

H3C 交换机

与服务器对接操作指导

资料版本：6W100-20230207

Copyright © 2023 新华三技术有限公司 版权所有，保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

除新华三技术有限公司的商标外，本手册中出现的其它公司的商标、产品标识及商品名称，由各自权利人拥有。

本文档中的信息可能变动，恕不另行通知。

前言

本文档主要用来介绍产品与服务器的对接场景，以及对接参数的配置，指导用户完成对接操作。前言部分包含如下内容：

- [读者对象](#)
- [本书约定](#)
- [文档使用前提](#)
- [资料意见反馈](#)

读者对象






本手册主要适用于如下工程师：

- 具有一定网络技术基础的网络规划人员
- 负责网络配置和维护，且具有一定网络技术基础的网络管理员

本书约定





1. 各类标志









本书采用各种醒目标志来表示在操作过程中应该特别注意的地方，这些标志的意义如下：

 警告	该标志后的注释需给予格外关注，不当的操作可能会对人身造成伤害。
 注意	提醒操作中应注意的事项，不当的操作可能会导致数据丢失或者设备损坏。
 提示	为确保设备配置成功或者正常工作而需要特别关注的操作或信息。
 说明	对操作内容的描述进行必要的补充和说明。
 窍门	配置、操作、或使用设备的技巧、小窍门。

2. 图标约定

本书使用的图标及其含义如下：

	该图标及其相关描述文字代表一般网络设备，如路由器、交换机、防火墙等。
	该图标及其相关描述文字代表一般意义下的路由器，以及其他运行了路由协议的设备。
	该图标及其相关描述文字代表二、三层以太网交换机，以及运行了二层协议的设备。
	该图标及其相关描述文字代表无线控制器、无线控制器业务板和有线无线一体化交换机的无线控制引擎设备。

	该图标及其相关描述文字代表无线接入点设备。
	该图标及其相关描述文字代表无线终结单元。
	该图标及其相关描述文字代表无线终结者。
	该图标及其相关描述文字代表无线Mesh设备。
	该图标代表发散的无线射频信号。
	该图标代表点到点的无线射频信号。
	该图标及其相关描述文字代表防火墙、UTM、多业务安全网关、负载均衡等安全设备。
	该图标及其相关描述文字代表防火墙插卡、负载均衡插卡、NetStream插卡、SSL VPN插卡、IPS插卡、ACG插卡等安全插卡。

3. 示例约定

由于设备型号不同、配置不同、版本升级等原因，可能造成本手册中的内容与用户使用的设备显示信息不一致。实际使用中请以设备显示的内容为准。

本手册中出现的端口编号仅作参考，并不代表设备上实际具有此编号的端口，实际使用中请以设备上存在的端口编号为准。

文档使用前提

本文档不严格与具体软、硬件版本对应，如果使用过程中与产品实际情况有差异，请以设备实际情况为准。

本文档中的配置均是在实验室环境下进行的配置和验证，配置前设备的所有参数均采用出厂时的缺省配置。如果您已经对设备进行了配置，为了保证配置效果，请确认现有配置和本文档中举例的配置不冲突。

资料意见反馈

如果您在使用过程中发现产品资料的任何问题，可以通过以下方式反馈：

E-mail: info@h3c.com

感谢您的反馈，让我们做得更好！

目 录

1 与服务器对接操作指导	1
1.1 与 Linux 服务器 Bonding 对接操作指导	1
1.1.1 Bonding 工作模式简介	1
1.1.2 互通性分析	1
1.1.3 配置指导	1
1.1.4 与 Linux 服务器 Bonding 对接案例（采用模式 1）	2
1.1.5 与 Linux 服务器 Bonding 对接案例（采用模式 4）	5
1.2 与 Linux 服务器 LLDP/DCBX 对接操作指导	10
1.2.1 互通性分析	10
1.2.2 组网需求	11
1.2.3 配置步骤	11
1.2.4 验证配置	16
1.3 与 BMP 服务器对接操作指导	18
1.3.1 BMP 简介	18
1.3.2 互通性分析	18
1.3.3 组网需求	18
1.3.4 配置步骤	18
1.3.5 验证配置	20

1 与服务器对接操作指导

1.1 与Linux服务器Bonding对接操作指导

1.1.1 Bonding 工作模式简介

Linux Bonding 提供了 7 种工作模式，不同模式具有不同的流量分担及链路备份策略：

- 模式 0：轮循均衡模式，按顺序依次在成员端口间发送数据包，能够实现负载均衡。
- 模式 1：主备模式，只有一个设备处于活动状态，当且仅当活动端口故障时另一个端口才转为主设备。
- 模式 2：异或均衡模式，通过计算公式 $[(\text{报文源 MAC XOR 报文目的 MAC}) \% \text{成员数}]$ 来决定从哪个端口发送报文。
- 模式 3：广播模式，该机制要求数据包向每个成员端口均发送一份。
- 模式 4：动态聚合模式，成员设备根据 802.3ad 协议决定本端状态，成员必须具有相同的双工、速率，同时需要对端设备支持 802.3ad。
- 模式 5：发送负载均衡模式，根据端口的发送利用率来决定从哪个端口发送报文。
- 模式 6：负载均衡模式，根据端口发送及接收利用率来决定从哪个端口发送报文。

1.1.2 互通性分析

表1 与 Linux 服务器 Bonding 对接互通性分析

H3C	Linux 服务器	互通结论
无需配置	模式1	可以互通
配置动态聚合模式	模式4	可以互通



说明

H3C 推荐您采用模式 1 或模式 4。

1.1.3 配置指导

1. ARP/ND

服务器网卡使用 bond1 工作模式时，服务器网卡需要支持接口发生 inactive-active 状态变化时，发送 ARP/ND 给接入交换机以便交换机刷新 ARP/ND 表项并生成主机路由。

服务器网卡使用 bond4 工作模式时，当服务器侧聚合组任一成员口发生 down->up 的状态变化时，都要发送 ARP/ND 给接入交换机以便交换机刷新 ARP/ND 表项并生成主机路由。

