework 22
3.3.4 Application Scenario 23
4 Deployment Scenarios and Solutions 24
4.1 Chain Retail and Restaurant Scenarios 27
4.1.1 Scenario Description 27
4.1.2 Requirements and Pain Points 27
4.1.3 Network Topology 28
4.1.4 Solution 28
4.2 Small and Midsize Office Scenarios 30
4.2.1 Scenario Description 30
4.2.2 Requirements and Pain Points 30
4.2.3 Network Topology 31
4.2.4 Solution 31
4.3 Prosumer Scenarios 33
4.3.1 Scenario Description 33
4.3.2 Requirements and Pain Points 33
4.3.3 Network Topology 34
4.3.4 Solution 34
5 Representative Models 36
6 Future Outlook 37
6.1 Technology Design and Limitations 37
6.2 Future Directions 38
7 Appendix 39
7.1 Glossary 39
7.2 References 40

1 Overview

1.1 Introduction

With the rapid growth of real-time audio/video communications, online collaboration, cloud applications, and other services in enterprise networks, limited bandwidth resources are facing increasing allocation challenges. Quality of Service (QoS) technologies provide end-to-end service quality assurance for critical services through traffic classification, bandwidth control, and priority scheduling on network devices, ensuring key services can still obtain the required bandwidth and low latency during network congestion.

1.2 Background

In traditional enterprise networks, bandwidth resources are typically allocated in a fair-competition or best-effort manner. When multiple traffic flows share the same uplink, large-traffic transmissions (e.g., file downloads and system backups) can quickly fill the gateway's egress queue, causing packet loss and high latency in real-time applications such as audio/video communications. Meanwhile, enterprise networks are carrying an increasingly diverse set of services, and different services have significantly different network quality requirements: voice and video conferencing require low latency and low jitter, while general web browsing and file transfers tolerate some delay.

The Omada platform, as an integrated networking solution for SMBs and commercial users, is expected to provide flexible and efficient QoS capabilities on gateways. The existing QoS and bandwidth control features fall short in terms of configurations and functions, failing to meet the needs for fine-grained traffic management and differentiated service assurance.

1.3 Fusion Gateway QoS

Fusion Gateway QoS is a newly designed QoS solution for Fusion series gateways. It is intended to integrate the existing QoS and Bandwidth Control modules and achieve unified traffic management. Fusion Gateway QoS is built on Linux's original Traffic Control (TC) framework and, together with a multi-dimensional traffic identification engine, achieves deep integration of three core capabilities: "traffic identification + bandwidth control + priority scheduling."

Fusion Gateway QoS supports traffic classification based on multiple dimensions, including device, network, IP address, DSCP markings, port, application, and domain name. It offers two rate-limiting modes: Shared and Individual. Using a weighted round-robin (DRR) queue scheduling strategy, it ensures high-priority traffic is forwarded first when the network is congested.

1.4 Evolution of Omada Gateway QoS

1.4.1 Historical Background

On ER series gateways, QoS configuration consists of five sections: **Gateway QoS Service, Bandwidth Control, Class Rule, VoIP Prioritization, and Tag Outbound Traffic**. The overall design centers on "bandwidth control + class-based traffic identification," with the following key characteristics:

- **Gateway QoS Service:** Uses custom service objects (protocol + source/destination port ranges + ICMP type/code + protocol number) as rule match conditions. Predefined entries cannot be edited or deleted, and additional entries can be created.
- **Bandwidth Control:** Allocates upstream/downstream bandwidth on the WAN by percentage across four levels (Class 1/2/3/Others). Supports ingress/egress/bidirectional direction control, UDP bandwidth control (including percentage), and outbound TCP ACK prioritization.
- **Class Rule:** Rules are matched from top to bottom by sequence number. Once a rule is matched, traffic is assigned to the corresponding class and no further matching is performed. Match dimensions include IP version, local/remote addresses (e.g., IPv4/IPv6 address groups), DSCP, and service entries.
- **VoIP Prioritization:** Provides a priority switch for traffic to SIP UDP ports.
- **Tag Outbound Traffic:** Allows setting DSCP or Precedence values for traffic in each class.

Overall, QoS on ER series gateways is **organized based on classes**. Define the classes first, then classify traffic into those classes, and finally apply bandwidth quotas and outbound marking on a per-class basis. This mechanism was suitable when services were mainly defined by "subnet + port." However, with today's application-driven traffic and increasingly dynamic networks, its static, coarse-grained, fixed-class structure shows clear limitations: dynamic ports and encrypted traffic weaken port-based identification; fixed percentages do not adapt well to bandwidth fluctuations; priority scheduling capabilities are limited; and it is difficult to meet modern applications' requirements for differentiated QoS.

1.4.2 Key Drivers of the Evolution

Fusion Gateway QoS was proposed in response to fundamental changes in enterprise traffic structures and management requirements:

1. **Application-driven traffic:** Application-level traffic such as video conferencing, cloud gaming, and NAS transcoding has become dominant, requiring identification and scheduling to move up to the application layer.
2. **Priority assurance as a necessity:** Non-critical traffic must be controlled, while

critical services must have their forwarding order guaranteed. This requires tight integration of identification, rate limiting, and scheduling.

3. **Higher deployment efficiency requirements:** In multi-WAN and multi-site scenarios, step-by-step configuration is costly and error-prone, calling for a more unified deployment approach.

4. **Increasingly dynamic network conditions:** Bandwidth fluctuations and time-dependent service differences require more flexible scheduling and policy orchestration.

1.4.3 QoS Technology Evolution Path

Aspect	ER Gateway QoS	Fusion Gateway QoS	Upgrade
Architecture	A class-centered bandwidth allocation system	A unified policy system integrating identification, control, and scheduling	Upgrades from a rate-limiting tool to a complete QoS system
Identification level	Service entries (port-level) + address groups + DSCP	Application/application category + device/network/IP/port/domain/protocol	Upgrades from Layer 4 port-based identification to application-level identification
Policy mechanism	Bandwidth limit only + VoIP exception	Prioritize/Limit/Prioritize and Limit	Completes the priority scheduling capability
Rate limit granularity	Class-level percentage + UDP ratio	Shared/Individual + Speedburst + independent uplink/downlink limits	Moves from class-level to per-client quantification
UDP control	UDP bandwidth ratio (class-level)	WAN-level UDP Bandwidth Quota (UDP ratio during congestion) + rule-level limit/prioritize and limit (fixed UDP cap supported)	Evolves from "class-level UDP ratio" to a combined mechanism of "WAN-level quota + rule-level fine-grained limiting"
QoS marking	DSCP/Precedence marking per class	Match DSCP + per-rule outbound DSCP remarking	More precise marking and end-to-end consistency across the network
Scope	Single WAN, class-based	Single WAN or all WANs	Supports unified orchestration for multi-WAN
Time orchestration	None	Schedule effective time periods	Enables time-based differentiated policies

Configuration	Predefined services/address groups + multi-page step-by-step configuration	One-click enablement per WAN + rule-based orchestration	Simplifies the configuration workflow

1.4.4 Key Improvements of Fusion Gateway QoS

1. **From a "class-centric" model to a unified policy system:** Traditional ER gateway QoS is organized around fixed classes. Identification, rate limiting, and marking are all performed on a per-class basis, resulting in a relatively fixed structure and closed loop. Fusion Gateway QoS uses flexible rules as the basic unit and integrates identification, control, and scheduling into a unified policy system. It can apply prioritization, rate limiting, or both to different services within the same network, significantly improving the policy space and flexibility.
2. **From port-level identification to application-level identification:** Traditional ER gateway QoS relies on service entries (e.g., ports/protocols) + address groups + DSCP for identification, and cannot effectively differentiate traffic with dynamic ports or encrypted traffic. Fusion Gateway QoS supports application- and application category-level identification, covering Layers 2 to 7 and better matching modern service patterns.
3. **Completed priority scheduling capability:** Traditional ER gateway QoS lacks systematic control over forwarding order for critical services during congestion, with prioritization mainly reflected only through specific toggles such as VoIP. Fusion Gateway QoS ensures prioritized forwarding for critical applications through priority policies, while constraining non-critical traffic through rate-limiting policies to achieve differentiated resource allocation.
4. **More fine-grained and unified rate limiting:** Traditional ER gateway QoS mainly uses class-level bandwidth percentages and UDP ratios, with granularity remaining at the class level. Fusion Gateway QoS supports Shared and Individual rate-limiting modes, enabling rate limiting down to individual clients. With the Speedburst mechanism, it balances throughput and fairness, and together with independent uplink/downlink limits, significantly improves precision.
5. **More precise and end-to-end QoS marking:** Both support outbound marking, but traditional ER gateway QoS applies DSCP/Precedence uniformly per class, resulting in coarse granularity. Fusion Gateway QoS can both classify traffic based on existing DSCP markings and precisely rewrite DSCPs per rule, enabling consistent marking across the entire network and supporting end-to-end quality assurance across devices and subnets.

6. **Improved deployment and O&M efficiency:** Traditional ER gateway QoS requires pre-configuring service entries and address groups, and then performing step-by-step setup across multiple pages. Fusion Gateway QoS provides one-click enablement per WAN and unified, rule-based orchestration with scheduling, significantly reducing configuration and maintenance costs in multi-site and multi-WAN scenarios.
7. **UDP traffic control evolves from "class-level ratio" to a combined mechanism of "WAN-level quota + rule-level limiting":** Traditional ER gateway QoS constrains UDP bandwidth share by class-level percentages during congestion, with control remaining at the traffic-type level. Fusion Gateway QoS limits the overall UDP bandwidth ratio on a congested WAN through a WAN-level UDP Bandwidth Quota, and also applies fixed caps to specific UDP traffic through rules using Limit/Prioritize and Limit. These mechanisms operate at different granularities—"WAN-level ratio" and "rule-level fixed limit"—and work together to meet UDP traffic control requirements at different levels.

1.4.5 Future Trends

Given Fusion Gateway QoS's comprehensive advantages in architecture, identification, scheduling, rate limiting, and marking, the QoS capabilities of the ER series gateways will be upgraded in phases according to model planning. The specific supported scope and introduction timeline are subject to each model's release notes. Looking ahead, this system will continue to evolve in areas such as rule library expansion, adaptive scheduling, and more fine-grained real-time service identification. These are industry-level exploration directions and do not constitute product feature commitments. The actual roadmap is subject to the official release notes of each version.

