

# 知 EVPN分布式网关组网部分leaf-6800开启dhcp enable (dhcp relay生效) 后服务器无法pxe装机

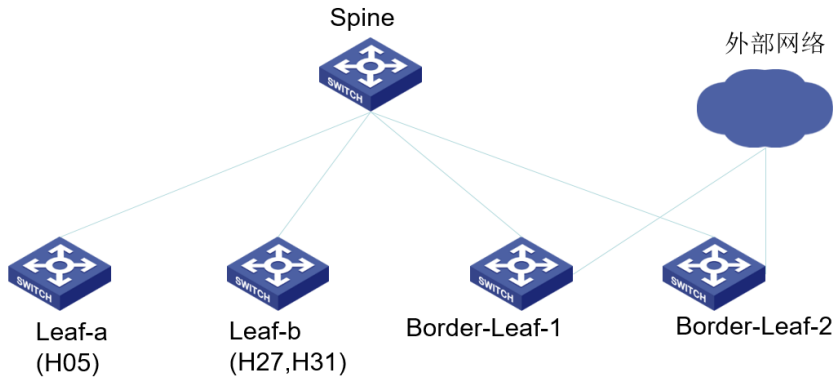
DHCP/DHCP Relay VxLAN EVPN 孙冰 2022-02-09 发表

## 组网及说明

设备: S6800-54QF

版本: R2612P02

组网: LEAF-a对应组网中原有leaf设备 (H05补丁或无补丁), LEAF-b对应新增leaf设备 (H27补丁或H31+H30补丁), leaf作为dhcp relay, 指定的dhcp服务器在外部网络。



Leaf配置如下:

```
#
dhcp enable
#
interface Vsi-interface10
ip binding vpn-instance 10
ip address 10.8.254.254 255.255.0.0
mac-address 0001-0001-0001
dhcp select relay
dhcp relay server-address 10.15.0.245
dhcp relay information enable (可选)
dhcp relay source-address interface LoopBack1
dhcp relay request-from-tunnel discard
distributed-gateway local
#
interface LoopBack0
ip address 10.15.0.32 255.255.255.255
#
interface LoopBack1
ip binding vpn-instance 10
ip address 10.8.255.32 255.255.255.255
#
```

## 问题描述

客户反馈leaf-b中H31+H30补丁的设备下连服务器pxe装机会失败，而leaf-b中H27补丁的设备和leaf-a的设备下连服务器pxe装机是可以成功的。失败后卡在如下界面：

```
Intel(R) Boot Agent XE v2.4.16
Copyright (C) 1997-2017, Intel Corporation

PXE-E61: Media test failure, check cable
PXE-M0F: Exiting Intel Boot Agent.

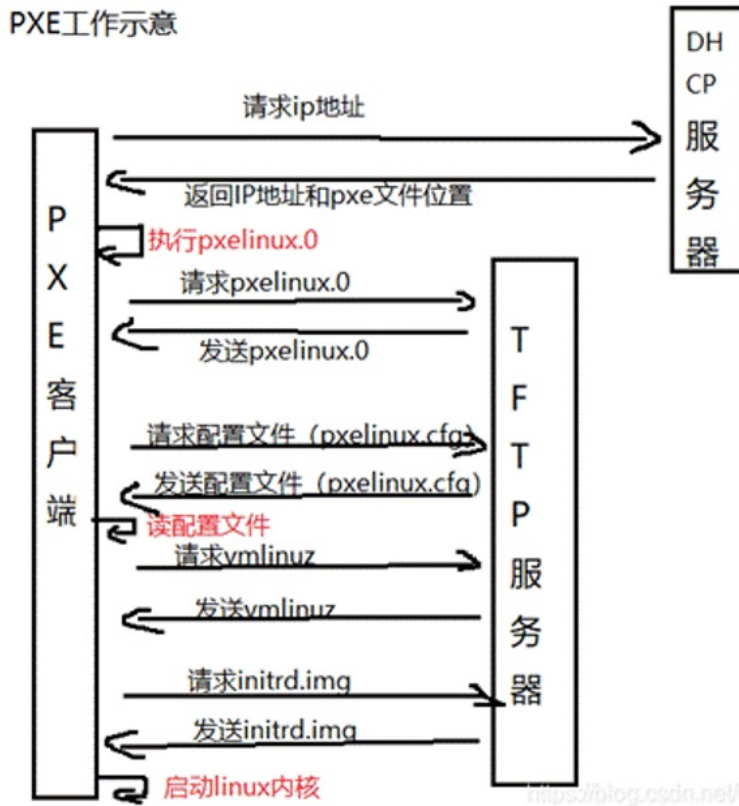
Booting from FlexBoot v3.5.210 (PCI 03:00.1)
FlexBoot v3.5.210 (PCI 03:00.1) starting execution...ok
FlexBoot initialising devices...
Initialising completed.

FlexBoot v3.5.210
Features: DNS HTTP iSCSI TFTP ULAN ELF MBOOT PXE bzImage COMBOOT Menu PXEXT
net0: ec:0d:9a:d3:00:b1
Using ConnectX-4Lx on 0000:03:00.1 (open)
  [Link:up, TX:0 TXE:0 RX:0 RXE:0]
Configuring (net0 ec:0d:9a:d3:00:b1)..... ok
net0: 10.8.252.163/255.255.0.0 gw 10.8.254.254
net0: fe80::ec0d:9aff:fed3:b1/64
Next server: 10.8.248.1
Filename: grub/grub-x86_64.efi
tftp://10.8.248.1/grub/grub-x86_64.efi... ok
grub-x86_64.efi : 243679 bytes [PXE-NBP (may be EFI?)]
```