2. FEC 设置

FEC (forward error correction, 前向纠错) 在数据发送端为数据报文附加纠错信息, 接收端利用纠错信息纠正数据报文传输过程中产生的误码。该技术可以有效降低信道误码率, 提高信号质量, 从而延伸物理介质的最远传输距离, 但也会带来一些传输延时。如果两端的 FEC 模式不匹配, 则物理链路无法连通, 所以如果网卡和交换机的 FEC 模式不匹配, 请按如下步骤配置 FEC 模式:

- (1) 查看当前网口的支持的 FEC 模式, 执行 `ethtool --show-fec <网口名>`
- (2) 如需修改网口的 FEC 模式, 执行 `ethtool --set-fec <网口名> encoding off/baser/rs/auto` (配置立即生效, 重启后恢复)
- (3) 如需修改网口的 FEC 模式且重启后生效, 可修改 `rc.local` 文件。编辑 `/etc/rc.d/rc.local` 文件, 写入 shell 命令:
 - `ethtool --set-fec <网口名> encoding off/baser/rs/auto`
 - 启动 rc-local 服务: `systemctl enable rc-local`
 - 重启服务器



说明

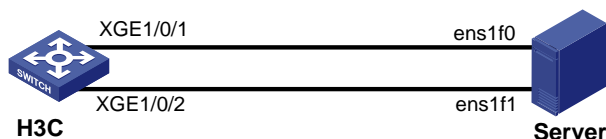
关于服务器的 ARP/ND 和 FEC 设置的详细说明, 请参见对应服务器的用户手册。

1.1.4 与 Linux 服务器 Bonding 对接案例 (采用模式 1)

1. 组网需求

如图 1 所示, Linux 服务器的两个网卡连接至交换机的不同接口。用户希望服务器和交换机对接的网卡形成主备, 当一个宕掉另一个可以由备份转换为主设备。

图1 与 Linux 服务器 Bonding 对接配置组网图 (采用模式 1)



2. 配置思路

- 网卡的 `ens7f3` 用作管理口, `ens1f0` 和 `ens1f1` 配置 bond, 采用模式 1。
- 服务器直装 Linux 系统, 如果是基于 VMware ESXI 上安装 Linux 虚拟机, 逻辑上 Linux 服务器的网口并非直接与设备相连, 而是与 VMware ESXI 上创建的 vswitch 相连, 达不到预期结果。
- 交换机侧无需配置。

3. 配置步骤

- 配置服务器

服务器的具体信息如下:

项目	描述
服务器型号	H3C R4900 G5

项目	描述
操作系统	内核版本: Linux version 4.18.0-305.25.1 操作系统版本: CentOS Linux release 8.4.2105
网卡型号	18:00.0 Ethernet controller: Mellanox Technologies MT2894 Family [ConnectX-6 Lx] 18:00.1 Ethernet controller: Mellanox Technologies MT2894 Family [ConnectX-6 Lx]
网卡驱动版本	MLNX_OFED_LINUX-5.4-3.2.7.2.3-rhel8.4-x86_64
网卡固件版本	driver: mlx5_core version: 5.4-3.2.7.2.3 firmware-version: 26.31.2006 (MT_0000000531) expansion-rom-version: bus-info: 0000:18:00.0 bus-info: 0000:18:00.1
依赖包	yum -y install zlib-devel bzip2-devel yum -y install openssl-devel ncurses-devel sqlite-devel readline-devel tk-devel gdbm-devel db4-devel libpcap-devel xz-devel --skip-broken yum -y install createrepo pciutils gcc gcc-c++ flex bison yum -y install gtk2 atk cairo tcl tcsh tk yum -y install tcl tcsh gcc-gfortran tk python36 perl yum -y install -y kernel-modules-extra yum remove pcp-pmda-infiniband

创建一个虚拟网卡 bond0。创建文件 ifcfg-bond0，保存退出。

```
[root@server4 /] vim /etc/sysconfig/network-scripts/ifcfg-bond0
:wq
```

编辑文件 ifcfg-bond0，写入网卡配置，保存退出。

```
vim /etc/sysconfig/network-scripts/ifcfg-bond0
BONDING_OPTS="mode=1 miimon=100 updelay=100 downdelay=100"
TYPE=Bond
BONDING_MASTER=yes
PROXY_METHOD=none
BROWSER_ONLY=no
BOOTPROTO=none
IPADDR=55.50.129.129
PREFIX=25
GATEWAY=55.50.129.252
DEFROUTE=no
IPV4_FAILURE_FATAL=no
IPV6INIT=yes
IPV6_AUTOCONF=no
IPV6_DEFROUTE=yes
IPV6_FAILURE_FATAL=no
IPV6_PRIVACY=no
IPV6_ADDR_GEN_MODE=stable-privacy
```

```
IPV6ADDR=200::5/64
IPV6_DEFAULTGW=200::1
NAME=bond0
DEVICE=bond0
ONBOOT=yes
```



bond 口 IP 为数据网 IP。

编辑文件 `ifcfg-ens1f0`，写入网卡配置，保存退出。

```
[root@server4 /]# more ifcfg-ens1f0
DEVICE=ens1f0
TYPE=Ethernet
ONBOOT=yes
SLAVE=yes
MASTER=bond0
BOOTPROTO=none
```

编辑文件 `ifcfg-ens1f1`，写入网卡配置，保存退出。

```
[root@server4 /]# more ifcfg-ens1f1
DEVICE=ens1f1
TYPE=Ethernet
ONBOOT=yes
SLAVE=yes
MASTER=bond0
BOOTPROTO=none
```

重启网络服务。

```
# ifdown bond0
# ifup bond0
```

4. 验证配置

查看 `bond0` 状态，可以看到其工作在主备模式下，有两个成员端口。

```
[root@server4 /]# cat /proc/net/bonding/bond0
Ethernet Channel Bonding Driver: v4.18.0-305.25.1.el8_4.x86_64

Bonding Mode: fault-tolerance (active-backup)
Primary Slave: None
Currently Active Slave: ens1f0
MII Status: up
MII Polling Interval (ms): 100
Up Delay (ms): 100
Down Delay (ms): 100
Peer Notification Delay (ms): 0

Slave Interface: ens1f0
MII Status: up
Speed: 25000 Mbps
Duplex: full
```



```
Link Failure Count: 0
Permanent HW addr: 10:70:fd:7f:da:a6
Slave queue ID: 0
```

```
Slave Interface: ens1f1
MII Status: up
Speed: 25000 Mbps
Duplex: full
Link Failure Count: 0
Permanent HW addr: 10:70:fd:7f:da:a7
Slave queue ID: 0
```

