NXON AI FACTORY PTY. LTD

NXON.AI RoCE

Technology White Paper

Preface

Abstract

This document introduces the technical implementation scheme of RDMA, the challenges RoCEv2 technology poses to the network, and the corresponding strategies.

Keywords

RDMA, RoCE, InfiniBand, PFC, ECN

Terminology

Acronym / Term	Description
RDMA	Remote Direct Memory Access
IWARP	Internet Wide Area RDMA Protocol
RoCE	RDMA over Converged Ethernet
PFC	Priority-based Flow Control
ECN	Explicit Congestion Notification
IBTA	InfiniBand Trade Association
CNP	Congestion Notification Packets

NXON AI FACTORY PTY. LTD

Table of Contents

Overview *4*

1.1 The Birth of RDMA **4**

1.2 Mainstream RDMA Protocols **5**

Technical Implementation *7*

2.1 The Development of RoCE **7**

2.2 RoCEv2 Processing in the Network **8**

2.3 RoCEv2 Packet **9**

2.4 RoCEv2 Working Principle **11**

2.4.1 Congestion Generation 11

2.4.2 RoCE's Go-Back-N Retransmission 12

2.4.3 Building a Lossless Network 14

Typical Applications *17*

3.1 RDMA Solving Network Performance Issues in HPC Scenarios **17**

3.1.1 Scenario Introduction 17

3.1.2 Solution 17

Overview

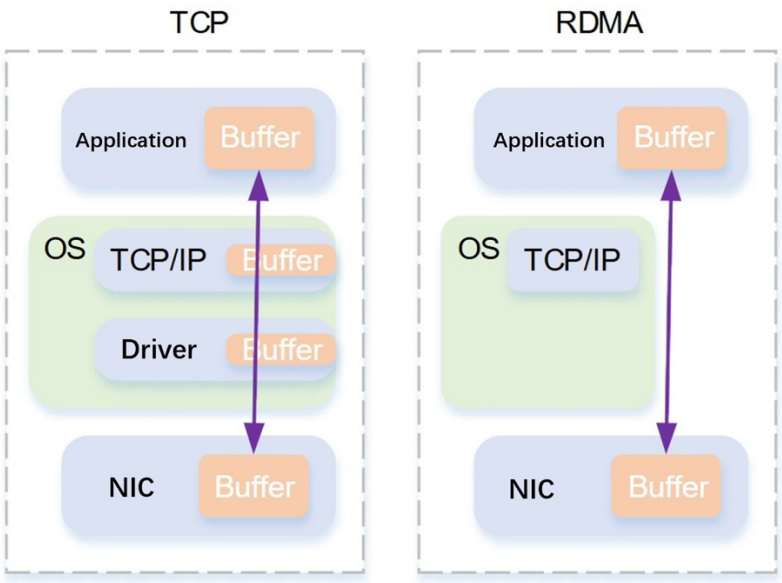
1.1 The Birth of RDMA

AI businesses, represented by AIGC (Artificial Intelligence Generated Content), are developing rapidly, and applications' demand for computing power is growing daily. The LLM model, as the core technology supporting AIGC applications, is also iterating quickly, and model parameters are increasingly expanding. It is reported that the GPT-4 model contains a total of 1.8 trillion parameters across 120 layers. Traditional single GPU servers cannot meet the training tasks for such huge parameters; whether it's video memory, storage space, or training efficiency, they are insufficient. Therefore, it is necessary to build a distributed training system to meet the demands of this task.

Distributed training builds a cluster with massive computing power and video memory capability through multiple GPU nodes. However, the overall computing power does not simply increase linearly with the addition of GPU nodes. The single computation time for model training includes the single-card computation time plus the inter-card communication time. The high-performance network connecting this super cluster directly determines the communication efficiency between GPU nodes, thereby affecting the throughput and performance of the entire AI cluster. Thus, building a high-performance AIGC network has become a key point for AI business clusters.

After decades of development and refinement, TCP/IP seems somewhat inadequate under the wave of AIGC high-performance networks. This is because when servers process TCP/IP packets, the CPU needs to participate multiple times in encapsulation and decapsulation operations, resulting in a fixed latency of tens of microseconds. Simultaneously, in large-traffic data center scenarios, CPU participation in data transmission consumes significant CPU resources. The once peerless TCP/IP has become a performance bottleneck. Therefore, RDMA (Remote Direct Memory Access) technology emerged. RDMA is a technology designed to reduce latency in data storage flows within traditional data centers. As shown in Figure 1-1, RDMA, supported by special network cards, can directly transfer data into the server's memory area, involving less CPU resources, effectively avoiding the consumption of server operating system context switching, reducing server memory usage and CPU overhead, and ultimately effectively reducing the transmission delay of data between the server network card and the internal memory area, while alleviating the overhead caused by large amounts of data transmission on the server.