1.5 Purpose of This Document

This document aims to:

- Offer a detailed introduction to the key technical concepts and fundamental principles of Fusion Gateway QoS.
- Explain the technical architecture of the three core modules: traffic identification, bandwidth control, and priority scheduling.
- Provide references for typical application scenarios and network deployment solutions.
- Help users fully understand and effectively use Fusion Gateway QoS to optimize network experience.

2 Key Technologies

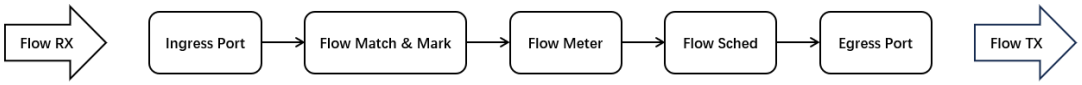


Figure 1-1 Traffic Processing Flow

Fusion Gateway QoS is designed based on three core modules: traffic identification, bandwidth control, and priority scheduling. Traffic identification classifies and marks network packets and serves as the prerequisite for Fusion Gateway QoS policy enforcement. Bandwidth control regulates the rate of different traffic types to prevent any single flow from consuming excessive bandwidth resources. Priority scheduling determines packet transmission order based on priority during congestion, ensuring critical services are forwarded first. This chapter introduces the concepts, advantages, and market demand for these three key technologies.

2.1 Traffic Identification

2.1.1 Basic Concepts

Traffic identification is a foundational capability of a QoS system. It analyzes packet header information, payload content, and behavioral characteristics to classify different network traffic into predefined categories, providing the basis for subsequent bandwidth control and priority scheduling. The accuracy of traffic identification directly affects the effectiveness of QoS policy enforcement.

Fusion Gateway QoS supports multi-dimensional traffic identification, including but not limited to device, network, IP address/subnet, DSCP markings, TCP/UDP port numbers, application, and domain. Administrators can define Fusion Gateway QoS rules flexibly using these match dimensions to meet actual business requirements.

2.1.2 Technical Advantages

Precise matching from multiple dimensions: Supports full-stack traffic identification from Layer 2 to Layer 7, covering IP addresses, ports, applications, domains, and more, meeting classification needs across different scenarios.

Flexible policy composition: Multiple match conditions can be flexibly used with AND logic, enabling administrators to define differentiated policies for specific users, applications, and destinations.

Efficient kernel-level processing: The Linux kernel netfilter framework allows packet classification directly in the protocol stack forwarding path, delivering high efficiency and low latency.

2.1.3 Market Demand

The growing diversity of enterprise network services increases network administrators' demand for traffic visibility and fine-grained control. Traditional port-based classification can no longer meet the requirements of modern applications. More and more applications use dynamic ports, encrypted transport, and CDN acceleration, which increases both the difficulty and importance of traffic identification. The market urgently needs QoS solutions with deep identification capabilities and flexible policy configuration to address bandwidth management challenges in diverse scenarios such as hybrid work, video conferencing, and cloud applications.

2.2 Bandwidth Control

2.2.1 Basic Concepts

Bandwidth Control refers to limiting the transmission rate of specific types of network traffic to ensure that any single type of traffic does not consume excessive bandwidth resources and affect other services. The core mechanism is the Token Bucket Algorithm, which precisely regulates packet transmission rates by controlling the token generation rate and bucket size.

Fusion Gateway QoS supports two rate-limiting modes: Shared and Individual. In Shared mode, all traffic of the same type shares a single bandwidth limit. In Individual mode, the system maintains bandwidth quotas separately based on "source IP + IP protocol (IPv4/IPv6)." That is, for the same source IP, IPv4 and IPv6 traffic each has an independent quota. This ensures each client (based on the source IP) is controlled independently per IP protocol and prevents a single client from monopolizing bandwidth. In addition, Fusion Gateway QoS supports Speedburst, which allows traffic to transmit at a rate exceeding the configured bandwidth limit for a short period, improving web browsing and file downloads.

2.2.2 Technical Advantages

Precise rate control: Uses the Token Bucket Algorithm to control the average packet transmission rate precisely, while adjusting burst tolerance through the bucket size.

Flexible rate-limiting modes: Shared mode is suitable for overall traffic control, while Individual mode ensures fairness for each client, meeting different scenario requirements.

Burst optimization: Speedburst allows short-term bursts while still limiting the long-term average rate, balancing user experience and control.

2.2.3 Market Demand

In scenarios where multiple users share the same internet uplink (e.g., offices, campus networks, and hotel Wi-Fi), huge downloads or video streaming by a small number of users

often consume bandwidth and degrade the overall network experience. Bandwidth control, through reasonable rate-limiting policies, maximizes network utilization while maintaining fairness. Especially for SMBs, where internet bandwidth is limited and must be managed carefully, bandwidth control has become an essential capability for gateway devices.

2.3 Priority Scheduling

2.3.1 Basic Concepts

Priority Scheduling is one of the core QoS capabilities. It determines packet transmission order based on packet priority during network congestion, ensuring higher-priority traffic is forwarded first. Unlike bandwidth control, priority scheduling focuses more on "assuring" service quality—prioritizing critical services when network resources are constrained.

Fusion Gateway QoS realizes priority scheduling based on the Linux kernel TC (Traffic Control) framework. TC is the traffic control infrastructure provided by the Linux kernel. By applying queueing disciplines (qdisc) on the egress queues of network interfaces, it supports classification, scheduling, and shaping of outbound packets.

2.3.2 Technical Advantages

End-to-end quality assurance: Provides weighted priority queue scheduling for latency-sensitive services such as voice and video transmission, maintaining a smooth experience under WAN congestion.

Mature framework support: Built on the native Linux kernel TC framework, which is mature, widely adopted, and backed by an active community, making it easier to extend and maintain.

2.3.3 Market Demand

In enterprise networks, different services have significantly different requirements for network quality. VoIP and video conferencing require low latency and low jitter, while file synchronization and backups can tolerate delay. A simple best-effort forwarding model can no longer meet the needs of enterprises running multiple services in parallel. Priority scheduling assigns different forwarding priorities to service classes, ensuring critical services maintain a stable user experience during congestion, and has become an essential feature of enterprise network devices.

3 Key Technical Principles

3.1 Traffic Identification

The core of traffic identification lies in parsing and matching protocol headers at different layers of network packets. In the TCP/IP stack, each packet carries abundant identifying information from the data link layer to the application layer. Layer 2 provides MAC addresses

and device information; Layer 3 provides source/destination IP addresses and DSCP service-class markings; Layer 4 provides TCP/UDP port numbers and protocol types; and at Layer 7, DPI (Deep Packet Inspection) can further identify specific application protocols.

3.1.1 Protocol

Fusion Gateway QoS leverages the Linux kernel netfilter framework and its match extension modules to apply classification rules at multiple layers of the protocol stack, providing full-stack identification from physical devices to the application layer.

Fusion Gateway QoS matches packets based on the following fields:

L2 fields: Source MAC address (SMAC field shown below)

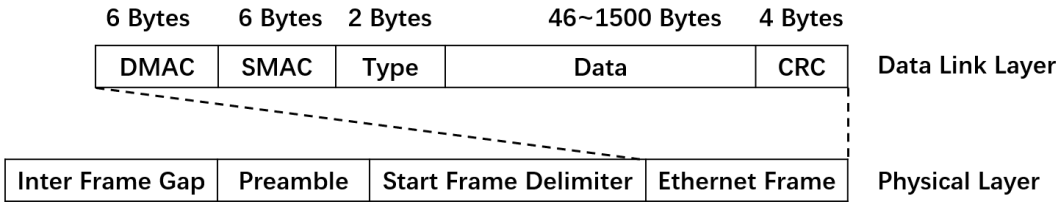


Figure 3-1 Data Link Layer SMAC Field

L3 fields: Source IP address, destination IP address, and Protocol (as shown below):

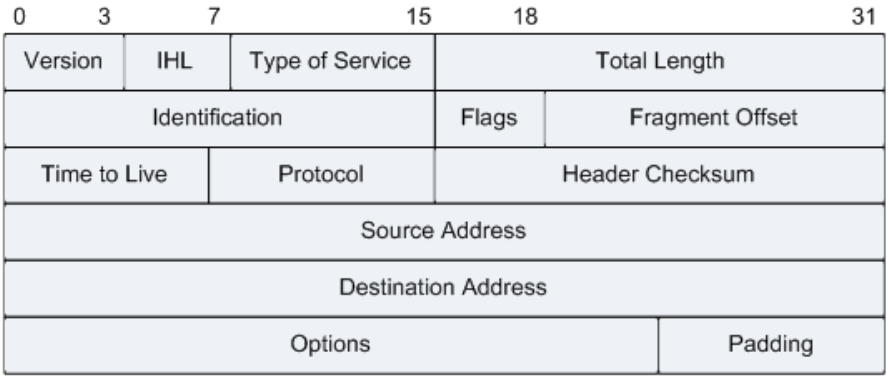


Figure 3-2 IPv4 Packet Format

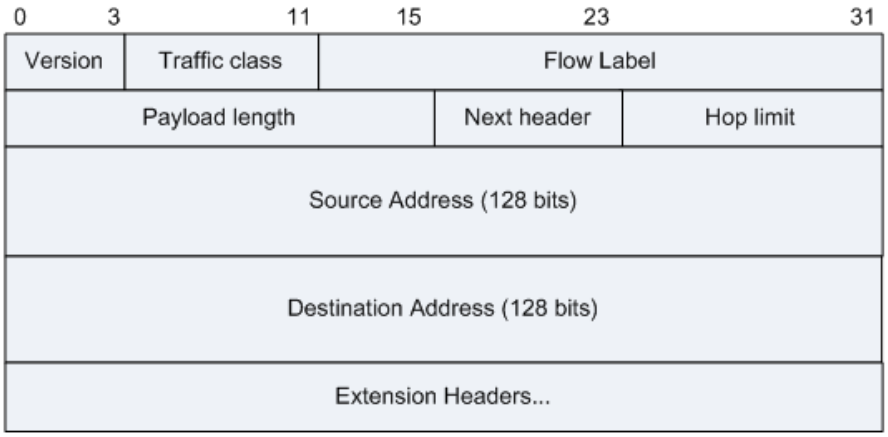


Figure 3-3 IPv6 Packet Format

DSCP field definition:

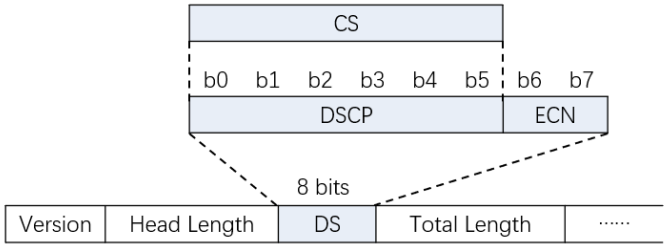


Figure 3-4 IPv4 DS Field

In IPv4 packets, the DS (Differentiated Services) field is 8 bits long. The first 6 bits (b0–b5) form the DSCP (Differentiated Services Code Point), and the last 2 bits (b6–b7) are used for ECN (Explicit Congestion Notification). RFC 2474 defines the DS field, replacing the earlier IPv4 ToS (Type of Service) field D/T/R semantics with a 6-bit DSCP used to select the Differentiated Services per-hop behavior (PHB). Network devices choose the corresponding QoS forwarding behavior based on the DSCP value. RFC 3168 defines ECN, occupying the last two bits (b6–b7) of the DS field, to signal congestion to the sender when network congestion occurs.