在 H3C 交换机上查看入流量。

```
<H3C> display counters rate inbound interface
Usage: Bandwidth utilization in percentage
Interface          Usage (%)   Total (pps)  Broadcast (pps)  Multicast (pps)
XGE1/0/1           25         2825519     --              --
XGE1/0/2           0          0           --              --
```

在 H3C 交换机上 shutdown 端口 Ten-GigabitEthernet1/0/1 后，再次查看入流量，发现流量已经切换到端口 Ten-GigabitEthernet1/0/2。

```
<H3C> display counters rate inbound interface
Usage: Bandwidth utilization in percentage
Interface          Usage (%)   Total (pps)  Broadcast (pps)  Multicast (pps)
XGE1/0/1           0          0           --              --
XGE1/0/2           100        2825703     --              --
```

在 H3C 交换机上恢复端口 Ten-GigabitEthernet1/0/1 后，再次查看入流量，因服务器工作在主备模式下，Ten-GigabitEthernet1/0/1 端口 down 后，ens1f0 自动设置为备设备，所以流量仍然在端口 Ten-GigabitEthernet1/0/2 上。

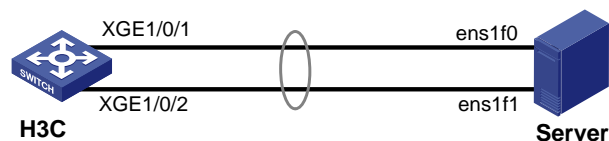
```
<H3C> display counters rate inbound interface
Usage: Bandwidth utilization in percentage
Interface          Usage (%)   Total (pps)  Broadcast (pps)  Multicast (pps)
XGE1/0/1           0          0           --              --
XGE1/0/2           100        2825508     --              --
```

1.1.5 与 Linux 服务器 Bonding 对接案例（采用模式 4）

1. 组网需求

如图 2 所示，Linux 服务器的两个网卡连接至交换机的不同接口。用户希望服务器和交换机对接时能够通过链路聚合提高接口利用率，实现负载均衡。

图2 与 Linux 服务器 Bonding 对接配置组网图（采用模式 4）



2. 配置思路

- 网卡的 ens7f3 用作管理口，ens1f0 和 ens1f1 配置 bond，采用模式 4。

- 服务器直装 Linux 系统，如果是基于 VMware ESXI 上安装 Linux 虚拟机，逻辑上 Linux 服务器的网口并非直接与设备相连，而是与 VMware ESXI 上创建的 vswitch 相连，达不到预期结果。
- 交换机侧配置动态聚合。

3. 配置步骤

- 配置交换机

创建二层聚合接口 1，并配置该接口为动态聚合模式。

```
<H3C> system-view
[H3C] interface bridge-aggregation 1
[H3C-Bridge-Aggregation1] link-aggregation mode dynamic
```

配置端口为边缘端口

```
[H3C-Bridge-Aggregation1] stp edged-port
[H3C-Bridge-Aggregation1] quit
```

分别将端口 Ten-GigabitEthernet1/0/1 和 Ten-GigabitEthernet1/0/2 加入到聚合组 1 中。

```
[H3C] interface Ten-GigabitEthernet 1/0/1
[H3C-Ten-GigabitEthernet1/0/1] port link-aggregation group 1
[H3C-Ten-GigabitEthernet1/0/1] quit
[H3C] interface ten-gigabitEthernet 1/0/2
[H3C-Ten-GigabitEthernet1/0/2] port link-aggregation group 1
[H3C-Ten-GigabitEthernet1/0/2] quit
```

- 配置服务器

服务器的具体信息如下：

项目	描述
服务器型号	H3C R4900 G5
操作系统	内核版本：Linux version 4.18.0-305.25.1 操作系统版本：CentOS Linux release 8.4.2105
网卡型号	18:00.0 Ethernet controller: Mellanox Technologies MT2894 Family [ConnectX-6 Lx] 18:00.1 Ethernet controller: Mellanox Technologies MT2894 Family [ConnectX-6 Lx]
网卡驱动版本	MLNX_OFED_LINUX-5.4-3.2.7.2.3-rhel8.4-x86_64
网卡固件版本	driver: mlx5_core version: 5.4-3.2.7.2.3 firmware-version: 26.31.2006 (MT_0000000531) expansion-rom-version: bus-info: 0000:18:00.0 bus-info: 0000:18:00.1
依赖包	yum -y install zlib-devel bzip2-devel yum -y install openssl-devel ncurses-devel sqlite-devel readline-devel tk-devel gdbm-devel db4-devel libpcap-devel xz-devel --skip-broken yum -y install createrepo pciutils gcc gcc-c++ flex bison yum -y install gtk2 atk cairo tcl tcsh tk yum -y install tcl tcsh gcc-gfortran tk python36 perl

项目	描述
	<pre> yum -y install -y kernel-modules-extra yum remove pc-pmda-infiniband </pre>

创建一个虚拟网卡 **bond0**。创建文件 **ifcfg-bond0**，保存退出。

```
[root@server4 /] vim /etc/sysconfig/network-scripts/ifcfg-bond0
:wq
```

编辑文件 **ifcfg-bond0**，写入网卡配置，保存退出。

```
vim /etc/sysconfig/network-scripts/ifcfg-bond0
BONDING_OPTS="mode=4 miimon=100 updelay=100 downdelay=100 xmit_hash_policy=layer3+4"
TYPE=Bond
BONDING_MASTER=yes
PROXY_METHOD=none
BROWSER_ONLY=no
BOOTPROTO=static
IPADDR=55.50.129.129
PREFIX=25
GATEWAY=55.50.129.252
DEFROUTE=no
IPV4_FAILURE_FATAL=no
IPV6INIT=yes
IPV6_AUTOCONF=no
IPV6_DEFROUTE=yes
IPV6_FAILURE_FATAL=no
IPV6_PRIVACY=no
IPV6_ADDR_GEN_MODE=stable-privacy
IPV6ADDR=200::5/64
IPV6_DEFAULTGW=200::1
NAME=bond0
DEVICE=bond0
ONBOOT=yes
```