Figure 1-1 TCP vs. RDMA Network Data Transmission Comparison



1.2 Mainstream RDMA Protocols

In the industry, the main network protocols supporting RDMA are: InfiniBand, iWARP (Internet Wide Area RDMA Protocol), and RoCE (RDMA over Converged Ethernet).

- InfiniBand: This is a high-speed network protocol specifically designed for high-performance computing data centers. It provides extremely high throughput and very low latency. Although InfiniBand performance is excellent, it has not been widely adopted because: InfiniBand requires dedicated hardware and network equipment, which is costly. Furthermore, InfiniBand is a complex, proprietary network protocol requiring specialized technical personnel for maintenance and has poor compatibility with the widely used IP Ethernet. This results in a small market share, a small ecosystem, and limited use cases for InfiniBand. In summary, InfiniBand offers excellent performance but has high costs, high maintenance complexity, and limited application scenarios.
- iWARP: This is a TCP-based RDMA protocol that uses the TCP protocol stack to transmit RDMA data.
- RoCE: This is a UDP-based RDMA protocol that uses the UDP protocol stack to transmit RDMA data. RoCE is divided into two versions, RoCEv1 and RoCEv2, with RoCEv2 supporting more features and higher performance. RoCEv2 technology's performance is basically on par with InfiniBand. Although static latency is slightly higher, AI businesses are more concerned with dynamic latency during operation, where the two technologies are comparable. The main advantages of RoCEv2 technology lie in its standardization, openness, and low cost (Capex & Opex). The recently established UEC (Ultra Ethernet Consortium) has also clearly stated its

intention to continuously improve network performance based on Ethernet technology standards to meet the evolving technical requirements in the AI field.

The three RDMA protocols each have their own characteristics. Analyzing them from the perspectives of performance, cost, delivery cycle, and sustainability, we get the comparison shown in Table 1-1. In the context of AIGC's high-traffic storage and computation scenarios, performance is a crucial indicator for technology selection. Due to iWARP's poor performance, it is not suitable for AIGC scenarios. As shown in Table 1-2, InfiniBand, due to its closed ecosystem and high cost, is also not chosen. Therefore, most TH5-chipset based switch manufacturers have selected RoCE as the technical solution for building AIGC networks.

Table 1-1 InfiniBand and RoCE Solution Comparison

Feature	InfiniBand	RoCE
Performance	Excellent+	Excellent
Cost	High	Low
Stability	Good	Relatively Good
Switch Requirement	InfiniBand Switch	Ethernet Switch

Table 1-2 InfiniBand vs. RoCE Technical Solution Comparative Analysis

Comparison Item	InfiniBand	RoCEv2
Access Bandwidth	400G/800G	400G/800G
Maximum Scale	16K x 400G Ports	32K x 400G Ports
Static Latency	1~2 µs	3~5 µs
Flow Control	Credit-based	PFC / ECN
Fault Recovery	Interconnect Self-healing	Route Fast Switchover
Bandwidth Utilization	96%	97%
Technology Ecosystem	Technologically Closed	Highly Open
Service Provider	Single Vendor	Multiple Software/Hardware Vendors
Delivery Lead Time	4-6 months	1-2 months
Procurement Cost	High	Low