To maintain compatibility with the legacy IP Precedence mechanism, RFC 2474 also defines the Class Selector (CS) code points. CS is a subset of DSCP values where the high 3 bits (b0–b2) can be any value and the low 3 bits (b3–b5) are fixed to 000, producing a 6-bit value of the form xxx000 (xxx ranges from 000/001/.../111, corresponding to CS0–CS7). CS values can be used for simple classification forwarding and priority handling.

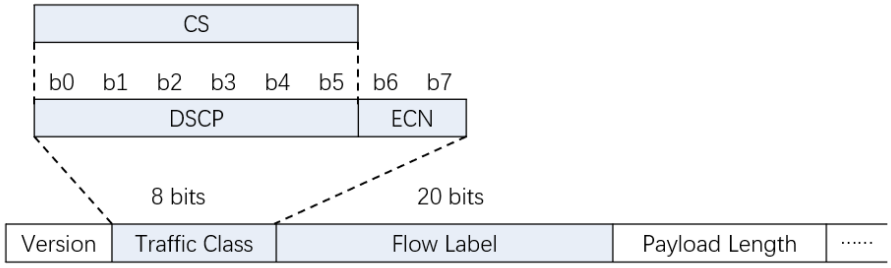


Figure 3-5 IPv6 Traffic Class and Flow Label Fields

In IPv6 packets, the fields relevant to Fusion Gateway QoS mainly include Traffic Class (TC, 8 bits) and Flow Label (FL, 20 bits). The TC field has the same semantics as the IPv4 DS field: the first 6 bits (b0–b5) are DSCP, used to select the differentiated services PHB, and network devices choose the corresponding QoS forwarding behavior based on the DSCP value. The last 2 bits (b6–b7) are ECN, defined by RFC 3168, to signal congestion to the sender during network congestion.

Class Selector (CS) code points also apply in IPv6 with the same meaning as in IPv4: the high 3 bits (b0–b2) can be any value and the low 3 bits (b3–b5) are fixed to 000 (value form xxx000, corresponding to CS0–CS7), enabling simple classification and priority handling. Fusion Gateway QoS supports DSCP values in the range 0–63, and values with the low 3 bits set to 000 can be interpreted under the CS semantics.

The FL field is 20 bits long and is used to identify packets belonging to the same flow. Together with the source and destination addresses, the flow label uniquely identifies a flow.

L4 fields: TCP/UDP port numbers (port fields shown below)

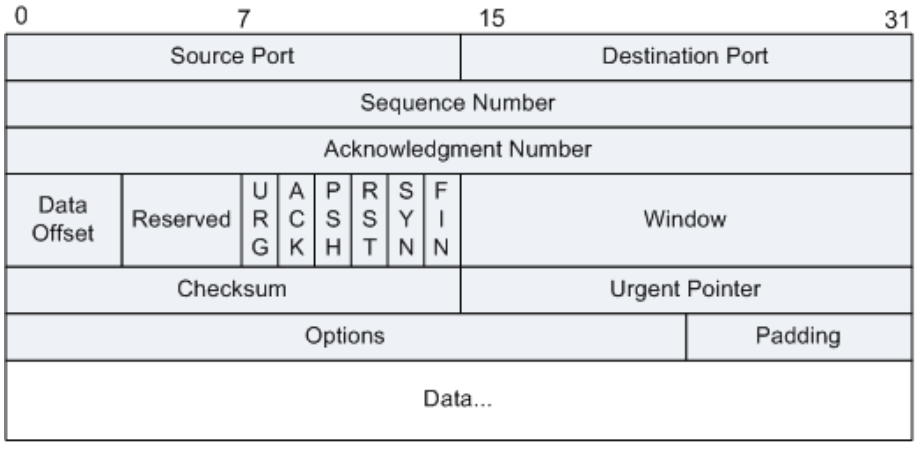


Figure 3-6 TCP/UDP Port Fields in IPv4 Packets

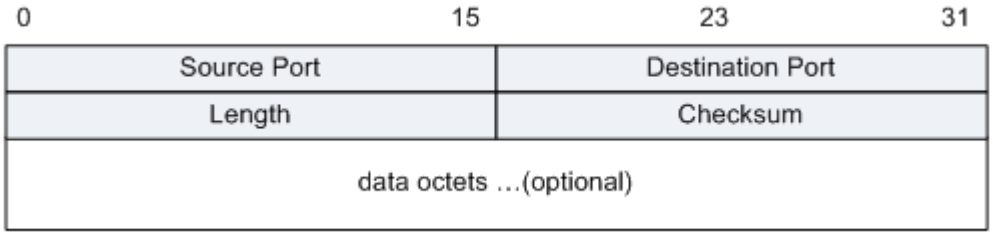


Figure 3-7 TCP/UDP Port Fields in IPv6 Packets

L7 fields: App ID identified via DPI. This requires the DPI module to identify the flow and write the result into the additional information of conntrack.

3.1.2 Principle

Traffic identification and classification are implemented through various iptables' match modules. Matching occurs when uplink and downlink traffic traverses the FORWARD chain. After a packet is identified, it is assigned a classification mark based on the configured policy:

- If the policy is Prioritize or Prioritize and Limit, the packet is marked as High Priority.
- If the policy is Limit, the packet is marked as Low Priority (or left unmarked).

Rule enforcement mechanism: Fusion Gateway QoS rules are matched top-down in order. The first rule hit takes effect, and subsequent rules are not evaluated. Therefore, important rules should be placed at the top.

3.1.3 Framework

Architecturally, the traffic identification module sits at the very front of the data plane of the Fusion Gateway QoS system, acting as an ingress classifier. It interacts with both the upper-layer configuration management system and the lower-layer scheduling system through standard interfaces:

- **Configuration interface:** Receives Fusion Gateway QoS rule configurations from the management platform and translates user policies into kernel match rules.
- **Classification output:** Passes classification results to the bandwidth control module and the scheduling module via packet marks.
- **Extensibility:** The modular design allows new matching dimensions to be added by introducing the corresponding match extensions, without affecting existing rules.

3.1.4 Application Scenario

Traffic identification capability directly determines the granularity and effectiveness of Fusion Gateway QoS policies. In real deployments, common classification scenarios include: classifying by device (e.g., guaranteeing bandwidth for video conferencing endpoints in a meeting room), by application (e.g., prioritizing Teams/Zoom traffic), by domain name (e.g.,

accelerating access to specific business systems), and by network (e.g., distinguishing guest networks from office networks). Multi-dimensional classification enables administrators to precisely control which traffic should be prioritized and which should be restricted, thereby achieving truly fine-grained network management.

3.2 Bandwidth Control

Bandwidth control allocates bandwidth to different users, applications, or traffic types according to predefined rules. The goal is to prevent some traffic from saturating the link and impacting other services.

3.2.1 Protocol

Bandwidth control does not rely on a single standard protocol. Instead, it is implemented through a combination of mechanisms to address different scenarios. Fusion Gateway QoS provides three bandwidth control features: Shared / Individual Mode, Speedburst, and Global bandwidth limiting.

- **Shared mode limiting:** Traffic of the same type shares a single bandwidth quota.
- **Individual mode limiting:** For traffic of the same type, an independent bandwidth quota is created for each client based on **Source IP (Client IP) + IP protocol (IPv4/IPv6)**. In other words, for the same client, IPv4 traffic and IPv6 traffic each have their own separate quota. Each client exclusively uses its bandwidth quota under the corresponding IP protocol.
- **Speedburst:** When disabled, traffic is strictly limited to the configured (rated) bandwidth. When enabled, traffic may temporarily exceed the configured rate while the token bucket has accumulated unused tokens (i.e., the bucket is not empty). However, the long-term average rate is still controlled within the configured bandwidth limit.
- **Global bandwidth limiting:** Traffic on the same WAN interface shares the total bandwidth limit.

3.2.2 Principle

The core of Shared/Individual mode is implemented using a single token bucket algorithm, applied on the FORWARD chain. The algorithm maintains a logical token bucket and adds tokens to the bucket at a fixed rate (i.e., each token represents a certain amount of transmission allowance). The bucket has a maximum capacity. When a packet arrives, it must consume tokens proportional to the packet size to be forwarded. If there are not enough tokens in the bucket, the portion exceeding the available tokens will be dropped.

The difference is:

- **Shared:** All flows matched by the rule share one token bucket.

- **Individual:** Token buckets are maintained independently per **source IP + IP protocol (IPv4/IPv6)** (i.e., for the same source IP, IPv4 and IPv6 each have a separate bandwidth quota).

Speedburst works as follows: when the token bucket has accumulated tokens that have not yet been consumed, packets can be forwarded at a rate above the configured limit for a short period. After that, the traffic is still constrained by the token generation rate, keeping the long-term average rate stable.

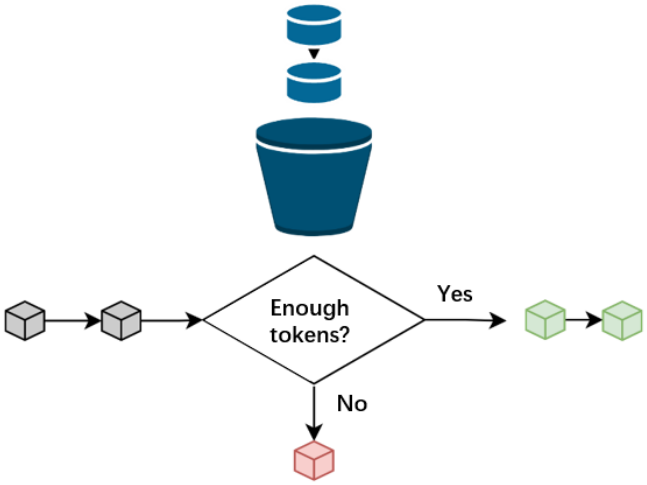


Figure 3-8 Token Bucket Algorithm

As shown in the figure, after packets enter the queue, they can be transmitted only when enough tokens are available in the token bucket (the bucket diagram in the middle). The system generates tokens at a fixed rate. If the tokens currently in the bucket are insufficient to cover the packet size, the portion exceeding the tokens will be dropped. This limits the long-term average sending rate to the configured bandwidth range while allowing a certain level of burst traffic. Both Shared and Individual limiting follow this token bucket algorithm. The only difference is the granularity at which token buckets are maintained (Individual maintains separate buckets per **source IP + IP protocol (IPv4/IPv6)**, while Shared uses one bucket for all flows).