说明

bond 口 IP 为数据网 IP。

编辑文件 **ifcfg-ens1f0**，写入网卡配置，保存退出。

```
[root@server4 /]# more ifcfg-ens1f0
DEVICE=ens1f0
TYPE=Ethernet
ONBOOT=yes
SLAVE=yes
MASTER=bond0
BOOTPROTO=none
```

编辑文件 **ifcfg-ens1f1**，写入网卡配置，保存退出。

```
[root@server4 /]# more ifcfg-ens1f1
```

```

DEVICE=ens1f1
TYPE=Ethernet
ONBOOT=yes
SLAVE=yes
MASTER=bond0
BOOTPROTO=none
# 重启网络服务。
# ifdown bond0
# ifup bond0

```

4. 验证配置

在交换机上查看聚合状态。

```

<H3C> display link-aggregation verbose
Loadsharing Type: Shar -- Loadsharing, NonS -- Non-Loadsharing
Port Status: S -- Selected, U -- Unselected, I -- Individual
Port: A -- Auto port, M -- Management port, R -- Reference port
Flags:  A -- LACP_Activity, B -- LACP_Timeout, C -- Aggregation,
        D -- Synchronization, E -- Collecting, F -- Distributing,
        G -- Defaulted, H -- Expired

```

```

Aggregate Interface: Bridge-Aggregation1
Creation Mode: Manual
Aggregation Mode: Dynamic
Loadsharing Type: Shar
Management VLANs: None
System ID: 0x6e, 2001-0000-0018

```

```

Local:
  Port                Status  Priority Index   Oper-Key      Flag
  XGE1/0/1(R)        S       32768  16391   40101        {ACDEF}
Remote:
  Actor                Priority Index   Oper-Key SystemID      Flag
  XGE1/0/2             255     1       21       0xffff, 1070-fd7f-dac2 {ABCDEF}

```

在服务器上查看 **bond0** 状态，可以看到其工作在动态聚合模式下，有两个成员端口，并携带 LACP 信息。

```

[root@server4/]# cat /proc/net/bonding/bond0
Ethernet Channel Bonding Driver: v4.18.0-305.25.1.el8_4.x86_64

```

```

Bonding Mode: IEEE 802.3ad Dynamic link aggregation
Transmit Hash Policy: layer3+4 (1)
MII Status: up
MII Polling Interval (ms): 100
Up Delay (ms): 100
Down Delay (ms): 100
Peer Notification Delay (ms): 0

```

```

802.3ad info
LACP rate: fast
Min links: 0

```

Aggregator selection policy (ad_select): stable
System priority: 65535
System MAC address: 10:70:fd:7f:da:a6
Active Aggregator Info:
 Aggregator ID: 1
 Number of ports: 2
 Actor Key: 21
 Partner Key: 40204
 Partner Mac Address: 20:01:00:00:00:03

Slave Interface: enslf0
MII Status: up
Speed: 25000 Mbps
Duplex: full
Link Failure Count: 1
Permanent HW addr: 10:70:fd:7f:da:a6
Slave queue ID: 0
Aggregator ID: 1
Actor Churn State: none
Partner Churn State: none
Actor Churned Count: 0
Partner Churned Count: 0
details actor lacp pdu:
 system priority: 65535
 system mac address: 10:70:fd:7f:da:a6
 port key: 21
 port priority: 255
 port number: 1
 port state: 63
details partner lacp pdu:
 system priority: 110
 system mac address: 20:01:00:00:00:03
 oper key: 40204
 port priority: 32768
 port number: 16391
 port state: 61

Slave Interface: enslf1
MII Status: up
Speed: 25000 Mbps
Duplex: full
Link Failure Count: 4
Permanent HW addr: 10:70:fd:7f:da:a7
Slave queue ID: 0
Aggregator ID: 1
Actor Churn State: none
Partner Churn State: none
Actor Churned Count: 0

```

Partner Churned Count: 0
details actor lacp pdu:
  system priority: 65535
  system mac address: 10:70:fd:7f:da:a6
  port key: 21
  port priority: 255
  port number: 2
  port state: 63
details partner lacp pdu:
  system priority: 110
  system mac address: 20:01:00:00:00:03
  oper key: 40204
  port priority: 32768
port number: 32775
port state: 61

```

在 H3C 交换机上查看入流量。

```

<H3C> display counters rate inbound interface
Usage: Bandwidth utilization in percentage
Interface          Usage (%)   Total (pps)  Broadcast (pps)  Multicast (pps)
BAGG1              20         1011085     --              --
XGE1/0/1          20         1011085     --              --
XGE1/0/2          0          0           --              --

```

在 H3C 交换机上 shutdown 端口 Ten-GigabitEthernet1/0/1 后，再次查看入流量，发现流量已经切换到端口 Ten-GigabitEthernet1/0/2。

```

<H3C> display counters rate outbound interface
Usage: Bandwidth utilization in percentage
Interface          Usage (%)   Total (pps)  Broadcast (pps)  Multicast (pps)
BAGG1              99         2809534     --              --
XGE1/0/1          0          0           --              --
XGE1/0/2          99         2809534     --              --

```

在 H3C 交换机上恢复端口 Ten-GigabitEthernet1/0/1 后，再次查看入流量，流量切回到端口 Ten-GigabitEthernet1/0/1。

```

<H3C> display counters rate inbound interface
Usage: Bandwidth utilization in percentage
Interface          Usage (%)   Total (pps)  Broadcast (pps)  Multicast (pps)
BAGG1              20         1011085     --              --
XGE1/0/1          20         1011085     --              --
XGE1/0/2          0          0           --              --