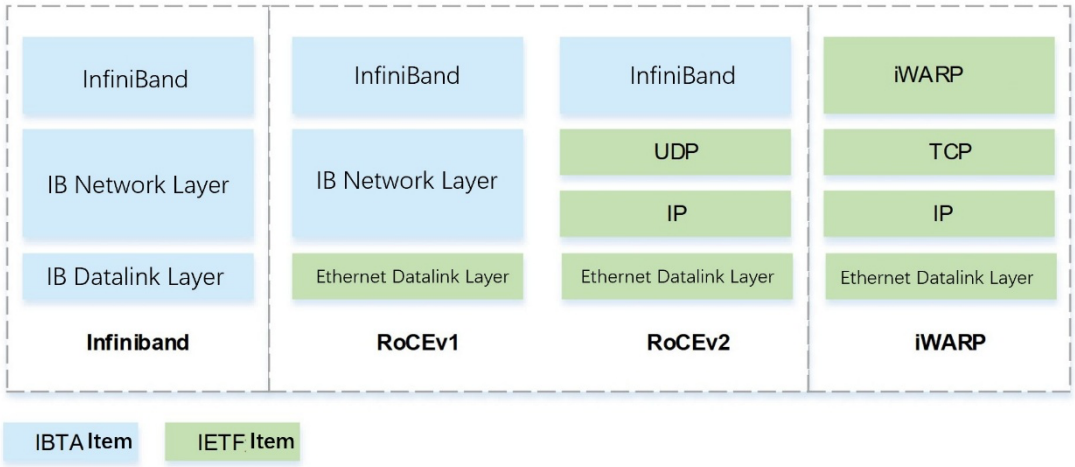
Technical Implementation

2.1 The Development of RoCE

RDMA was first proposed by the IBTA (InfiniBand Trade Association) and used in InfiniBand networks. InfiniBand networks operate with dedicated switches, dedicated network cards, and the proprietary InfiniBand protocol, resulting in a relatively closed ecosystem. Furthermore, as most enterprise services are based on Ethernet, this was not conducive to the iteration and promotion of RDMA. Therefore, the IBTA standards organization defined another set of character specifications, namely RoCE, enabling RDMA to run over Ethernet. RoCE is currently the most effective solution for reducing data processing latency in Ethernet.

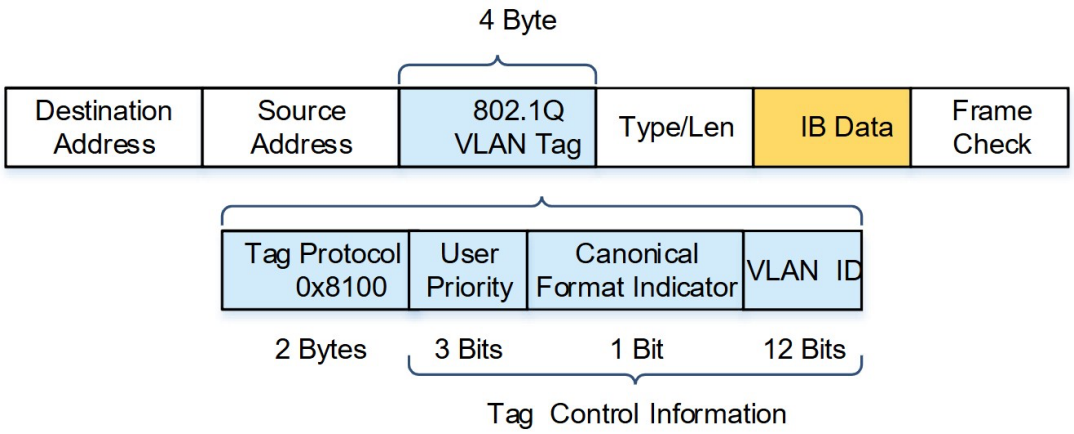
RoCE has two versions, RoCEv1 and RoCEv2. Their network models are shown in Figure 2-1.

Figure 2-1 RDMA Protocol Network Model



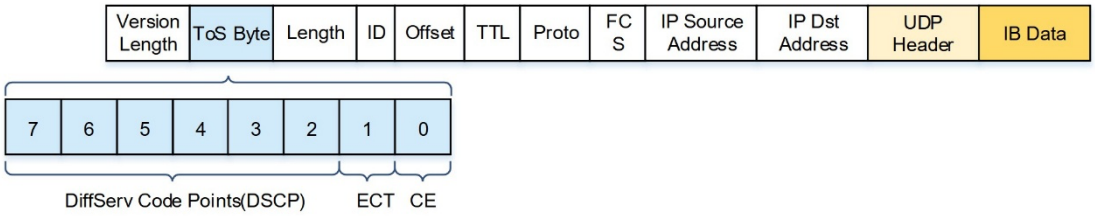
RoCEv1 is a link-layer protocol that encapsulates the RDMA data segment within the Ethernet data segment, plus the Ethernet header, thus belonging to Layer 2 packets. To classify it, only the PCP (Priority Code Point) field (3 bits) in the VLAN (IEEE 802.1q) header can be used to set the priority value, as shown in Figure 2-2. RoCEv1 allows any two hosts within the same broadcast domain to access each other directly.

Figure 2-2 RoCEv1 Frame Format



RoCEv2 encapsulates the RDMA data segment first into the UDP data segment, adds the UDP header, then the IP header, and finally the Ethernet header, belonging to Layer 3 packets, as shown in Figure 2-3. It is an Internet layer protocol, thus enabling routing functionality. For classification, either the PCP field in the Ethernet VLAN or the DSCP field in the IP header can be used.