Global bandwidth limiting: The core implementation is based on the **HTB (Hierarchical Token Bucket)** algorithm and is applied at the NIC egress (packet transmission) point. HTB is a hierarchical token bucket model that can limit the total outbound bandwidth of an interface.

Enforcement mechanism: Traffic that matches a rate-limiting policy (**Limit / Prioritize and Limit**) is limited based on its configured bandwidth cap and the Speedburst mode. In Shared

mode, all traffic matched by the rule shares one bandwidth quota. In Individual mode, traffic matched by the same rule maintains separate bandwidth quotas per **source IP + IP protocol (IPv4/IPv6)**. Both modes apply to the same rule and are switched via the **Bandwidth Control Mode** parameter.

3.2.3 Framework

From a system architecture perspective, the bandwidth control module acts as the rate gatekeeper in Fusion Gateway QoS. It sits after the traffic identification module and before the priority scheduling module. It receives classification marks from the traffic identification module and enforces independent rate control on different traffic classes according to the user's configured bandwidth policies.

3.2.4 Application Scenario

Typical enterprise use cases for bandwidth control mainly include overall traffic control or per-client rate limiting for traffic entering and leaving a WAN interface.

- **Prioritizing office video conferencing traffic:** Assign high priority to real-time interactive traffic such as video conferencing in the office network to ensure it is forwarded first when bandwidth is constrained, reducing stutter and latency.
  - **Recommended Fusion Gateway QoS rule:** Create a QoS rule with **Policy = Prioritize**. Set **Source = Device** and select video conferencing endpoints (or set **Source = Network** and select the office subnet). Set **Destination = App Category** and select conferencing application categories.
  - **Effect:** Conferencing traffic is sensitive to latency and jitter. Marking it as high priority ensures it is forwarded first during WAN congestion, reducing stutter, disconnections, and latency, and improving meeting quality.
- **Limiting large file downloads to prevent congestion:** Rate limit high-bandwidth, non-real-time traffic such as large file or large video downloads to prevent it from consuming excessive bandwidth and slowing down critical services.
  - **Recommended Fusion Gateway QoS rule:** Create a QoS rule with **Policy = Prioritize and Limit**. Configure **Upload Bandwidth Limit** and **Download Bandwidth Limit**. Set **Speedburst = Long** (recommended). Set **Source = Network** and select the office subnet. Set **Destination = App Category** and select P2P/video streaming-related categories. Set **Bandwidth Control Mode = Shared** so that this type of traffic from the subnet shares the configured bandwidth cap.
  - **Effect:** This type of traffic does not require low latency but tends to occupy bandwidth continuously at a high level. Setting a bandwidth cap via Prioritize and Limit allows it to proceed while preventing it from crowding out critical services (e.g., conferencing and POS). Using Shared keeps the entire download category within a

single total cap.

- **Limiting and isolating guest network traffic:** Restrict outbound traffic from the guest network to prevent a single guest or all guests together from squeezing office bandwidth.
  - **Recommended Fusion Gateway QoS rule:** Create a QoS rule with **Policy = Prioritize and Limit**. Configure **Upload Bandwidth Limit** and **Download Bandwidth Limit**. Set **Speedburst = Short** (recommended). Set **Source = Network** and select the guest subnet. Set **Destination = Any**. Set **Bandwidth Control Mode = Individual**. The system creates an independent token bucket per guest device based on "source IP + IP protocol (IPv4/IPv6)," meaning IPv4 and IPv6 traffic from the same guest device each has its own bandwidth quota. This prevents any single guest from monopolizing the entire bandwidth and improves fairness in multi-user environments.
  - **Effect:** Guest traffic is typically web browsing or messaging and is not sensitive to bursts. Short keeps rate limiting relatively strict. Individual ensures each guest is controlled independently, preventing one heavy user (e.g., large downloads, video playback, etc.) from monopolizing bandwidth and degrading overall performance.

3.3 Priority Scheduling

When congestion occurs on a routing interface, packets entering the high-priority queue are forwarded first. High- and low-priority queues are served using a fixed default weight ratio of 9:1 (this ratio is a built-in scheduling weight in Fusion Gateway QoS). The implementation is based on FQ_CoDel + DRR queues.

3.3.1 Protocol

FQ_CoDel (FlowQueue_CoDel) is a queue management mechanism that combines **Fair Queuing (FQ)** with the **Controlled Delay (CoDel)** algorithm. It first places packets from different flows into separate queues and then transmits them in turn using fair scheduling, preventing a single large flow from monopolizing bandwidth. Meanwhile, CoDel continuously monitors queue delay and proactively drops or marks packets when congestion occurs, controlling queue length and queuing delay. This effectively mitigates Bufferbloat and improves the experience of real-time services such as voice, video conferencing, and online gaming.

DRR (Deficit Round Robin) is an efficient fair scheduling algorithm. It maintains a **Deficit Counter** and a fixed **Quantum** for each flow. The scheduler serves flows in a round-robin manner. If a packet's size exceeds the remaining quota, that flow is skipped for the current round, and any unused (or insufficient) quota is carried over to the next round. This approach accommodates different packet sizes while ensuring that, over time, flows receive approximately fair bandwidth allocation. Due to its simple implementation and low

computational overhead, DRR is widely used in network devices and queue scheduling systems, and it is also one of the foundations for fair flow scheduling in FQ_CoDel.

3.3.2 Principle

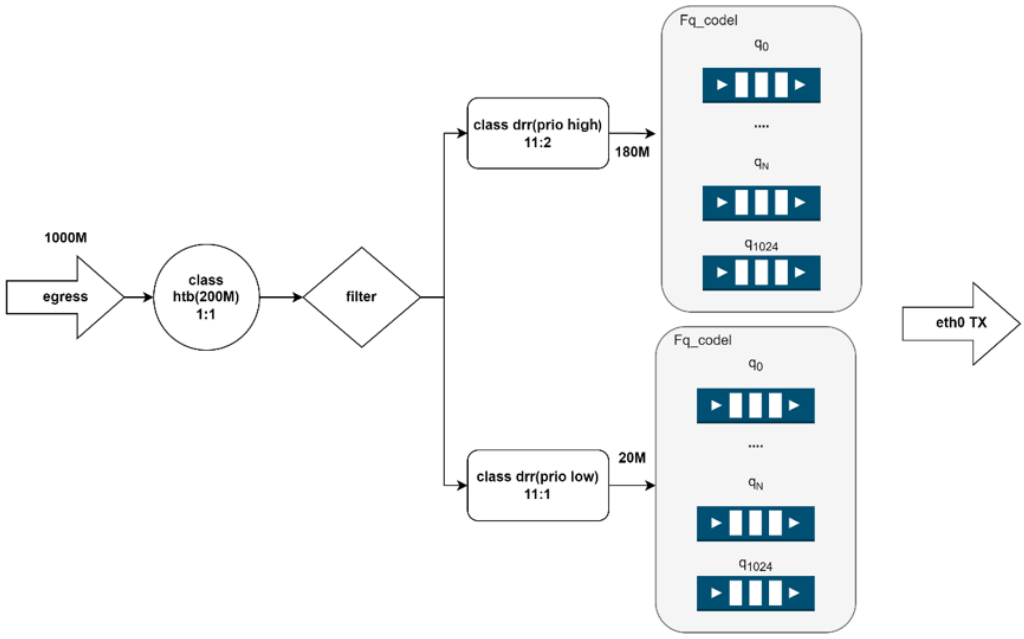


Figure 3-9 Priority Scheduling Example

As shown in the figure, the total egress bandwidth is limited to 200 Mbps by HTB. Traffic is then classified by filters into high-priority and low-priority groups, which enter two DRR queues respectively. Under each DRR queue, an FQ_CoDel instance is attached.

The two DRR queues share transmission opportunities using a fixed 9:1 weight allocation (high-priority queue weight = 9; low-priority queue weight = 1). This is a built-in default scheduling weight in Fusion Gateway QoS and is **not** a user-adjustable example parameter.

DRR uses round-robin scheduling. In each scheduling round, it allocates a byte budget (Deficit Counter) to each queue. If the remaining budget is insufficient to transmit the current packet, that queue is skipped for the round, and the unused (or insufficient) budget is carried over to the next round. For example, if the high-priority quantum is 9,000 bytes and the low-priority quantum is 1,000 bytes, and both traffic classes are continuously present, the high-priority queue receives about 90% of the transmission opportunities per round, while the low-priority queue receives about 10%.

Note that the 9:1 ratio is enforced strictly only under congestion when queues remain backlogged and tokens are insufficient. When bandwidth is sufficient and tokens are available, queues forward packets according to token availability and normal ordering, without forcing this ratio.

FQ_CoDel manages traffic within each priority class:

- The **FQ (Fair Queuing)** component hashes flows based on the 5-tuple (i.e., source address, destination address, source port, destination port, and protocol) and distributes them into many sub-queues. This prevents a single large download connection from filling the entire queue. For example, when a file download flow and a voice flow coexist, they are placed into different sub-queues and scheduled separately, ensuring small, real-time flows are not blocked for long periods by large flows—achieving fairness between flows.
- The **CoDel (Controlled Delay)** component addresses Bufferbloat. Traditional queues often detect congestion based on queue length, while CoDel focuses on how long packets stay in the queue. When CoDel detects that queuing delay continuously exceeds a target value (about 5 ms by default) for a period of time, it determines that persistent congestion is occurring and proactively drops a small number of packets or applies ECN marking. After receiving these congestion signals, TCP senders reduce their sending rates, which shortens the queue and restores normal queuing delay.

Enforcement mechanism: Based on the WAN port's **Bandwidth Control** configuration and **QoS Activation Threshold**, whether traffic on the WAN is subject to QoS rule matching and how strongly weighted priority scheduling intervenes is determined by the following three cases. Uplink and downlink are evaluated independently.

