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

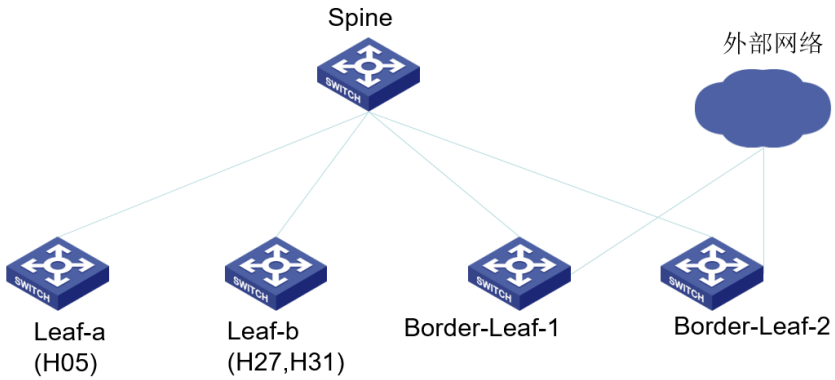
DHCP/DHCP Relay　VxLAN　EVPN　孙冰　2022-02-09 发表

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。

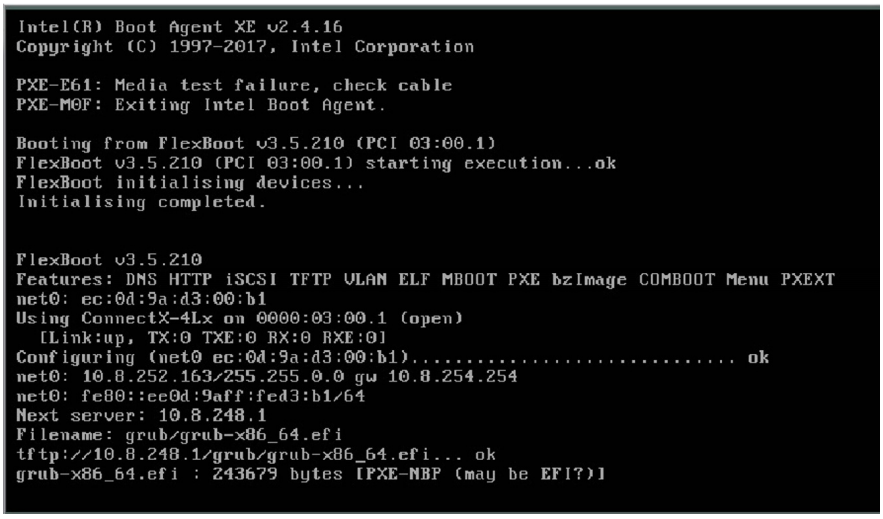


Leaf配置如下：

```
#
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

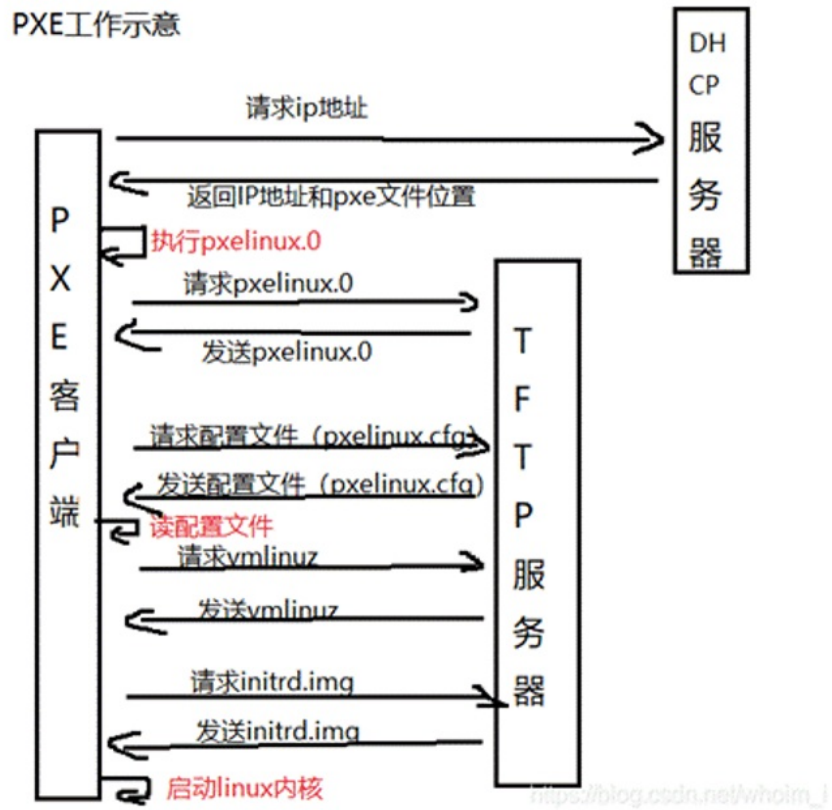
客户反馈leaf-b中H31+H30补丁的设备下连服务器pxe装机会失败，而leaf-b中H27补丁的设备和leaf-a的设备下连服务器pxe装机是可以成功的。失败后卡在如下界面：



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。**

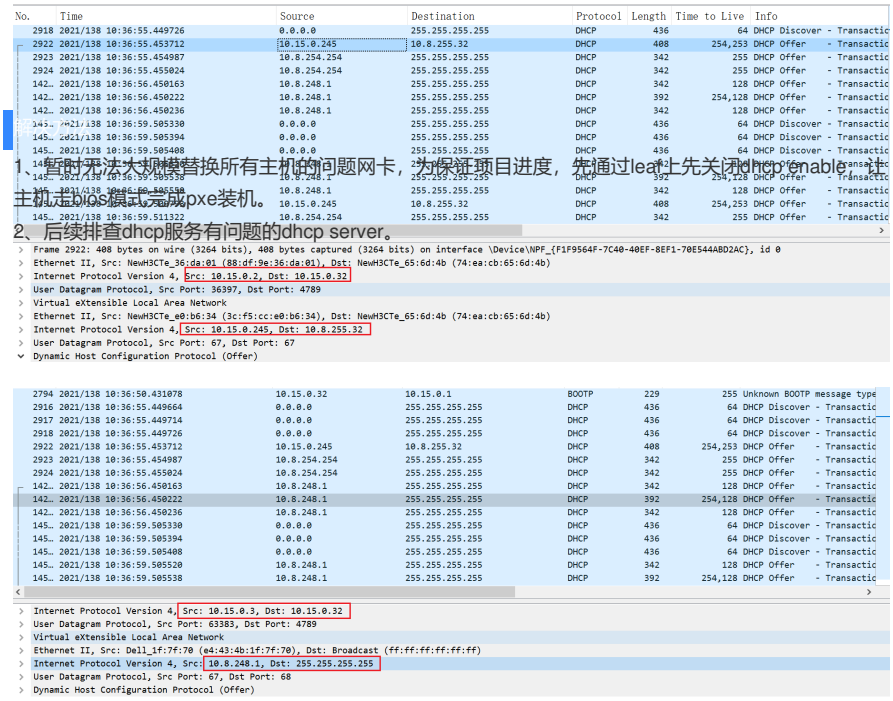
协调测试发现，如果换到服务器自带网卡，不管接入leaf是否开启dhcp enable，服务器都可以装机成功，故可以确定， Mellanox**网卡是导致问题的根因，接下来需要通过抓包进一步确定导致该现象的具体原因。**