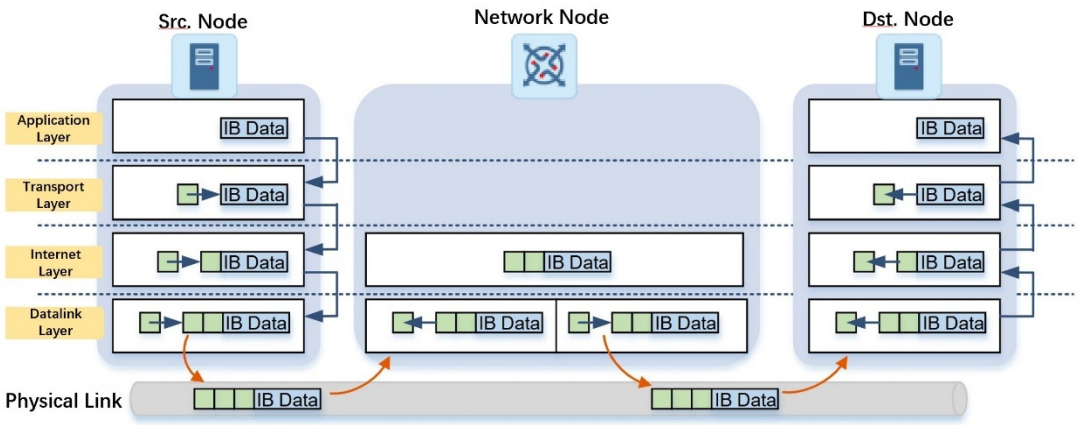
Figure 2-3 RoCEv2 Packet Format



2.2 RoCEv2 Processing in the Network

Figure 2-4 illustrates the processing of RoCEv2 packets between the server and the switch. In the diagram, RoCEv2 provides a low CPU load, low latency data read/write solution for the service sender and receiver. In this solution, only the servers involved in sending and receiving process the RoCEv2 packets. When network devices receive RoCEv2 packets, they only decapsulate up to the IP packet header; the UDP header and IB data segment are not decapsulated or parsed. Therefore, network devices only need to be responsible for the quality of the carrier network to meet the high sensitivity of RDMA to network quality.

Figure 2-4 RoCEv2 Packet Forwarding Process



2.3 RoCEv2 Packet

In the chapter on the development of RoCE, we learned about the position of the RoCEv2 data part within the IP packet. Next, we will further understand the structure of the RoCEv2 data part, as shown in Figure 2-5.

Figure 2-5 RoCEv2 Data Part Structure

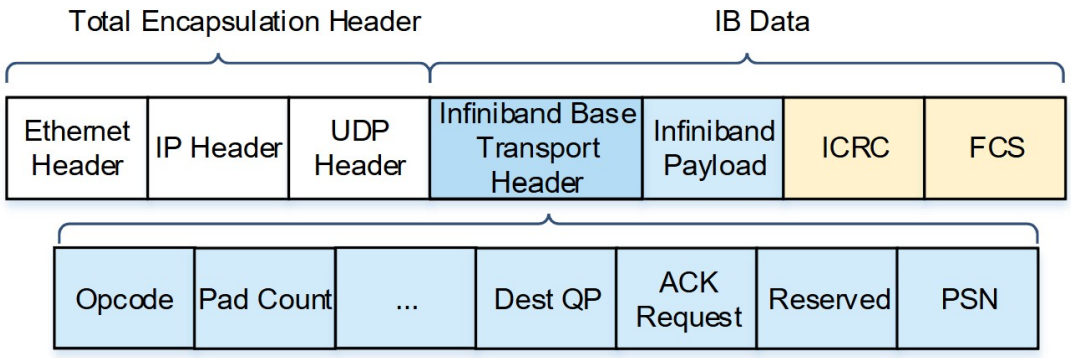


Table 2-1 RoCEv2 Packet Field Description

Field	Description
Ethernet Header	Contains source and destination MAC addresses, and the Ethernet Type field. - **Ethernet Type for RoCEv2:** 0x8915
IP Header	Contains source and destination IP addresses, and other standard IP layer protocol fields.

UDP Header	Contains source and destination ports, length, and checksum fields. - **Default RoCEv2 Destination Port:** 4791
InfiniBand Base Transport Header (BTH)	Carries the core RDMA control fields. (See Table 2-2 for a detailed breakdown).
InfiniBand Payload	The actual data payload of the RDMA message.
ICRC & FCS	Integrity Check Redundancy Code and Frame Check Sequence for error detection.