1） **Bandwidth Control is not enabled on the WAN port:** This is equivalent to enabling Bandwidth Control in that WAN interface and setting Upstream/Downstream Bandwidth to the actual negotiated uplink/downlink rates of the WAN port. For example, if three WAN interfaces all have Bandwidth Control disabled, with WAN1 uplink/downlink = 2.5 Gbps, WAN2 = 1 Gbps, and WAN3 = Auto (but the negotiated uplink/downlink rate in the environment is 100 Mbps), then this equates to enabling Bandwidth Control on all three WANs and setting Upstream/Downstream Bandwidth to 2.5 Gbps, 1 Gbps, and 100 Mbps, respectively. In this case, traffic traversing the WAN matches Fusion Gateway QoS rules from top to bottom until the first rule matches or no rules match. The matched rule determines traffic classification and handling (e.g., priority marking, rate limiting, etc.). High- and low-priority traffic is scheduled with the fixed 9:1 weight ratio. This ratio is enforced strictly only when tokens are insufficient in that WAN direction (i.e., congestion). When bandwidth is sufficient and tokens remain available, traffic is forwarded as long as tokens are available, without enforcing the ratio.

2） **Bandwidth Control is enabled, but QoS Activation Threshold is not selected:**

Traffic traversing the WAN matches Fusion Gateway QoS rules from top to bottom until the first rule matches or no rules match. The rule determines classification and handling. High- and low-priority traffic is scheduled using the predefined weight ratio (strictly enforced when tokens are insufficient; not enforced when tokens are sufficient).

3） **Bandwidth Control is enabled and QoS Activation Threshold is selected:** When traffic in a given WAN direction (i.e., uplink or downlink) reaches Bandwidth (Upstream/Downstream) × QoS Activation Threshold (e.g., 1,000 Mbps × 50% = 500 Mbps), traffic in that direction begins matching Fusion Gateway QoS rules from top to bottom until the first rule matches or no rules match, and is scheduled according to the weight ratio (strictly enforced under congestion when tokens are insufficient; when tokens are sufficient and bandwidth has not become a bottleneck, forwarding primarily follows "forward if tokens are available"). Before the threshold is reached, bandwidth in that direction is relatively sufficient, and traffic is forwarded in normal order without being constrained by the weight ratio (i.e., normal state).

3.3.3 Framework

From a system architecture perspective, the priority scheduling module sits at the very end of the Fusion Gateway QoS forwarding plane and is the final executor of all classification and rate-limiting policies. Key architectural considerations include:

- **Decoupling rate limiting and scheduling**

One core design decision is to decouple bandwidth limiting (i.e., rate limiting) from priority scheduling. Rate limiting is completed by an independent token bucket module in the rule chain, and packets enter the TC qdisc scheduler only after passing through the rate-limiting module.

- **Ingress scheduling and the IFB mechanism**

The TC framework's queuing and scheduling is primarily designed for egress. To implement priority scheduling for WAN ingress (i.e., download direction) traffic, Fusion Gateway QoS uses the Linux IFB (Intermediate Functional Block) virtual device technique. IFB is a Linux virtual network device whose key characteristic is "receive then transmit."

By configuring an ingress filter on the WAN interface, inbound packets are redirected to an IFB device via a mirred action. The packets are then re-queued on the IFB's egress direction, where standard TC qdisc rules can be attached for queuing and scheduling. A conntrack mark synchronization mechanism is also used to propagate classification marks to inbound packets, ensuring correct priority recognition on ingress.

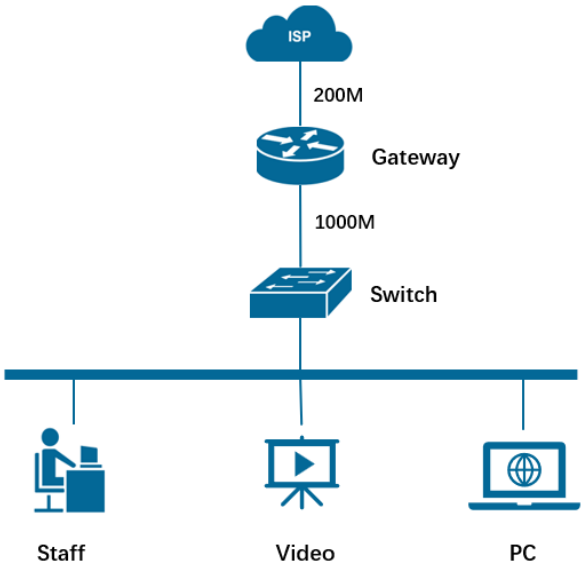
- **Strategy comparison**

In consumer routers and SMB QoS, three mainstream TC scheduling strategies are commonly used:

- **CAKE (Common Applications Kept Enhanced):** An intelligent all-in-one scheduler integrating shaping, queuing, and classification. It typically requires only bandwidth parameters and can automatically classify based on DSCP, making configuration simple but less flexible.
- **HFSC (Hierarchical Fair Service Curve):** Uses service curves to achieve precise bandwidth and delay allocation, supporting real-time (rt) and link-sharing (ls) guarantees. It provides the highest precision but also the highest CPU cost.
- **HTB + FQ_CoDel:** HTB provides flexible hierarchical bandwidth control, while FQ_CoDel provides fair queuing and Active Queue Management (AQM) within each queue, achieving a good balance between flexibility and performance.

Fusion Gateway QoS adopts **HTB + DRR + FQ_CoDel**. This divide-and-conquer architecture achieves an optimal balance among functional flexibility, performance overhead, and development complexity.

3.3.4 Application Scenario



A key enterprise use case for priority scheduling is prioritizing voice and video conferencing: mark video conferencing traffic (e.g., Zoom/Teams) as the highest priority. At the same time, configure a reasonable bandwidth cap for the high-priority queue to prevent it from over-consuming bandwidth and interrupting other services.

As shown in the figure, under normal conditions all endpoints share 200 Mbps. When congestion occurs, bandwidth is allocated according to priority and weights: high-priority traffic is forwarded first, and low-priority traffic uses the remaining resources without affecting critical services.

4 Deployment Scenarios and Solutions

This chapter describes how Fusion Gateway QoS is used in real deployments through three typical scenarios. Each scenario provides end-to-end guidance, including requirements analysis, solution design, and network topology.

➢ **Application Scope**

Fusion Gateway QoS rules apply only to uplink and downlink traffic that traverses a WAN port. Traffic between internal network devices does not pass through a WAN port, so the rules do not apply.

When traffic passes through a WAN port, the port's Bandwidth Control configuration will determine whether to match Fusion Gateway QoS rules. If so, traffic is matched against rule cards/entries from top to bottom in order. After a match is found, the system executes the action defined by the rule policy (e.g., priority forwarding or rate-limited forwarding). If no rules match, the traffic is forwarded directly. For this reason, it's recommended to place the most important QoS rule cards/entries at the top.

➢ **Limitations and Impact**

Fusion Gateway QoS trades some forwarding performance for guaranteed bandwidth and priority for critical services. Its limitations and impact are mainly reflected in the following areas:

- CPU utilization

  Fusion Gateway QoS must identify, classify, enqueue, and schedule traffic. Compared to the fast forwarding path when Fusion Gateway QoS is disabled, this consumes additional CPU resources. In deployments with multiple WAN ports or a large number of rule entries, gateway load may increase.

- Hardware acceleration impact

  With Fusion Gateway QoS enabled, some hardware offload features might not apply to traffic that matches rules. That traffic is processed by the CPU instead, which can reduce overall forwarding throughput. Evaluate whether to enable Fusion Gateway QoS based on actual load and forwarding performance requirements.

- Limited scope

Fusion Gateway QoS can schedule traffic only for WAN-port congestion that the gateway can control. It cannot influence network segments outside the gateway, such as wireless medium contention or upstream ISP congestion. The degree of improvement in transmission quality depends on end-to-end link conditions.

➢ **Key Parameters**

To avoid ambiguity, the following table provides definitions for key parameters.

Parameter	Description
Bandwidth Control	• **Configured per WAN port:** Each WAN port has its own Bandwidth Control settings. • **Bandwidth Control not selected:** All traffic that traverses this WAN port is matched against Fusion Gateway QoS rules from top to bottom until the first match is found, or until no rules match. In this option, Upstream Bandwidth and Downstream Bandwidth are set to the WAN port's negotiated uplink/downlink rate. • **Bandwidth Control selected:** Configure Upstream Bandwidth and Downstream Bandwidth manually based on business requirements.
Upstream Bandwidth/ Downstream Bandwidth	• **Prerequisite:** Bandwidth Control must be selected. • **Explanation:** Used as the baseline for calculating the QoS Activation Threshold and as the bandwidth reference for congestion detection. When uplink or downlink traffic on the WAN port approaches this bandwidth and tokens are insufficient, high- and low-priority traffic are scheduled with a fixed 9:1 weight. When bandwidth and tokens are sufficient, traffic is forwarded normally without enforcing the ratio. • **Recommendation:** Set them to the actual uplink and downlink rates provided by your ISP, not the WAN negotiated rate. Setting values too high limits the effectiveness of Fusion Gateway QoS rules, while setting them too low wastes bandwidth and reduces peak-time performance.
QoS Activation Threshold	• **QoS Activation Threshold not selected:** All traffic that traverses this WAN port is matched against Fusion Gateway QoS rule cards/entries from top to bottom until the first match is found, or until no rules match. • **QoS Activation Threshold selected:** If the WAN uplink and downlink traffic is below **Upstream Bandwidth/Downstream Bandwidth × QoS Activation Threshold**, traffic does not match Fusion Gateway QoS rules and is forwarded directly. When uplink or downlink traffic reaches or exceeds **Upstream Bandwidth/Downstream Bandwidth × QoS Activation Threshold**, all traffic in the corresponding direction is matched against Fusion

	Gateway QoS rule cards/entries from top to bottom. If a rule matches, the action defined by the rule policy is applied (e.g., priority forwarding or rate-limited forwarding). If no rules match, the traffic is forwarded directly.
UDP Bandwidth Quota	• **UDP Bandwidth Quota not selected:** The system does not specifically track UDP bandwidth share, and no UDP quota is applied. • **UDP Bandwidth Quota selected:** When WAN uplink or downlink traffic reaches or exceeds the configured **Upstream Bandwidth/Downstream Bandwidth**, the system limits the percentage of WAN bandwidth that UDP traffic can consume in that direction. When usage is below the configured **Upstream Bandwidth/Downstream Bandwidth**, UDP traffic is allowed to exceed the quota. If you want a fixed UDP bandwidth limit, create a rule that matches UDP traffic and set **Policy** to **Limit**. • **Recommendation:** Enable this option to prevent UDP-heavy traffic (often P2P or streaming) from consuming bandwidth needed by critical services. Note that many online games use UDP; setting the UDP quota too low can negatively affect game traffic. Use caution in gaming scenarios.