```

1.2 与Linux服务器LLDP/DCBX对接操作指导

1.2.1 互通性分析

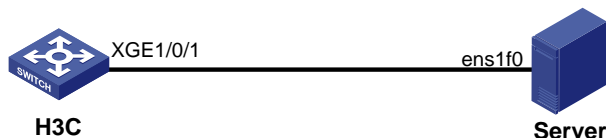
表2 与 Linux 服务器 LLDP/DCBX 对接互通性分析

H3C	Linux 服务器	互通结论
支持	支持	可以互通

1.2.2 组网需求

如图 3 所示，H3C 交换机通过接口 Ten-GigabitEthernet1/0/1 与服务器的网卡相连。用户希望交换机与服务器之间指定 802.1p 优先级的流量无丢包。

图3 与 Linux 服务器 LLDP/DCBX 对接案例



1.2.3 配置步骤

- 配置交换机

全局开启 LLDP 功能。

```
<H3C> system-view
[H3C] lldp global enable
```

在接口 Ten-GigabitEthernet1/0/1 上开启 LLDP 功能，并允许发布 DCBX TLV。

```
[H3C] interface Ten-GigabitEthernet 1/0/1
[H3C-Ten-GigabitEthernet1/0/1] lldp enable
[H3C-Ten-GigabitEthernet1/0/1] lldp tlv-enable dot1-tlv dcbx
```

在接口 Ten-GigabitEthernet1/0/1 上配置优先级信任模式为信任报文自带的 dscp 优先级。

```
[H3C-Ten-GigabitEthernet1/0/1] qos trust dscp
```

在接口 Ten-GigabitEthernet1/0/1 上配置 PFC 功能的开启模式为自动协商模式，并对 802.1p 优先级 5 开启 PFC 功能。

```
[H3C-Ten-GigabitEthernet1/0/1] priority-flow-control auto
[H3C-Ten-GigabitEthernet1/0/1] priority-flow-control no-drop dot1p 5
```

- 配置服务器

服务器的具体信息如下：

项目	描述
服务器型号	H3C R4900 G5
操作系统	内核版本: Linux version 4.18.0-305.25.1 操作系统版本: CentOS Linux release 8.4.2105
网卡型号	18:00.0 Ethernet controller: Mellanox Technologies MT2894 Family [ConnectX-6 Lx] 18:00.1 Ethernet controller: Mellanox Technologies MT2894 Family [ConnectX-6 Lx]
网卡驱动版本	MLNX_OFED_LINUX-5.4-3.2.7.2.3-rhel8.4-x86_64
网卡固件版本	driver: mlx5_core version: 5.4-3.2.7.2.3 firmware-version: 26.31.2006 (MT_0000000531) expansion-rom-version:

项目	描述
	bus-info: 0000:18:00.0 bus-info: 0000:18:00.1
依赖包	<pre> yum -y install zlib-devel bzip2-devel yum -y install openssl-devel ncurses-devel sqlite-devel readline-devel tk-devel gdbm-devel db4-devel libpcap-devel xz-devel --skip-broken yum -y install createrepo pciutils gcc gcc-c++ flex bison yum -y install gtk2 atk cairo tcl tcsh tk yum -y install tcl tcsh gcc-gfortran tk python36 perl yum -y install -y kernel-modules-extra yum remove pcp-pmda-infiniband </pre>



说明

配置服务器前需要将服务器的 ECN 去使能，预防 ECN 在 PFC 生效前生效。

服务器使能 LLDP 和 DCBX

详细配置请参考: [Flow Control - MLNX_EN v5.4-3.6.8.1 LTS - NVIDIA Networking Docs](#)

启动 mst

```

[root@server4 /]# mst start
Starting MST (Mellanox Software Tools) driver set
Loading MST PCI module - Success
[warn] mst_pciconf is already loaded, skipping
Create devices
Unloading MST PCI module (unused) - Success

```

查看 mst 状态

```

[root@server4 /]# mst status
MST modules:
-----
    MST PCI module is not loaded
    MST PCI configuration module loaded

MST devices:
-----
/dev/mst/mt4119_pciconf0      - PCI configuration cycles access.
                               domain:bus:dev.fn=0000:4b:00.0 addr.reg=88 data.reg=92
cr_bar.gw_offset=-1
                               Chip revision is: 00
/dev/mst/mt4127_pciconf0    - PCI configuration cycles access.
                               domain:bus:dev.fn=0000:18:00.0 addr.reg=88 data.reg=92
cr_bar.gw_offset=-1
                               Chip revision is: 00

```

查看 lldp 和 dcbx 状态

```

[root@server4 mst]# mlxconfig -d /dev/mst/mt4127_pciconf0 q
Device #1:

```



```

-----
Device type:    ConnectX6LX
Name:          MCX631102AN-ADA_Ax
Description:    ConnectX-6 Lx EN adapter card; 25GbE ; Dual-port SFP28; PCIe 4.0 x8; No Crypto
Device:        /dev/mst/mt4127_pciconf0

```

```

Configurations:
MEMIC_BAR_SIZE          0
MEMIC_SIZE_LIMIT       _256KB(1)
LLDP_NB_DCBX_P1        False(0)
LLDP_NB_RX_MODE_P1     OFF(0)
LLDP_NB_TX_MODE_P1     OFF(0)
LLDP_NB_DCBX_P2        False(0)
LLDP_NB_RX_MODE_P2     OFF(0)
LLDP_NB_TX_MODE_P2     OFF(0)
DCBX_IEEE_P1           True(1)
DCBX_CEE_P1            True(1)
DCBX_WILLING_P1        True(1)
DCBX_IEEE_P2           True(1)
DCBX_CEE_P2            True(1)
DCBX_WILLING_P2        True(1)

```

修改 LLDP 和 DCBX 参数

```

[root@server4 /]# mlxconfig -d /dev/mst/mt4127_pciconf0 set LLDP_NB_DCBX_P1=TRUE
LLDP_NB_TX_MODE_P1=2 LLDP_NB_RX_MODE_P1=2 LLDP_NB_DCBX_P2=TRUE LLDP_NB_TX_MODE_P2=2
LLDP_NB_RX_MODE_P2=2

```

Device #1:

```

-----
Device type: ConnectX6LX
Name: MCX631102AN-ADA_Ax
Description: ConnectX-6 Lx EN adapter card; 25GbE ; Dual-port SFP28; PCIe 4.0 x8; No Crypto
Device: /dev/mst/mt4127_pciconf0

```

```

Configurations:
LLDP_NB_DCBX_P1        True(1)   True(1)
LLDP_NB_TX_MODE_P1     OFF(0)   ALL(2)
LLDP_NB_RX_MODE_P1     OFF(0)   ALL(2)
LLDP_NB_DCBX_P2        False(0)  True(1)
LLDP_NB_TX_MODE_P2     OFF(0)   ALL(2)
LLDP_NB_RX_MODE_P2     OFF(0)   ALL(2)

```

Apply new Configuration? (y/n) [n] : y

Applying... Done!