Table 2-2 InfiniBand Base Transport Header (BTH) Field Description

Field	Description
Opcode	Indicates the RoCEv2 packet type and operation mode. **Common Operation Modes:** • **ConnectMsg:** Used for CM (Communication Management) connection establishment. • **Send:** Sender requests to transfer data to the remote end (no receiver address specified). • **Read:** Sender requests to read data from the remote end. • **Write:** Sender writes data to a specified address in the remote memory. • **ACK:** Response to a sender's request. There are two modes of ACK. - **ACK:** Successful receipt. - **NAK:** Indicates packet loss. • **CNP:** Used for control and status management after connection is established.
Pad Count	Specifies the number of padding bytes added to the InfiniBand Payload.
Dest QP	The destination Queue Pair. A QP is the basic RDMA communication unit, consisting of a Send Queue (SQ) and Receive Queue (RQ), and identifies a RoCEv2 flow.
ACK Request (A)	A flag that, when set, requests an acknowledgment (ACK) from the receiver for this packet.
PSN	The Packet Sequence Number. Used to detect packet loss by checking for consecutive numbers. A non-consecutive PSN triggers a NAK packet.

2.4 RoCEv2 Working Principle

Since network switches only need to focus on the quality of the carrier network and do not need to process RDMA packets, our discussion on the principle will focus on the causes of congestion in RDMA networks and the problems caused by retransmissions due to congestion.

2.4.1 Congestion Generation

There are many reasons for congestion. Listed below are three key and common reasons in data center scenarios:

- Oversubscription Ratio

When designing data center network architecture, considering both cost and benefit, asymmetric bandwidth design is often adopted, meaning uplink and downlink bandwidths are inconsistent. The switch's oversubscription ratio is simply the total input bandwidth divided by the total output bandwidth. For some previous-generation 40G 48-port switches, the downlink bandwidth available for server input may be \(48 \times 10G = 480G \), and the uplink output bandwidth maybe \(6 \times 40G = 240G \). Thus, the chassis oversubscription ratio is 2:1.

This means that when the total uplink packet sending rate of the downstream servers exceeds the total uplink bandwidth, congestion will occur on the uplink ports.

However, switches deployed in NXON.AI AIDCs, such as Ruijie RG-S6990-128QC2XS, supports 128x 400G ports which are available to work as either downlink or uplink ones. It makes the chassis oversubscription ratio is 1:1.

- ECMP

Current data center networks mostly use Fabric architecture and employ ECMP to build multiple equal-cost load-balanced links. It is simple to set a disturbance factor and HASH select a link for forwarding, but this process does not consider whether the selected link itself is congested. ECMP does not have a congestion awareness mechanism; it simply distributes flows to different links for forwarding, which may exacerbate congestion on links that are already congested.

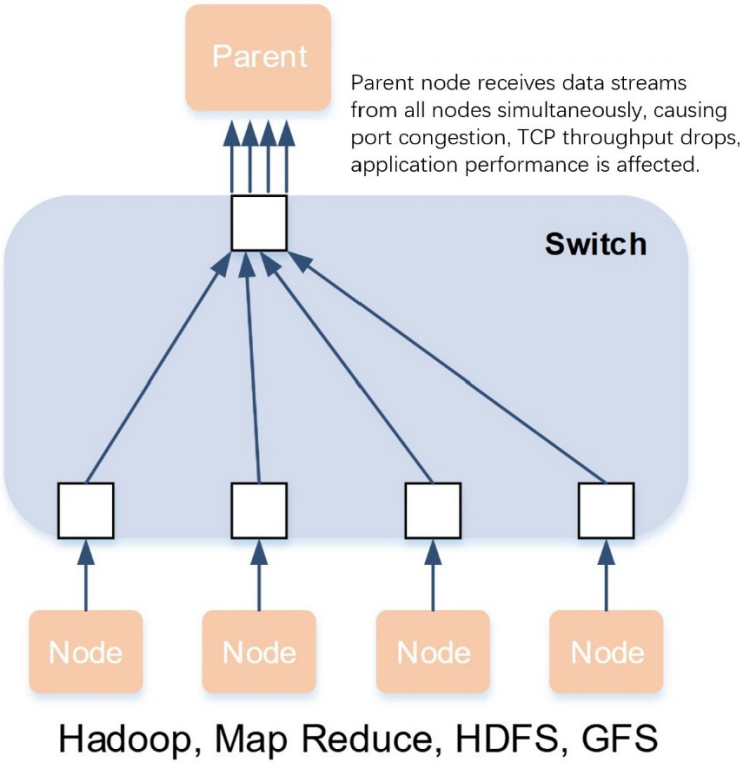
However, switches deployed in NXON.AI AIDCs support more advancing load-balancing technics then ECMP, such as AILB and RALB.