正常PXE装机没有最后一句话直接就到装机界面了。

## 过程分析

首先了解下pxe装机的过程：



pxe过程与交换机的ztp类似，服务器主机上电开机，发送dhcp请求报文获取ip地址，dhcp服务器根据dhcp请求报文携带的信息回复dhcp offer给主机，提供pxe启动文件地址和目录信息（Legacy BIOS模式对应pxelinux.0文件，UEFI模式对应grub.efi文件），主机再通过tftp获取对应pxe文件完成自动装机。回到问题本身，根据客户反馈打了不同补丁的S6800在开启dhcp enable后，下连服务器可以dhcp获取地址，但pxe装机有不同的表现，

于是找了一台未上业务的leaf开启dhcp enable打不通补丁进行测试，测试结果如下：

- 1、H05，装机失败
- 2、H27，装机失败
- 3、H31+H30，装机失败

**发现装机结果与补丁版本无关，H27的设备测试现象与客户反馈不一致，检查配置发现之前装机成功的H27设备没有开启dhcp enable。**

于是继续测试：

- 1、H05,undo dhcp enable 装机成功
- 2、H27,undo dhcp enable 装机成功
- 3、H31+H30,undo dhcp enable装机成功

**以上测试确认，leaf下联服务器pxe装机成功与否，与leaf是否开启dhcp enable有关。但是组网中确实有一部分leaf开启dhcp enable时下联服务器是可以装机成功的，与客户沟通装机失败的服务器与成功的服务器之间有何区别，客户确认后反馈，失败的服务器均采用了第三方Mellanox网卡接入。**

**协调测试发现，如果换到服务器自带网卡，不管接入leaf是否开启dhcp enable，服务器都可以装机成功，故可以确定，Mellanox网卡是导致问题的根因，接下来需要通过抓包进一步确定导致该现象的具体原因。**

leaf的dhcp enable未开启，dhcp relay就不会生效，说明dhcp请求报文广播后通过其他leaf的网关发送到了另外的某个提供dhcp服务的服务器。

抓包分析如下，

(1) dhcp enable时，Mellanox网卡主机装机成功，通过抓包发现，有两个不同源ip的dhcp offer报文，从dhcp relay配置看10.15.0.245是指定的dhcp服务器地址，10.8.248.1是dhcp server侧配置文件（dhcpd.conf）指定的tftp服务器地址，即option next server ip address。但10.8.248.1也作为dhcp服务器地址回复了dhcp报文，说明10.8.248.1这台服务器既开启了tftp服务作为tftp server，同时也开启了dhcp服务作为另外一台dhcp server。从dhcp server 10.15.0.245回给dhcp relay 10.8.255.32的单播dhcp offer报文中，携带的boot file 引导文件信息是grub/grub-x86\_64.efi。而从tftp server 10.8.248.1广播过