-I- Please reboot machine to load new configurations.

验证 LLDP 和 DCBX 修改成功

```

[root@server4 /]# mlxconfig -d /dev/mst/mt4127_pciconf0 q

```

Device #1:

```

-----
Device type:    ConnectX6LX
Name:          MCX631102AN-ADA_Ax
Description:    ConnectX-6 Lx EN adapter card; 25GbE ; Dual-port SFP28; PCIe 4.0 x8; No Crypto

```

Device: /dev/mst/mt4127_pciconf0

Configurations:	Next Boot
MEMIC_BAR_SIZE	0
MEMIC_SIZE_LIMIT	_256KB(1)
LLDP_NB_DCBX_P1	True(1)
LLDP_NB_RX_MODE_P1	ALL(2)
LLDP_NB_TX_MODE_P1	ALL(2)
LLDP_NB_DCBX_P2	True(1)
LLDP_NB_RX_MODE_P2	ALL(2)
LLDP_NB_TX_MODE_P2	ALL(2)
DCBX_IEEE_P1	True(1)
DCBX_CEE_P1	True(1)
DCBX_WILLING_P1	True(1)
DCBX_IEEE_P2	True(1)
DCBX_CEE_P2	True(1)
DCBX_WILLING_P2	True(1)

重启 firmware

```
[root@server4 /]# mlxfwreset -d /dev/mst/mt4127_pciconf0 --level 3 reset
```

Requested reset level for device, /dev/mst/mt4127_pciconf0:

3: Driver restart and PCI reset

Continue with reset?[y/N] y

-I- Sending Reset Command To Fw -Done

-I- Stopping Driver -Done

-I- Resetting PCI -Done

-I- Starting Driver -Done

-I- Restarting MST -Done

-I- FW was loaded successfully.

修改 DCBX 模式为 firmware

```
[root@server4 /]# mlnx_qos -i ens1f0 -d get
```

DCBX mode: Firmware controlled

Priority trust state: pcp

default priority:

Receive buffer size (bytes): 0,156096,0,0,0,0,0,0,

Cable len: 7

PFC configuration:

priority	0	1	2	3	4	5	6	7
enabled	0	0	0	0	0	0	0	0
buffer	1	1	1	1	1	1	1	1

tc: 1 ratelimit: unlimited, tsa: vendor

priority: 0

tc: 0 ratelimit: unlimited, tsa: vendor

priority: 1

tc: 2 ratelimit: unlimited, tsa: vendor

priority: 2

tc: 3 ratelimit: unlimited, tsa: vendor

```

        priority: 3
tc: 4 ratelimit: unlimited, tsa: vendor
        priority: 4
tc: 5 ratelimit: unlimited, tsa: vendor
        priority: 5
tc: 6 ratelimit: unlimited, tsa: vendor
        priority: 6
tc: 7 ratelimit: unlimited, tsa: vendor
        priority: 7
[root@server4 /]# mlnx_qos -i ens1f0 -d fw
DCBX mode: Firmware controlled
Priority trust state: pcp
default priority:
Receive buffer size (bytes): 0,156096,0,0,0,0,0,0,
Cable len: 7
PFC configuration:
        priority    0    1    2    3    4    5    6    7
        enabled      0    0    0    0    0    0    0    0
        buffer       1    1    1    1    1    1    1    1
tc: 1 ratelimit: unlimited, tsa: vendor
        priority: 0
tc: 0 ratelimit: unlimited, tsa: vendor
        priority: 1
tc: 2 ratelimit: unlimited, tsa: vendor
        priority: 2
tc: 3 ratelimit: unlimited, tsa: vendor
        priority: 3
tc: 4 ratelimit: unlimited, tsa: vendor
        priority: 4
tc: 5 ratelimit: unlimited, tsa: vendor
        priority: 5
tc: 6 ratelimit: unlimited, tsa: vendor
        priority: 6
tc: 7 ratelimit: unlimited, tsa: vendor
        priority: 7

```

修改 mlnx_qos 信任 dscp

```

[root@server4 /]# mlnx_qos -i ens1f0 --trust dscp
DCBX mode: Firmware controlled
Priority trust state: dscp
dscp2prio mapping:
        prio:0 dscp:07,06,05,04,03,02,01,00,
        prio:1 dscp:15,14,13,12,11,10,09,08,
        prio:2 dscp:23,22,21,20,19,18,17,16,
        prio:3 dscp:31,30,29,28,27,26,25,24,
        prio:4 dscp:39,38,37,36,35,34,33,32,
        prio:5 dscp:47,46,45,44,43,42,41,40,
        prio:6 dscp:55,54,53,52,51,50,49,48,
        prio:7 dscp:63,62,61,60,59,58,57,56,

```

```

default priority:
Receive buffer size (bytes): 0,156096,0,0,0,0,0,0,
Cable len: 7
PFC configuration:
    priority    0   1   2   3   4   5   6   7
    enabled     0   0   0   0   0   0   0   0
    buffer      1   1   1   1   1   1   1   1
tc: 1 ratelimit: unlimited, tsa: vendor
    priority: 0
tc: 0 ratelimit: unlimited, tsa: vendor
    priority: 1
tc: 2 ratelimit: unlimited, tsa: vendor
    priority: 2
tc: 3 ratelimit: unlimited, tsa: vendor
    priority: 3
tc: 4 ratelimit: unlimited, tsa: vendor
    priority: 4
tc: 5 ratelimit: unlimited, tsa: vendor
    priority: 5
tc: 6 ratelimit: unlimited, tsa: vendor
    priority: 6
tc: 7 ratelimit: unlimited, tsa: vendor
    priority: 7