Speedburst	• **Prerequisite:** If traffic matches a rule whose **Policy** is **Limit** or **Prioritize and Limit**, the Speedburst rate-limiting behavior applies.
	• **Modes:**
	  • **Off:** Limits traffic to the configured bandwidth. Even if tokens accumulate in the token bucket, bursting is not allowed. Both instantaneous and sustained rates are strictly capped at the configured bandwidth.
	  • **Short:** Allows short bursts when tokens have accumulated. The burst window is short and the overshoot is small, minimizing the impact on rate-limiting strictness. Recommended for web browsing, POS, and interactive queries.
	  • **Long:** Allows longer bursts and absorbs more instantaneous traffic. Recommended for large transfers such as downloads, cloud sync, and streaming.
	• **Recommendation:** Use **Short** for bursty, latency-sensitive traffic (e.g., web browsing, POS transactions, and interactive queries) to improve responsiveness without relaxing rate-limiting discipline too much. Use **Long** for sustained high-volume traffic (e.g., downloads and cloud sync) when user experience is the priority. Use **Off** when strict control is required.

Bandwidth Control Mode	• **Prerequisite:** If traffic matches a rule whose **Policy** is **Limit** or **Prioritize and Limit**, the Bandwidth Control Mode behavior applies. • **Modes:**

	• **Shared:** All clients/flows that match the rule share a single bandwidth allowance. The total bandwidth of all matched traffic combined is capped at the bandwidth limit configured for the rule, regardless of client or source IP. Use this to control the total volume for a traffic type. • **Individual:** Bandwidth allowances are maintained separately per source IP + IP protocol (IPv4/IPv6). Each "source address + IP version" gets its own limited bandwidth, which prevents a single client from consuming all bandwidth and improves per-client fairness.

4.1 Chain Retail and Restaurant Scenarios

4.1.1 Scenario Description

Chain retail stores and restaurants (e.g., fast food restaurants, convenience stores, and branded specialty stores) operate multiple branches under centralized management from headquarters. Each branch typically deploys the following devices:

- POS terminals for transaction processing
- Digital signage displays for menu presentation and advertising
- Security cameras for in-store monitoring
- Wi-Fi APs that provide wireless access for employees and customers
- Back-office PCs for inventory and store management

During peak business hours, multiple types of traffic compete for limited WAN bandwidth. POS transactions must complete reliably, digital signage must play smoothly, security monitoring must remain stable for cloud/remote viewing, and customers (i.e., Wi-Fi users) may also stream video or download large files.

Store traffic can be divided into two categories:

- **Remote/cloud services over the WAN port** (e.g., POS transactions uploaded to headquarters or an acquirer system, digital signage content downloaded from the cloud or headquarters, and cloud storage/remote playback for surveillance video). This traffic is eligible for Fusion Gateway QoS.
- **Local services** (e.g., NVR local recording storage and playback, or signage content delivered within the store LAN). This traffic does not traverse the WAN port, so Fusion Gateway QoS does not apply.

4.1.2 Requirements and Pain Points

Requirement	Pain Point
POS transaction reliability	When the store uploads transaction data to headquarters or an acquirer system, WAN congestion can cause failures or timeouts,

	leading to customer dissatisfaction and revenue loss.
Smooth digital signage playback	When signage downloads content from the cloud or headquarters, insufficient bandwidth can cause slow updates or frozen playback.
Stable remote monitoring	During peak hours, remote viewing, cloud storage uploads, or remote playback can experience stutter, frame drops, or disconnects.
Customer Wi-Fi management	Customer Wi-Fi traffic (e.g., video streaming and large downloads) can consume WAN bandwidth and crowd out the store's critical traffic.
Consistent configuration across stores	Stores require a unified Fusion Gateway QoS policy, and per-store manual configuration is not practical.
No dedicated IT staff	Store employees may not be able to configure complex network policies.

4.1.3 Network Topology

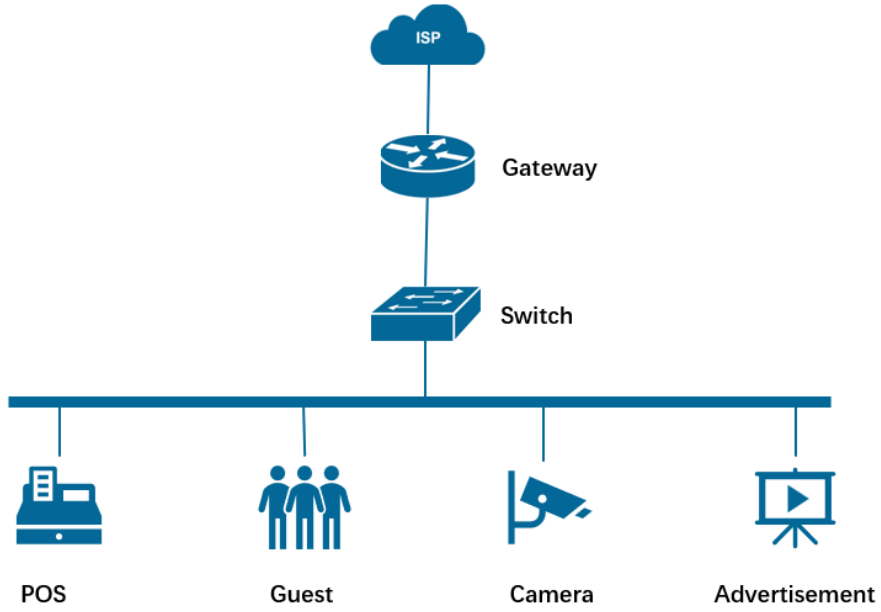


Figure 4-1: Chain retail/restaurant network topology

4.1.4 Solution

➢ **Overview**

1. **Easy configurations**

   Create and enable Fusion Gateway QoS rules to take effect immediately. A store can get baseline protection by creating and enabling Fusion Gateway QoS rules (e.g., setting POS traffic to **Prioritize** and customer Wi-Fi to **Limit**) without requiring dedicated IT staff.

   For more accurate congestion detection and rate-limiting behavior, configure

Upstream Bandwidth/Downstream Bandwidth under **Bandwidth Control** as the bandwidth baseline, and configure **QoS Activation Threshold** as needed (if it is not configured, rules are always active). With **QoS Activation Threshold** configured, traffic in a given direction starts matching rules only when WAN uplink or downlink traffic reaches (threshold %) × (configured bandwidth in that direction).

If **Bandwidth Control** is not configured, Fusion Gateway QoS rules are still always active, and the system uses the WAN port's negotiated rate as the bandwidth reference. Enable **UDP Bandwidth Quota** as needed to prevent P2P/streaming traffic from consuming bandwidth during peak hours.

2. **Full-spectrum traffic identification**

   Fusion Gateway QoS rules can match traffic sources by Device, Network, IP (e.g., IP Subnet, IP Range, and IP Group), and Port. Destinations can be classified by App, App Category, Domain, and IP. You can also filter by IPv4/IPv6 and protocol lists (e.g., TCP, UDP, ICMP, etc.). This lets you precisely distinguish WAN traffic for POS, digital signage, security monitoring, and customer Wi-Fi.

3. **Differentiated priority guarantees**

   Rules can apply to a specific WAN port or to all WAN ports. Choose among three policies: **Prioritize** (priority forwarding for POS, signage, and surveillance), **Limit** (rate-limit customer Wi-Fi), and **Prioritize and Limit** (prioritize critical traffic while keeping it under control). Use **Schedule** to switch rules between peak and off-peak hours. When the store WAN port is congested, Fusion Gateway QoS prioritizes forwarding for WAN-based POS, signage, and surveillance traffic to reduce the risk of transaction timeouts and video stutter (improvements are limited to WAN congestion points that the gateway can control).

4. **Granular rate limiting and consistent marking**

   For rate-limiting policies, configure **Upload/Download Bandwidth Limit** and the **Speedburst** mode (Off/Short/Long). In **Advanced**, select **Shared** or **Individual** to control customer traffic as a whole or per device.

   You can also use **Match DSCP** to match existing DSCP (Differentiated Services Code Point) markings, and rewrite DSCP for outbound packets per rule (each rule can set its own outbound DSCP value). This helps ensure a consistent "critical traffic first, guest traffic limited" policy across branches. Configure the outbound DSCP value and scope per rule.

- **Configuration Steps**

For detailed steps, refer to the FAQ: How to Use the Gateway QoS Function with an Omada Fusion Gateway

- **QoS Rule Management**

- Rule evaluation order: Rules take effect in order. Put higher-priority rules first. You can drag rule cards/list entries to change rule order and priority. For the same traffic class, only the first matching rule is applied.
- Enable or disable a rule: Use the rule's **Status** setting to control whether a rule is active. Turning **Status** on enables the rule; turning **Status** off disables the rule.

4.2 Small and Midsize Office Scenarios

4.2.1 Scenario Description

Small and midsize office networks typically deploy multiple wireless access points (APs) and 10–30 clients, including desktops, laptops, VoIP phones, video conferencing systems, printers, and NAS. During business hours, multiple types of traffic share a single WAN uplink, including video conferencing, VoIP, cloud applications (e.g., SaaS/CRM/ERP), file sync, system patches, and guest Wi-Fi.

Office traffic can be grouped into two categories:

- **Remote/cloud services over the WAN port** (e.g., video conferencing, VoIP calls, SaaS/CRM/ERP cloud applications, file sync and cloud backup, system updates, and guest internet access). This traffic is eligible for Fusion Gateway QoS.
- **Local services** (e.g., direct NAS access within the office network, LAN file sharing, and printing). This traffic does not traverse the WAN port, so Fusion Gateway QoS does not apply.

Pay special attention to NAS traffic: Internal sharing access (e.g., office devices accessing the NAS directly on the LAN) does not use the WAN port and is not affected by Fusion Gateway QoS. Remote/cloud access (e.g., offsite users accessing the NAS remotely, or NAS-to-cloud sync/backup) traverses the WAN port and is eligible for Fusion Gateway QoS.

Without Fusion Gateway QoS, background bulk transfers (e.g., file sync and cloud backup) or a single client performing a large download can quickly consume outbound bandwidth. This can degrade real-time services such as video meetings and VoIP calls (e.g., stutter, audio drops, or disconnects), directly impacting business communication. Fusion Gateway QoS addresses this by prioritizing WAN-based office traffic when bandwidth is congested.