- TCP Incast

TCP Incast is a Many-to-One communication model. Under the major trend of data center cloudification, this communication model often occurs, especially in distributed storage and computing applications implemented in a Scale-Out manner, including Hadoop, MapReduce, HDFS, etc.

As shown in Figure 2-6, when a Parent Server sends a request to a group of nodes (server cluster or storage cluster), all nodes in the cluster receive the request simultaneously and respond almost at the same time. Many nodes send TCP data streams to one machine (Parent Server) simultaneously, creating a "micro-burst flow," causing insufficient buffer in the switch port connected to the Parent Server, leading to congestion.

Figure 2-6 TCP Incast Traffic Model



As mentioned earlier, RDMA is different from TCP; it requires a lossless network. For ordinary micro-burst traffic, the switch's Buffer can play a role, queuing the burst packets in the buffer. However, because increasing switch Buffer capacity is very costly, its effect is limited. Once the queued packets in the buffer become too many, packet loss will still occur.

To achieve end-to-end lossless forwarding and avoid packet loss caused by Buffer overflow in switches, the switch must introduce other mechanisms, such as flow control. By controlling the traffic on the link, the pressure on the switch Buffer is reduced, thus avoiding packet loss.

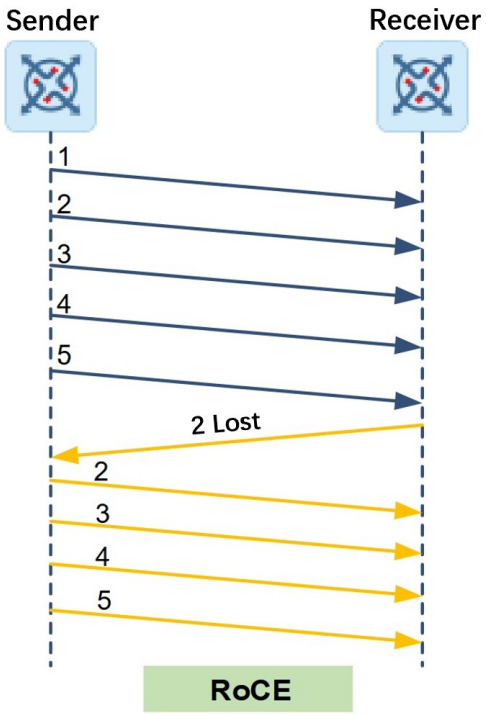
2.4.2 RoCE's Go-Back-N Retransmission

Packet loss will trigger packet retransmission. In RDMA, the cost of retransmission is huge. We can understand the hidden dangers brought by RoCE during retransmission by comparing the behavior of RoCE and iWARP during retransmission.

- iWARP is based on the TCP protocol, which has mechanisms like sliding windows and acknowledgment to achieve reliable transmission.

- RoCE is based on the UDP protocol, so reliable transmission can only be implemented at the application layer, significantly reducing efficiency. During retransmission, RoCE can only process packets in order. Therefore, the retransmission method it uses is the go-back-N method, meaning every packet after the lost packet sequence number must be retransmitted. As shown in Figure 2-7, the sender sends 5 data packets in total. Assuming the packet with sequence number 2 is lost, and the other packets are received normally by the receiver. In iWARP, the receiver will request retransmission of the packet with sequence number 2. In RoCE, all packets starting from sequence number 2 will be retransmitted, even though packets 3, 4, and 5 were not lost.

Figure 2-7 RoCE's Go-Back-N Retransmission



Thus, it is significant for RDMA work efficiency. Common situations that may cause RoCE retransmission include: ACK timeout, out-of-order packets, and receiving NAK packets.

- ACK Timeout

The RDMA receiver or responder confirms the requester's read/write requests through ACK

packets. If an ACK packet is not received for a long time, it is assumed that packet loss or congestion has occurred, triggering retransmission.

- Out-of-Order

RDMA does not support out-of-order reassembly like TCP. Therefore, once received packets are out of order, retransmission is also triggered.

- Receiving NAK

There are many situations that can cause receiving a NAK packet; both packet loss and out-of-order are triggering reasons. Additionally, when the responder queue is temporarily unable to process the request packet, it will send an RNR NAK (Receiver Not Ready NAK) to the requester. In summary, receiving a NAK indicates an abnormality in RDMA data transmission. Therefore, monitoring the number of NAK packets in the network can help operations personnel understand the operational status of the RDMA network.