```

1.2.4 验证配置

查看交换机侧 LLDP 邻居。

```

<H3C> display lldp neighbor-information list
Chassis ID : * -- -- Nearest nontpmr bridge neighbor
              # -- -- Nearest customer bridge neighbor
              Default -- -- Nearest bridge neighbor
Local Interface Chassis ID      Port ID      System Name
XGE1/0/1          ec0d-9ad4-48fa  ec0d-9ad4-48f8  -

```

查看服务器的 PFC 优先级自动协商到 5 队列。

```

[root@server4 ~]# mlnx_qos -i ens1f0
DCBX mode: Firmware controlled
Priority trust state: dscp
dscp2prio mapping:
    prio:0 dscp:07,06,05,04,03,02,01,00,
    prio:1 dscp:15,14,13,12,11,10,09,08,
    prio:2 dscp:23,22,21,20,19,18,17,16,
    prio:3 dscp:31,30,29,28,27,26,25,24,
    prio:4 dscp:39,38,37,36,35,34,33,32,
    prio:5 dscp:47,46,45,44,43,42,41,40,
    prio:6 dscp:55,54,53,52,51,50,49,48,
    prio:7 dscp:63,62,61,60,59,58,57,56,
default priority:
Receive buffer size (bytes): 20016,156096,0,0,0,0,0,0,

```

```

Cable len: 7
PFC configuration:
    priority    0  1  2  3  4  5  6  7
    enabled     0  0  0  0  0  1  0  0
    buffer      0  0  0  0  0  1  0  0
tc: 0 ratelimit: unlimited, tsa: ets, bw: 2%
    priority: 0
tc: 1 ratelimit: unlimited, tsa: ets, bw: 4%
    priority: 1
tc: 2 ratelimit: unlimited, tsa: ets, bw: 6%
    priority: 2
tc: 3 ratelimit: unlimited, tsa: ets, bw: 8%
    priority: 3
tc: 4 ratelimit: unlimited, tsa: ets, bw: 9%
    priority: 4
tc: 5 ratelimit: unlimited, tsa: ets, bw: 17%
    priority: 5
tc: 6 ratelimit: unlimited, tsa: ets, bw: 25%
    priority: 6
tc: 7 ratelimit: unlimited, tsa: ets, bw: 29%
    priority: 7

```

查看交换机侧的流量队列。

```

<H3C> display qos queue-statistics interface Ten-GigabitEthernet 1/0/1 outbound
Interface: Twenty-FiveGigE1/0/1
Direction: outbound
Forwarded: 24731576 packets, 26864670580 bytes
Dropped: 0 packets, 0 bytes
Queue 0
  Forwarded: 0 packets, 0 bytes, 0 pps, 0 bps
  Dropped: 0 packets, 0 bytes
  Current queue length: 0 packets
Queue 1
  Forwarded: 0 packets, 0 bytes, 0 pps, 0 bps
  Dropped: 0 packets, 0 bytes
  Current queue length: 0 packets
Queue 2
  Forwarded: 0 packets, 0 bytes, 0 pps, 0 bps
  Dropped: 0 packets, 0 bytes
  Current queue length: 0 packets
Queue 3
  Forwarded: 0 packets, 0 bytes, 0 pps, 0 bps
  Dropped: 0 packets, 0 bytes
  Current queue length: 0 packets
Queue 4
  Forwarded: 0 packets, 0 bytes, 0 pps, 0 bps
  Dropped: 0 packets, 0 bytes
  Current queue length: 0 packets
Queue 5

```

```

Forwarded: 24731572 packets, 26864670088 bytes, 2822493 pps, 24527467704 bps
Dropped: 0 packets, 0 bytes
Current queue length: 5 packets
Queue 6
Forwarded: 0 packets, 0 bytes, 0 pps, 0 bps
Dropped: 0 packets, 0 bytes
Current queue length: 0 packets
Queue 7
Forwarded: 4 packets, 492 bytes, 0 pps, 0 bps
Dropped: 0 packets, 0 bytes
Current queue length: 0 packets

```

1.3 与BMP服务器对接操作指导

1.3.1 BMP 简介

BGP 协议只能记录 BGP 会话和 BGP 路由的当前状态，无法直接收集到会话状态变化和路由更新的过程，通过配置 BMP（BGP Monitoring Protocol，BGP 监控协议）特性，监控服务器可以对网络中设备上 BGP 会话的运行状态进行实时监控，包括对等体关系的建立与解除、路由信息等，以方便网络管理员更加细致地了解 BGP 运行状况。

1.3.2 互通性分析

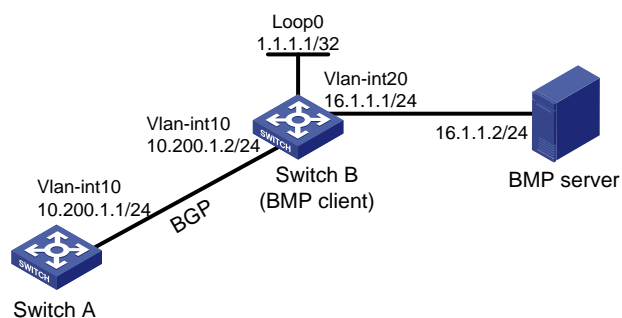
表3 互通性分析

H3C	BMP Server	互通结论
支持	支持	可以互通

1.3.3 组网需求

如图 4 所示，Switch A 和 Switch B 之间建立 BGP 会话，Switch B 上配置 BMP 功能对其 BGP 运行状态以及路由变化进行监控，并将 BMP 信息传递给 BMP Server。

图4 与 BMP 服务器对接配置组网图



1.3.4 配置步骤

- 配置 SwitchA

```

<SwitchA> system-view
[SwitchA] interface vlan-interface 10
[SwitchA-vlan-interface10] ip address 10.200.1.1 255.255.255.0
[SwitchA] bgp 45090
[SwitchA-bgp-default] peer 10.200.1.2 as-number 45090
[SwitchA-bgp-default] address-family ipv4 unicast
[SwitchA-bgp-default-ipv4] peer 10.200.1.2 enable
[SwitchA-bgp-default-ipv4] quit
[SwitchA-bgp-default] quit