来的dhcp offer报文中，携带的boot file 引导文件信息是pxelinux.0。

No.	Time	Source	Destination	Protocol	Length	Time to Live	Info
2918	2021/1/30 10:36:55.449726	0.0.0.0	255.255.255.255	DHCP	436	64	DHCP Discover - Transacti
2922	2021/1/30 10:36:55.453712	10.15.0.245	10.8.255.32	DHCP	488	254,253	DHCP Offer - Transacti
2923	2021/1/30 10:36:55.454987	10.8.254.254	255.255.255.255	DHCP	342	255	DHCP Offer - Transacti
2924	2021/1/30 10:36:55.455024	10.8.254.254	255.255.255.255	DHCP	342	255	DHCP Offer - Transacti
142.	2021/1/30 10:36:56.450163	10.8.248.1	255.255.255.255	DHCP	342	128	DHCP Offer - Transacti
142.	2021/1/30 10:36:56.450222	10.8.248.1	255.255.255.255	DHCP	392	254,128	DHCP Offer - Transacti
142.	2021/1/30 10:36:56.450236	10.8.248.1	255.255.255.255	DHCP	342	128	DHCP Offer - Transacti
145.	2021/1/30 10:36:59.505330	0.0.0.0	255.255.255.255	DHCP	436	64	DHCP Discover - Transacti
145.	2021/1/30 10:36:59.505394	0.0.0.0	255.255.255.255	DHCP	436	64	DHCP Discover - Transacti
145.	2021/1/30 10:36:59.505408	0.0.0.0	255.255.255.255	DHCP	436	64	DHCP Discover - Transacti
145.	2021/1/30 10:36:59.505538	10.8.248.1	255.255.255.255	DHCP	392	254,128	DHCP Offer - Transacti
145.	2021/1/30 10:36:59.511322	10.8.254.254	255.255.255.255	DHCP	342	255	DHCP Offer - Transacti

1. 暂时无法大规模替换所有主机的问题网卡，为核项目进度，先通过leat先关闭dhcpenable，让主机走BIOS模式完成pxe装机。  
2. 后续排查dhcp服务有问题的dhcp server。

```
> Frame 2922: 488 bytes on wire (3264 bits), 488 bytes captured (3264 bits) on interface \Device\NPF_{F1F9564F-7C40-40EF-8EF1-70E544A8D2AC}, id 0
> Ethernet II, Src: NewH3CTe_36:da:01 (88:df:9e:36:da:01), Dst: NewH3CTe_65:6d:4b (74:ea:cb:65:6d:4b)
> Internet Protocol Version 4, Src: 10.15.0.2, Dst: 10.15.0.32
> User Datagram Protocol, Src Port: 36397, Dst Port: 4789
> Virtual extensible Local Area Network
> Ethernet II, Src: NewH3CTe_e0:b6:34 (3c:f5:cc:e0:b6:34), Dst: NewH3CTe_65:6d:4b (74:ea:cb:65:6d:4b)
> Internet Protocol Version 4, Src: 10.15.0.245, Dst: 10.8.255.32
> User Datagram Protocol, Src Port: 67, Dst Port: 67
v Dynamic Host Configuration Protocol (Offer)
2794 2021/1/30 10:36:50.431078 10.15.0.32 10.15.0.1 BOOTP 229 255 Unknown BOOTP message type
2916 2021/1/30 10:36:55.449664 0.0.0.0 255.255.255.255 DHCP 436 64 DHCP Discover - Transacti
2917 2021/1/30 10:36:55.449714 0.0.0.0 255.255.255.255 DHCP 436 64 DHCP Discover - Transacti
2918 2021/1/30 10:36:55.449726 0.0.0.0 255.255.255.255 DHCP 436 64 DHCP Discover - Transacti
2922 2021/1/30 10:36:55.453712 10.15.0.245 10.8.255.32 DHCP 488 254,253 DHCP Offer - Transacti
2923 2021/1/30 10:36:55.454987 10.8.254.254 255.255.255.255 DHCP 342 255 DHCP Offer - Transacti
2924 2021/1/30 10:36:55.455024 10.8.254.254 255.255.255.255 DHCP 342 255 DHCP Offer - Transacti
142. 2021/1/30 10:36:56.450163 10.8.248.1 255.255.255.255 DHCP 342 128 DHCP Offer - Transacti
142. 2021/1/30 10:36:56.450222 10.8.248.1 255.255.255.255 DHCP 392 254,128 DHCP Offer - Transacti
142. 2021/1/30 10:36:56.450236 10.8.248.1 255.255.255.255 DHCP 342 128 DHCP Offer - Transacti
145. 2021/1/30 10:36:59.505330 0.0.0.0 255.255.255.255 DHCP 436 64 DHCP Discover - Transacti
145. 2021/1/30 10:36:59.505394 0.0.0.0 255.255.255.255 DHCP 436 64 DHCP Discover - Transacti
145. 2021/1/30 10:36:59.505408 0.0.0.0 255.255.255.255 DHCP 436 64 DHCP Discover - Transacti
145. 2021/1/30 10:36:59.505528 10.8.248.1 255.255.255.255 DHCP 342 128 DHCP Offer - Transacti
145. 2021/1/30 10:36:59.505538 10.8.248.1 255.255.255.255 DHCP 392 254,128 DHCP Offer - Transacti
> Internet Protocol Version 4, Src: 10.15.0.3, Dst: 10.15.0.32
> User Datagram Protocol, Src Port: 63383, Dst Port: 4789
> Virtual extensible Local Area Network
> Ethernet II, Src: Dell_j1:7f:70 (e4:43:4b:1f:7f:70), Dst: Broadcast (ff:ff:ff:ff:ff:ff)
> Internet Protocol Version 4, Src: 10.8.248.1, Dst: 255.255.255.255
> User Datagram Protocol, Src Port: 67, Dst Port: 68
> Dynamic Host Configuration Protocol (Offer)
```

```
Next server IP address: 10.8.248.1
Relay agent IP address: 0.0.0.0
Client MAC address: Mellanox_d3:00:b5 (ec:0d:9a:d3:00:b5)
Client hardware address padding: 000000000000000000
Server host name not given
Boot file name: pxelinux.0
Magic cookie: DHCP
> Option: (53) DHCP Message Type (Offer)
> Option: (54) DHCP Server Identifier (10.8.248.1)
> Option: (51) IP Address Lease Time
```