2.4.3 Building a Lossless Network

The stability and reliability of network structure directly affect RDMA performance. Building a lossless network is necessary to maximize RDMA's advantages. The key to building a lossless network is congestion avoidance and operations monitoring.

1. Congestion Avoidance

Congestion avoidance primarily relies on two network flow control technologies: PFC (Priority-based Flow Control) and ECN (Explicit Congestion Notification). PFC operates by having a switch monitor its buffer for high-priority RDMA traffic. Once the buffer level hits a predefined high threshold (XOFF), the switch sends a "Pause Frame" command to the upstream device, forcing it to immediately halt transmission. This pause allows the switch to drain its buffer. Transmission automatically resumes once the buffer drains below a low threshold (XON), either because the pause timer expired or an explicit "resume" command is sent. This process ensures a lossless link for RDMA traffic.

ECN (Explicit Congestion Notification) is a proactive, end-to-end mechanism that manages network congestion before it becomes severe. Switches experiencing growing congestion mark packets—instead of dropping them—by setting a "Congestion Experienced" bit in their headers, aka CNP (Congestion Notification Packets) packet. The receiver relays this signal back to the sender, which then proactively reduces its transmission rate. This feedback loop allows the network to maintain high utilization while minimizing queue buildup and the need for drastic measures like PFC.

Table 2-3 Comparison between PFC and ECN

Feature	**PFC (Priority Flow Control)**	**ECN (Explicit Congestion Notification)**
Goal	**Prevent Packet Loss** (Create losslessness)	**Manage Congestion** (Maintain fairness & efficiency)
Scope	**Hop-by-Hop (Link-Level)**	**End-to-End**
Mechanism	**Reactive "Pause"** (Stop sending now!)	**Proactive "Slow Down"** (Reduce your rate)
Trigger	Buffer is **full** (XOFF threshold)	Buffer is **filling** (Marking threshold)
Analogy	A traffic cop stopping cars at a jammed intersection.	A GPS rerouting you due to reported traffic ahead.

2. Operations Monitoring

From the introduction above, we know that CNP packets and NAK packets reflect network congestion status and RDMA operational status, respectively. Therefore, we use CNP packets and NAK packets as statistical objects for RDMA.

- How to Match Packets

Figure 2-8 shows the format of a CNP packet. Its characteristics are: packet type is UDP, destination port number is 4791, and the Opcode field in the RDMA packet's BTH is 0x81.

Figure 2-8 Matching CNP Packet

Figure 2-9 shows a NAK packet. It is an RDMA UDP packet with destination port 4791, where the Opcode field in the BTH is 0x11, and the Syndrome field in the AETH is 0x60.

NOTE: AETH is a type of packet header used for ACK packets. The Syndrome field in the AETH is used to indicate whether the requested operation was successful.

Figure 2-9 Matching NAK Packet

Use ACL to match the CNP/NAK packet characteristic rules, and then set the action to Counter to obtain statistical values. Based on the CNP/NAK packet characteristics described above, the matching rules are summarized in Table 2-3.

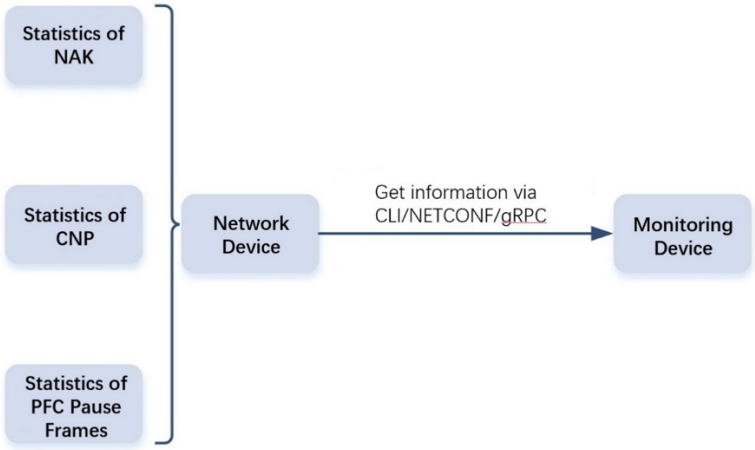
Table 2-4 CNP/NAK Matching Rules Summary

Feature	CNP (Congestion Notification Packet)	NAK (Negative Acknowledgment)
Match Object	Physical Port + Aggregate Port	Physical Port + Aggregate Port
Matching Rules	**Packet Type:** UDP (IPv4) **Destination Port:** 4791 **BTH Opcode Field:** 0x81	**Packet Type:** UDP (IPv4) **Destination Port:** 4791 **BTH Opcode Field:** 0x11 **AETH Syndrome Field:** 0x60
Match Action	Counter	Counter

- Statistics and Reporting

The RDMA statistics data model is shown in Figure 2-10. After the device obtains NAK and CNP packet statistics, users can retrieve the statistical information via CLI, NETCONF, or gRPC for analysis.