```

- 配置 SwitchB

```

<SwitchB> system-view
[SwitchB] interface vlan-interface 10
[SwitchB-vlan-interface10] ip address 10.200.1.2 255.255.255.0
[SwitchB] bgp 45090
[SwitchB-bgp-default] peer 10.200.1.1 as-number 45090
[SwitchB-bgp-default] peer 10.200.1.1. bmp server 1
[SwitchB-bgp-default] address-family ipv4 unicast
[SwitchB-bgp-default-ipv4] peer 10.200.1.1 enable
[SwitchB-bgp-default-ipv4] quit
[SwitchB-bgp-default] quit
[SwitchB] bmp server 1
[SwitchB-bmpserver-1] server address 16.1.1.2 port 5000
[SwitchB-bmpserver-1] server connect-interface loopback0
[SwitchB-bmpserver-1] route-mode adj-rib-out
[SwitchB-bmpserver-1] route-mode loc-rib
[SwitchB-bmpserver-1] statistics-interval 10

```

- 配置 BMP Server

服务器的具体信息如下:

项目	描述
服务器型号	H3C R4900
宿主机系统	Vmware ESXi 6.0
虚拟机系统	linux 内核版本3.10.0-693.5.2.el7.x86_64 #1 SMP Fri Oct 20 20:32:50 UTC 2017 x86_64 x86_64 x86_64 GNU/Linux
CentOS版本	CentOS Linux release 7.4.1708 (Core)
BMP软件	openbmpd (www.openbmp.org) version : 0.14.0-0

配置 BMP Server 和 SwichB 之间路由可达 (略)。

#在 BMP Server 上安装 OPENBMP 软件, 具体步骤如下:

- 安装 ova 文件
- 运行 docker:

```

docker run -d --name=openbmp_aio \ -e KAFKA_FQDN=localhost \ -v
/var/openbmp/mysql:/data/mysql \ -v /var/openbmp/config:/config \ -p 3306:3306 -p 2181:2181
-p 9092:9092 -p 5000:5000 -p 8001:8001 \ openbmp/aio

```



说明

如上命令需要整理在一行内下发，不能有换行。

1.3.5 验证配置

在交换机上查看 BGP 监控服务器的信息。

```
<SwitchB> display bgp bmp server 1
BMP server number: 1
Server VPN instance name: --
Server address: 16.1.1.2 Server port: 5000
Client address: 16.1.1.1 Client port: 34481
BMP server state: Connected Up for 00h09m42s
TCP source interface has been configured
```

Message statistics:

```
Total messages sent: 285751
    INITIATION: 1
    TERMINATION: 0
    STATS-REPORT: 464
        PEER-UP: 15
        PEER-DOWN: 0
        ROUTE-MON: 285271
```

BMP monitor BGP peers:

```
10.200.1.1
```

在服务器上查看当 BGP 邻居建立、撤销时，BGP 监控服务器上收到的信息。

```
[root@openbmp ~]# docker exec openbmp_aio tail -f /var/log/openbmpd.log
```

```
2022-12-22T04:01:48.223803 | INFO      | runServer          | Initializing server
2022-12-22T04:01:49.328034 | INFO      | runServer          | Ready. Waiting for connections
2022-12-22T04:01:59.480328 | INFO      | runServer          | Accepted new connection; active
connections = 1
2022-12-22T04:01:59.480373 | INFO      | runServer          | Client Connected => 1.1.1.1|:12815,
sock = 10
2022-12-22T04:02:02.588530 | INFO      | ClientThread       | Thread started to monitor BMP
from router 1.1.1.1| using socket 10 buffer in bytes = 15728640
2022-12-22T04:02:02.589027 | INFO      | ReadIncomingMsg    | 1.1.1.1|: Init message received
with length of 208
2022-12-22T04:02:02.589083 | INFO      | handleInitMsg      | Init message type 1 and length
162 parsed
2022-12-22T04:02:02.589098 | INFO      | handleInitMsg      | Init message type 1 = H3C Comware
Platform Software, Software Version 7.1.070, Feature 2809
H3C S12508X-AF
Copyright (c) 2004-2021 New H3C Technologies Co., Ltd. All rights reserved.
2022-12-22T04:02:02.589107 | INFO      | handleInitMsg      | Init message type 2 and length
7 parsed
```



```

2022-12-22T04:02:02.589115 | INFO      | handleInitMsg      | Init message type 2 = kalia-2
2022-12-22T04:02:02.589123 | INFO      | handleInitMsg      | Init message type 0 and length
27 parsed
2022-12-22T04:02:02.589131 | INFO      | handleInitMsg      | Init message type 0 = bgp instance
name:
default
2022-12-22T04:02:02.589138 | INFO      | ReadIncomingMsg    | Router ID hashed with hash_type:
1

2022-12-22T04:02:19.431566 | INFO      | ReadIncomingMsg    | 1.1.1.1|: PEER UP Received, local
addr=:::0 remote addr=:::0 -----SwitchA 和 SwitchB 的bgp peer 建立
2022-12-22T04:02:19.431647 | NOTICE   | parsePeerUpInfo    | Peer info message type 0 is not
implemented
2022-12-22T04:02:19.431875 | INFO      | ReadIncomingMsg    | 1.1.1.1|: PEER UP Received, local
addr=10.200.1.2:179 remote addr=10.200.1.1:55674
2022-12-22T04:02:29.437774 | INFO      | ReadIncomingMsg    | 1.1.1.1|: PEER UP Received, local
addr=:::0 remote addr=:::0
2022-12-22T04:02:29.437845 | NOTICE   | parsePeerUpInfo    | Peer info message type 0 is not
implemented
2022-12-22T04:02:29.438906 | NOTICE   | parseAttr_AsPath   | 10.200.1.1 rtr=1.1.1.1|: Could
not parse the AS PATH due to update message buffer being too short when using ASN octet size
4 (4 > 2)
2022-12-22T04:02:29.438934 | NOTICE   | parseAttr_AsPath   | 10.200.1.1: rtr=1.1.1.1|:
switching encoding size to 2-octet
2022-12-22T04:02:46.083438 | NOTICE   | parsePeerDownEventHdr | sock=16 : 10.200.1.1: BGP
peer down notification with reason code: 3 -----SwitchA 和 SwitchB 的bgp peer 撤
销

```