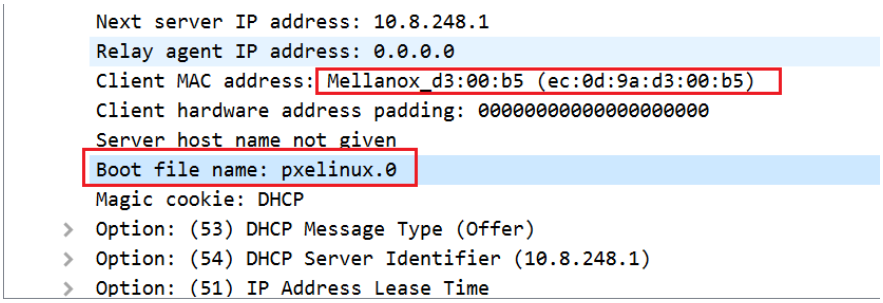
leaf的dhcp enable未开启，dhcp relay就不会生效，说明dhcp 请求报文广播后通过其他leaf的网关发送到了另外的某个提供dhcp服务的服务器。

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

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



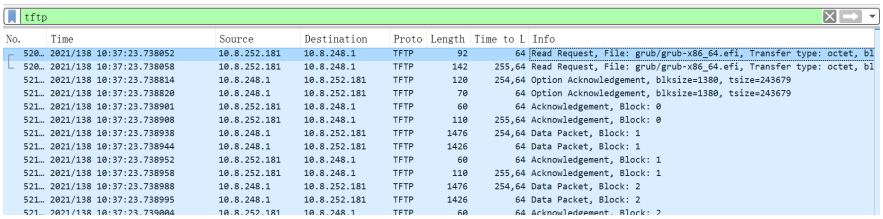


1、暂时无法大规模替换所有主机的问题网卡，为保证项目进度，先通过leaf上先关闭dhcp relay，让主机走bios模式完成pxe装机。

2、后续排查dhcp服务有问题的dhcp server。

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

。



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optin 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                "/pxelinux.0";
    default-lease-time         21600;
```