Figure 2-10 RDMA Statistics Data Model



In addition to CNP and NAK packets, PFC Pause frames can also reflect network congestion status. Operations systems often also count the number of PFC Pause frames in the network.

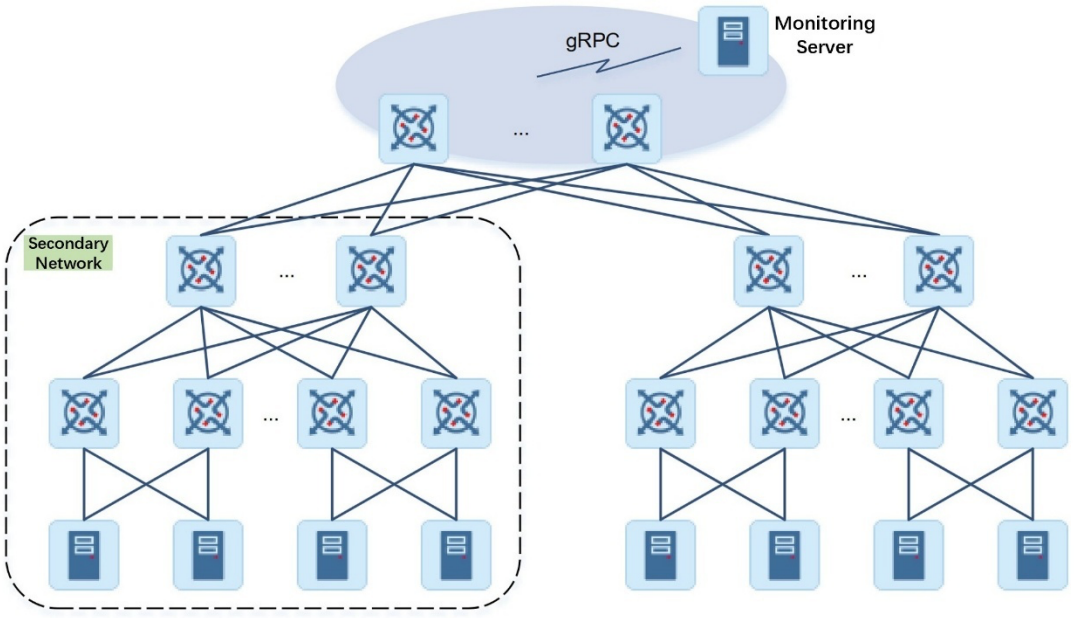
Typical Applications

3.1 RDMA Solving Network Performance Issues in HPC Scenarios

3.1.1 Scenario Introduction

HPC (High Performance Computing) utilizes aggregated computing power to process complex scientific computing problems that standard workstations cannot handle, including simulation, modeling, and rendering in scientific research and industry. It integrates the computing power of multiple units and distributes data and operations accordingly to solve the computational bottlenecks of a single server node. The demands of HPC on the network are: high throughput and low latency.

Figure 3-1 RDMA Application Scenario Network Diagram



3.1.2 Solution

(1) To avoid packet loss caused by factors like multi-level flow control scheduling latency during congestion, it is recommended to adopt a two-tier network for the RDMA service deployment POD. Simultaneously, it is recommended to select single-chip products for the equipment selection.

(2) In business scenarios where multiple services coexist, i.e., RDMA coexists with traditional IP services, network nodes need to classify services into different QoS queues based on service

type and corresponding DSCP values. Use SP+DRR congestion management to ensure RDMA traffic is scheduled with priority.

(3) Deploy PFC function at the traffic scheduling ingress to notify the peer to stop sending during ingress congestion.

(4) Deploy ECN function at the traffic scheduling egress to notify the server via ECN marking during egress congestion.

(5) The monitoring server monitors CNP packets, which are feedback from RDMA servers regarding congestion. Deploy ERSPAN on the devices to capture these packets.

(6) Each node device sends device status information, including cache status, congestion status, etc., via the gRPC channel.