4.2.2 Requirements and Pain Points

Requirement	Pain Point
Real-time service quality	Video conferencing, VoIP, and cloud applications (e.g., SaaS/CRM/ERP) can stutter or experience packet loss over the WAN, impacting business communication.
Separation of real-time and bulk traffic	High-volume WAN transfers (e.g., cloud backup, file sync, and system patches) compete with real-time traffic for bandwidth.

Requirement	Pain Point
Fairness across users	A single client's large downloads can monopolize the WAN uplink and degrade other users' experiences.
Guest network controls	Guests streaming media or downloading large files consume bandwidth needed for office traffic.
Offsite/cloud NAS access	Remote NAS access and NAS cloud sync can be slow or unresponsive.
Low operational overhead	No dedicated network engineer; policies must be template-based and easy to configure.

4.2.3 Network Topology

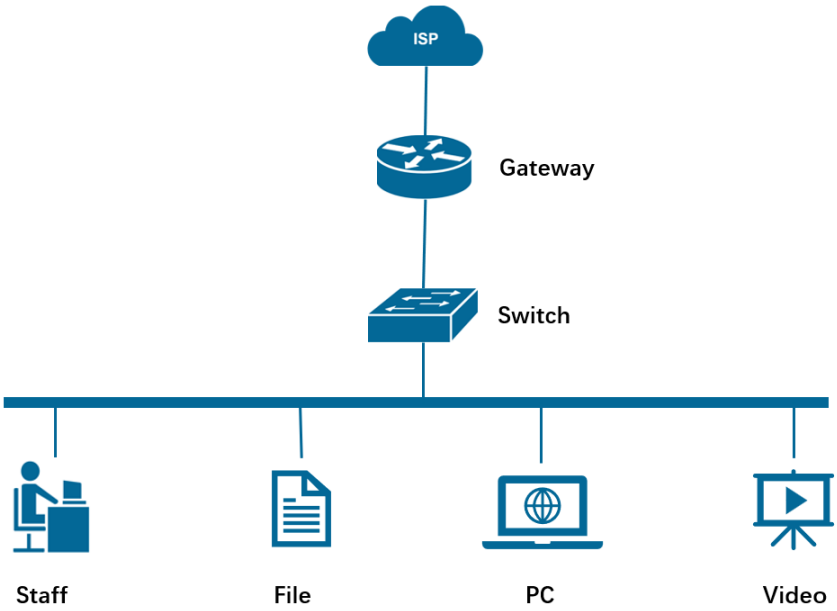


Figure 4-2: Small and midsize office network topology

4.2.4 Solution

- **Overview**

  1. **Easy configurations**

     Create and enable Fusion Gateway QoS rules to take effect immediately. The office network can get baseline protection by creating and enabling Fusion Gateway QoS rules (e.g., set video conferencing to **Prioritize**, and set P2P/guest traffic to **Limit**) without a dedicated network engineer.

     For more accurate congestion detection and rate-limiting behavior, configure **Upstream Bandwidth/Downstream Bandwidth** under **Bandwidth Control** as the bandwidth baseline, and configure **QoS Activation Threshold** as needed (if it is not configured, rules are always active). With **QoS Activation Threshold** configured, traffic in a given direction starts matching rules only when WAN uplink or downlink traffic reaches (threshold %) × (configured bandwidth in that direction).

If **Bandwidth Control** is not configured, Fusion Gateway QoS rules are still always active, and the system uses the WAN port's negotiated rate as the bandwidth reference. Enable **UDP Bandwidth Quota** as needed to reduce peak-hour impact from large downloads and streaming traffic.

2. **Full-spectrum traffic identification**

   Fusion Gateway QoS rules can match traffic sources by Device, Network, IP (e.g., IP Subnet, IP Range, and IP Group), and Port. Destinations can be classified by App, App Category, Domain, Port, and IP. You can also filter by IPv4/IPv6 and protocol lists (e.g., TCP/UDP). This lets you precisely distinguish WAN traffic for video conferencing, cloud applications, ERP, and the guest network.

3. **Differentiated priority guarantees**

   Rules can apply to a specific WAN port or to all WAN ports. Use one of three policies: **Prioritize** (e.g., video conferencing and online collaboration first), **Limit** (e.g., guest Wi-Fi / P2P rate limiting), and **Prioritize and Limit** (e.g., prioritize critical traffic while keeping it under control). Use **Schedule** to switch policies between business hours and after-hours.

   When the office WAN port is congested, Fusion Gateway QoS prioritizes video conferencing and cloud application traffic to improve calling and collaboration experience (improvements are limited to WAN congestion points that the gateway can control).

5. **Granular rate limiting and consistent marking**

   For rate-limiting policies, configure **Upload/Download Bandwidth Limit** and the **Speedburst** mode (e.g., Off/Short/Long). In **Advanced**, select **Shared** or **Individual** to control office traffic as a whole or per device.

   You can also use **Match DSCP** to match existing DSCP (Differentiated Services Code Point) markings, and rewrite DSCP for outbound packets per rule (each rule can set its own outbound DSCP value). This helps ensure a consistent "conferencing traffic first, office traffic orderly, guest traffic limited" policy. Configure the outbound DSCP value and scope per rule.

➢ **Configuration Steps**

For detailed steps, refer to the FAQ: How to Use the Gateway QoS Function with an Omada Fusion Gateway

➢ **QoS Rule Management**

- **Rule evaluation order:** Rules take effect in order. Place higher-priority rules first. You can drag rule cards/list entries to change rule order and priority. For the same traffic class, only the first matching rule is applied.
- **Enable or disable a rule:** Use the rule's **Status** setting to control whether a rule is

active. Turning **Status** on enables the rule; turning **Status** off disables the rule.

4.3 Prosumer Scenarios

4.3.1 Scenario Description

Prosumer users are tech enthusiasts who use SMB-class devices at home to achieve better network performance. A typical prosumer home network includes:

- Game consoles and high-performance gaming PCs
- 4K streaming devices (e.g., smart TVs, Apple TV, and media players)
- Smart home IoT devices (e.g., lights, sensors, and thermostats)
- NAS personal cloud storage
- Remote work devices
- Home security cameras

Prosumer network requirements are demanding: online gaming needs ultra-low latency, 4K streaming needs sustained bandwidth, and background activities (e.g., NAS sync, cloud backup, and IoT firmware updates) must not interfere with real-time applications.

Home traffic can be grouped into two categories:

- **Remote/cloud services over the WAN port** (e.g., online gaming, 4K streaming, NAS-to-cloud sync and backup, cloud backup uploads, offsite access to the NAS or cameras, IoT firmware updates, and remote work). This traffic is eligible for Fusion Gateway QoS.
- **Local services** (e.g., direct NAS access on the LAN, local smart home automation, and local viewing of live camera feeds). This traffic does not traverse the WAN port, so Fusion Gateway QoS does not apply.

Pay special attention to NAS traffic: Direct LAN access to the NAS does not use the WAN port and is not affected by Fusion Gateway QoS. NAS-to-cloud sync/backup and offsite access to the NAS traverse the WAN port and are eligible for Fusion Gateway QoS.

Across WAN-based services, large background sync and cloud backups often consume uplink bandwidth and disrupt real-time gaming and streaming. This can cause latency spikes, buffering, and reduced video quality. Fusion Gateway QoS addresses this by prioritizing latency-sensitive traffic, such as gaming and streaming, when bandwidth is congested.

4.3.2 Requirements and Pain Points

Requirement	Pain Point
Low-latency gaming	Background uploads/downloads over the WAN can interfere with online games and cause latency spikes.

Requirement	Pain Point
Smooth 4K streaming	Other traffic can consume WAN bandwidth, increasing buffering and reducing video quality.
Cloud backups without disrupting real-time apps	NAS-to-cloud sync and cloud backup uploads consume uplink bandwidth and disrupt gaming/streaming.
Remote access to home devices	Offsite access to the NAS and home cameras can be slow or unresponsive.
Work-from-home quality	When someone is gaming or streaming and saturates WAN uplink/downlink, video calls can stutter.
Low-maintenance operation	Users want fine-grained control without having to tweak settings every day.

4.3.3 Network Topology

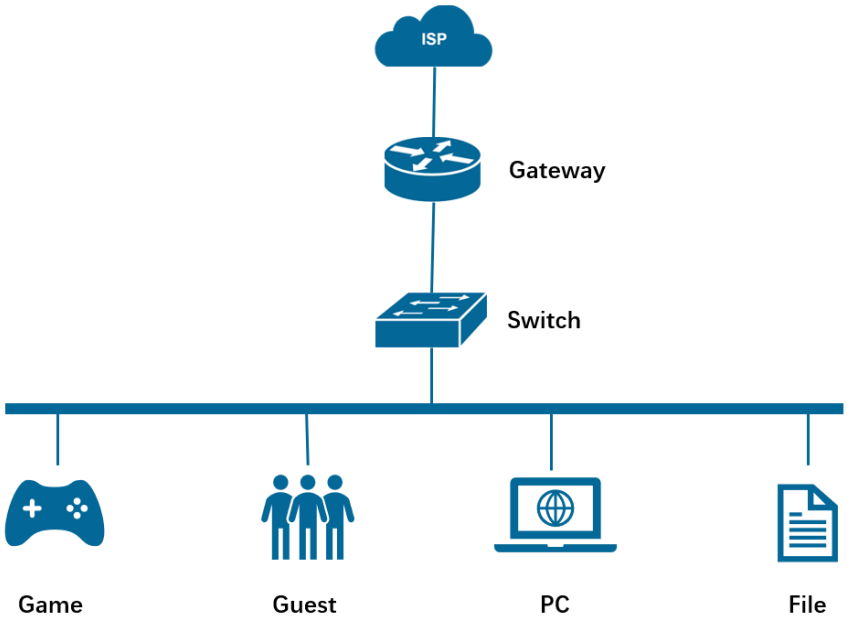


Figure 4-3: Prosumer network topology

4.3.4 Solution

- **Overview**

  1. **Easy configurations**

     Create and enable Fusion Gateway QoS rules to take effect immediately. A home network can get baseline protection by creating and enabling Fusion Gateway QoS rules (e.g., set gaming traffic to **Prioritize** and background downloads to **Limit**).

     For more accurate congestion detection and rate-limiting behavior, configure

Upstream Bandwidth/Downstream Bandwidth under **Bandwidth Control** based on your ISP's actual service rate, and configure **QoS Activation Threshold** as needed (if it is not configured, the rules are always active). With **QoS Activation Threshold** configured, traffic in a given direction starts matching rules only when WAN uplink or downlink traffic reaches (threshold %) × (configured bandwidth in that direction).