之后查看后续的tftp报文交互，分到地址的待pxe装机的主机同时得到了grub-x86\_64.efi和pxelinux.0的信息，tftp优先请求了grub-x86\_64.efi文件进行后续的EFI模式的pxe装机，最后的结果是pxe装机失败。

No.	Time	Source	Destination	Proto	Length	Time to Live	Info
520.	2021/1/30 10:37:23.738052	10.8.252.181	10.8.248.1	TFTP	92	64	Read Request, File: grub/grub-x86_64.efi, Transfer type: octet, bl
520.	2021/1/30 10:37:23.738058	10.8.252.181	10.8.248.1	TFTP	142	255,64	Read Request, File: grub/grub-x86_64.efi, Transfer type: octet, bl
521.	2021/1/30 10:37:23.738054	10.8.248.1	10.8.252.181	TFTP	120	254,64	Option Acknowledgement, blksize=1380, tsize=243679
521.	2021/1/30 10:37:23.738020	10.8.248.1	10.8.252.181	TFTP	70	64	Option Acknowledgement, blksize=1380, tsize=243679
521.	2021/1/30 10:37:23.738001	10.8.252.181	10.8.248.1	TFTP	60	64	Acknowledgement, Block: 0
521.	2021/1/30 10:37:23.738008	10.8.252.181	10.8.248.1	TFTP	110	255,64	Acknowledgement, Block: 0
521.	2021/1/30 10:37:23.738038	10.8.248.1	10.8.252.181	TFTP	1476	254,64	Data Packet, Block: 1
521.	2021/1/30 10:37:23.738044	10.8.248.1	10.8.252.181	TFTP	1426	64	Data Packet, Block: 1
521.	2021/1/30 10:37:23.738052	10.8.252.181	10.8.248.1	TFTP	60	64	Acknowledgement, Block: 1
521.	2021/1/30 10:37:23.738058	10.8.252.181	10.8.248.1	TFTP	110	255,64	Acknowledgement, Block: 1
521.	2021/1/30 10:37:23.738088	10.8.248.1	10.8.252.181	TFTP	1476	254,64	Data Packet, Block: 2
521.	2021/1/30 10:37:23.738095	10.8.248.1	10.8.252.181	TFTP	1426	64	Data Packet, Block: 2
521.	2021/1/30 10:37:23.738004	10.8.252.181	10.8.248.1	TFTP	60	64	Acknowledgement, Block: 2

通过dhcp服务器10.8.0.245侧的配置可以看出，dhcp服务器是将dhcp discover报文中的option60字段的pxe-system-type也即option93作为判断条件，分配给待pxe装机的主机不同的引导文件的：pxe-system-type为0006, 0007, 0009时分配grub/grub-x86.efi, 0002时分配ia64/elilo.efi, 其他情况分配pxelinux.0文件。通过抓包看mellanox网卡的主机发出的dhcp discover的option 93对应的是0000, dhcp offer报文应该携带是pxelinux.0参数，但实际上却是错误的回复了grub/grub-x86.efi，导致主机进行了efi模式装机最终失败。

分析怀疑mellanox网卡不支持efi模式的pxe装机，其次10.8.0.245这台服务器的dhcp服务功能有问题。

```
subnet 10.8.0.0 netmask 255.255.0.0 {
    option routers          10.8.254.254;
    option domain-name-servers 8.8.8.8;
    option subnet-mask     255.255.0.0;
    range dynamic-bootp   10.8.250.0 10.8.253.250;
    filename                "/pxlinux.0";
    default-lease-time     21600;
```