If **Bandwidth Control** is not configured, Fusion Gateway QoS rules remain active, and the system uses the WAN port's negotiated rate as the bandwidth reference. Because gaming traffic is often carried over UDP, use caution when enabling **UDP Bandwidth Quota** and avoid setting the quota too low.

2. **Full-spectrum traffic identification**

   Fusion Gateway QoS rules can match traffic sources by Device, Network, IP (e.g., IP Subnet, IP Range, and IP Group), and Port. Destinations can be classified by App, App Category, Domain, Port, and IP. You can also filter by IPv4/IPv6 and protocol lists (e.g., UDP/TCP/ICMP). This lets you precisely distinguish WAN traffic for gaming, NAS, streaming, smart home services, and video calls.

3. **Differentiated priority guarantees**

   Rules can apply to a specific WAN port or to all WAN ports. Use one of three policies: **Prioritize** (e.g., low-latency priority for online games), **Limit** (e.g., rate-limit background downloads and update tasks), and **Prioritize and Limit** (e.g., prioritize gaming/streaming while keeping it under control). Use **Schedule** to switch rules between gaming hours and work hours.

   When the home WAN port is congested, Fusion Gateway QoS prioritizes gaming and streaming traffic to reduce latency spikes and buffering (e.g., improvements are limited to WAN congestion points that the gateway can control).

4. **Granular rate limiting and consistent marking**

   For rate-limiting policies, configure **Uplink/Downlink Bandwidth Limit** and the **Speedburst** mode (e.g., Off/Short/Long). For large downloads outside gaming hours, **Long** is recommended to absorb bursts. In **Advanced**, select **Shared** or **Individual** to control traffic as a whole or per device.

   You can also use **Match DSCP** to match existing DSCP (Differentiated Services Code Point) markings, and rewrite DSCP for outbound packets per rule (each rule can set its own outbound DSCP value). This helps ensure a customized policy for best-in-class gaming latency, smooth NAS/streaming, and stable smart home connectivity. Configure the outbound DSCP value and scope per rule.

➢ **Configuration Steps**

For detailed steps, refer to the FAQ: How to Use the Gateway QoS Function with an Omada Fusion Gateway

➢ **QoS Rule Management**

- **Rule evaluation order:** Rules take effect in order. Place higher-priority rules first. You can drag rule cards/list entries to change rule order and priority. For the same traffic class, only the first matching rule is applied.
- **Enable or disable a rule:** Use the rule's **Status** setting to control whether a rule is active. Turning **Status** on enables the rule; turning **Status** off disables the rule.

5 Representative Models

The Fusion Gateway QoS capabilities described in this document apply to all Fusion Gateway series products that have been officially released.

Fusion Gateways include a built-in controller. All released models are configured and managed through Controller Mode only. Standalone Mode is not provided.

QoS capabilities will be added to traditional ER series gateways based on each model's roadmap. For the supported scope and availability, see each model's release notes.

The following table provides an overview of the currently released Fusion Gateway series models:

Series	Model	Image	Target Scenarios	Firmware Versions	Management Mode
Fusion	Fusion 2.5G PoE V1		• Small Chain Stores • Small Restaurants • Small Offices • Medical Clinics	• Fusion 2.5G PoE(UN)_V1_1.0.10 Build 20260814 • Fusion 2.5G PoE(UN)_V1_1.0.0 Build 20260611	Controller Mode
	Fusion 2.5G V1 Fusion 2.5G V2		• Small Chain Stores • Small Restaurants • Small Offices • Medical Clinics	V1: • Fusion 2.5G(UN)_V1_1.0.30 Build 20260721 • Fusion 2.5G(UN)_V1_1.0.20 Build 20260425 V2: • Fusion 2.5G(UN)_V2_1.0.10 Build 20260814 • Fusion 2.5G(UN)_V2_1.0.0 Build 20260611	Controller Mode

	Fusion G+ V2		• Small Chain Stores • Small Restaurants • Small Offices • Medical Clinics	• Fusion G+(UN)_V2_1.0.10 Build 20260814 • Fusion G+(UN)_V2_1.0.0 Build 20260611	Controller Mode
Fusion Pro	Fusion Pro 2.5G PoE		• Restaurants • Cafes • High-End Residential • Small Retail Shops • Small Clinics	Coming Soon	Controller Mode
	Fusion Pro 2.5G		• Restaurants • Cafes • High-End Residential • Small Retail Shops • Small Clinics	Coming Soon	Controller Mode

6 Future Outlook

6.1 Technology Design and Limitations

Currently, Fusion Gateway QoS integrates traffic identification, bandwidth control, and priority scheduling (see Chapter 3 for details). On the scheduling side, the current solution uses a combination of **HTB + DRR + FQ_CoDel**: HTB provides hierarchical bandwidth limits, DRR prioritizes the high-priority queue using a 9:1 weight ratio, and FQ_CoDel implements fair queuing and Active Queue Management (AQM) within each queue. Overall, the architecture strikes a balance between **functional flexibility, performance overhead, and implementation complexity**, enabling different policy combinations for different scenarios such as retail, small-to-medium offices, and prosumers.

However, this solution faces some limitations:

1) **Policies rely on manual configuration:** Bandwidth limits and DSCP-based classification are manually configured through rules, and queue parameters must be adjusted per scenario.
2) **A learning curve for non-professional users:** Retail stores and small offices often lack dedicated network administrators, making manual tuning costly.
3) **Cannot fully eliminate extreme contention:** Weighted scheduling can significantly improve the experience of critical services under typical congestion, but it cannot

completely eliminate bandwidth contention on an overloaded link.

4) **Resource and forwarding performance overhead:** Fusion Gateway QoS requires traffic classification, queuing, and scheduling, which consume CPU resources. Under high bandwidth, large rule sets, or multi-WAN scenarios, it may increase gateway load. Fusion Gateway QoS may also affect the effective scope of hardware acceleration (e.g., offloading), which can impact actual forwarding performance.

6.2 Future Directions

Based on the limitations described in 6.1, future directions for Fusion Gateway QoS may focus on reducing configuration cost, enhancing intelligent identification and adaptive scheduling, and improving assurance accuracy for real-time services. The following directions are exploratory discussions toward these goals and are industry-level topics, and are not subject to planning or commitments for any specific product.

- **Further simplification of scheduling and configuration:** To address the limitations of "manual configuration dependency" and "high entry barrier for non-professional users," reducing configuration complexity remains an ongoing industry focus. Some platforms are exploring **integrated scheduling mechanisms that combine rate limiting, queuing, and traffic classification**, which require only a small number of parameters (e.g., bandwidth) to automatically classify and schedule traffic based on predefined service levels, thereby reducing reliance on manual configuration. Such mechanisms typically trade some tuning flexibility for simplicity.
- **Smarter traffic identification:** The continued growth of encrypted traffic and emerging applications poses challenges to classification based on static characteristics (e.g., ports), and also increases the workload of manually maintaining rules. To handle this, the industry is broadly exploring improvements in identification accuracy using machine learning and behavioral analysis, and using continuously updated signature/feature libraries to enhance recognition of new applications.
- **Application-aware and link-aware scheduling:** To address bandwidth contention under extreme congestion, one possible direction is to evolve from relatively fixed priority scheduling toward dynamic adjustment **based on real-time quality requirements and link conditions**, enabling more adaptive end-to-end quality assurance. In multi-link scenarios, scheduling may also be integrated with link quality monitoring. This can dynamically select egress paths based on link status and service requirements, improving overall reliability.
- **Fine-grained assurance for media and real-time services:** For real-time services, such as voice and video, current solutions often focus on control signaling. For data flows carrying actual media payloads, the industry is still exploring more precise identification and assurance methods—for example, identifying media ports locally based on

connection negotiation information and applying corresponding processing—to further improve call and audio/video experience.

- **End-to-end coordination and intelligent O&M**
  - ✓ **End-to-end quality coordination:** The industry widely discusses unifying the QoS marking framework across gateways, switches, and wireless access devices to provide consistent assurance from wired to wireless and mitigate quality degradation caused by wireless airtime contention.
  - ✓ **Intelligent assisted tuning:** Big data and cloud-based analytics can derive optimal resource configurations from traffic characteristics, lowering the tuning threshold and serving as a typical approach for introducing AI-assisted network O&M.

7 Appendix

7.1 Glossary

Term	Description
QoS (Quality of Service)	Quality of Service. With limited bandwidth resources, QoS allocates bandwidth and priority to different services to provide end-to-end service quality assurance.
TC (Traffic Control)	The traffic control framework provided by the Linux kernel. Supports traffic classification, scheduling, shaping, and monitoring.
qdisc (Queueing Discipline)	A queueing discipline. The core component of the TC framework that manages packet enqueue and dequeue behavior.
Prio qdisc	A strict-priority queueing discipline that transmits packets in descending priority order (from high to low).
DRR (Deficit Round Robin)	Deficit Round Robin. A weighted round-robin scheduling algorithm that allocates scheduling opportunities to queues according to configured weights.
IFB (Intermediate Functional Block)	An intermediate functional block device used to redirect ingress traffic so that egress scheduling can be applied.
FQ_CoDel (Fair Queuing + Controlled Delay)	A queue management algorithm that combines fair queuing with Active Queue Management (AQM) to control latency.
CIR (Committed Information Rate)	The committed information rate: the long-term average rate limit in a token bucket algorithm.
PIR (Peak Information Rate)	The peak information rate: the short-term burst rate limit in a two-rate token bucket model.
SP (Strict Priority)	A strict-priority scheduling policy that always transmits packets from the highest-priority queue first.
DSCP (Differentiated Services Code Point)	A field in the IP header used to mark service class/priority.

7.2 References

[1] OpenWrt Project. Smart Queue Management (SQM)[EB/OL]. (2019-01-01)[2026-09-02]. https://openwrt.org/docs/guide-user/network/traffic-shaping/start.

[2] Kerrisk M. tc(8) — Linux manual page — Linux traffic control (TC) tool[EB/OL]. (2019-01-01)[2026-09-02]. https://man7.org/linux/man-pages/man8/tc.8.html.

[3] Hubert B. TC HTB — Linux HTB (Hierarchy Token Bucket) queuing discipline manual[EB/OL]. [2026-09-02]. https://linux.die.net/man/8/tc-htb.

[4] Heinanen J, Guerin R. A single rate three color marker[S]. RFC 2697, Internet Engineering Task Force (IETF), 1999-09.

[5] Heinanen J, Guerin R. A two rate three color marker[S]. RFC 2698, Internet Engineering Task Force (IETF), 1999-09.