跨 DC 间 VXLAN 互联的 MTU 规划建议

文档版本 01

发布日期 2020-03-24

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： https://e.huawei.com

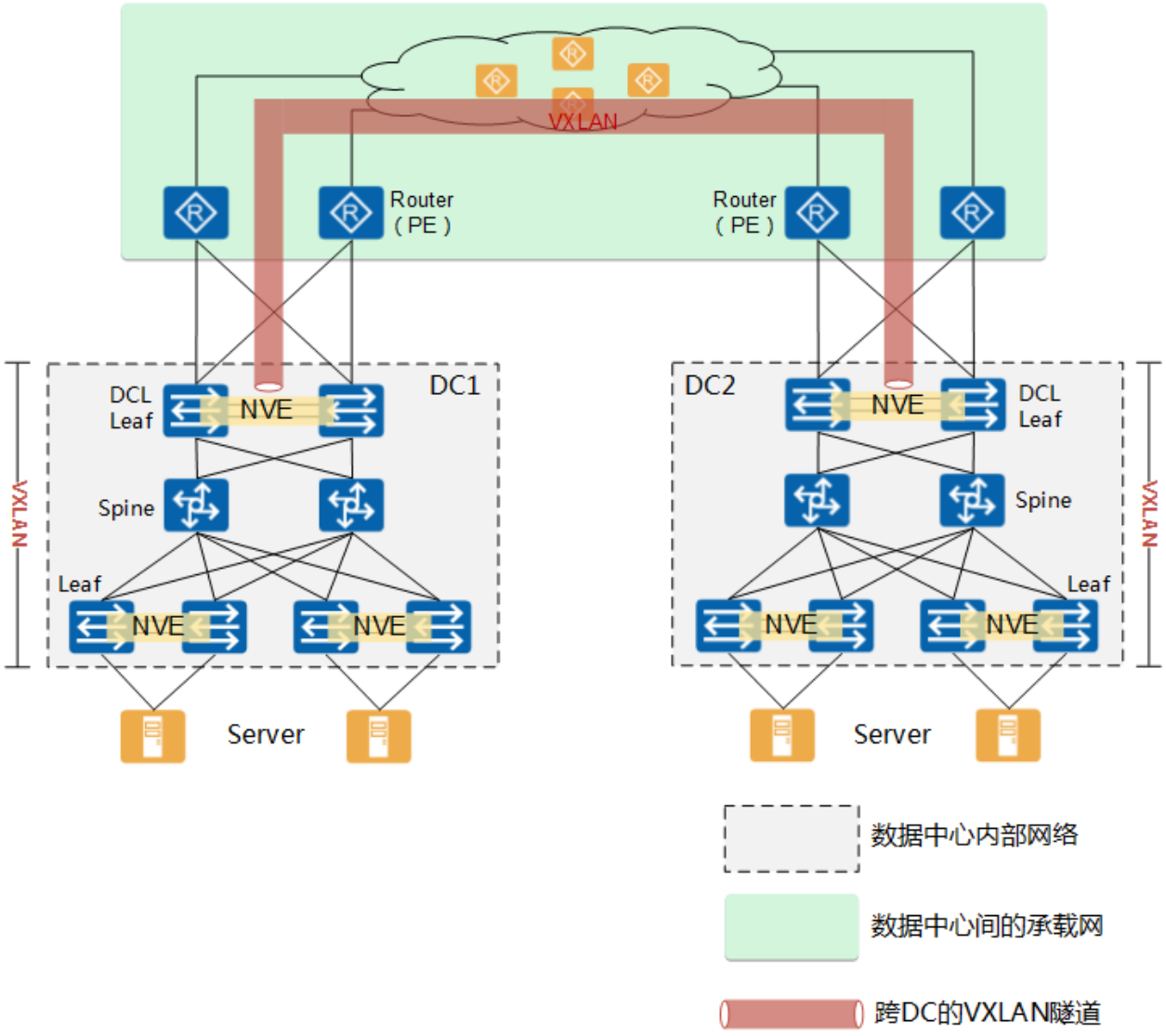
目 录

1 跨 DC 间 VXLAN 互联的 MTU 规划建议

1.1 场景说明

两个数据中心，通过广域承载网络互联，实现服务器之间跨DC的通信。其中灰色部分是DC内部网络，由CloudEngine系列交换机（下文简称CE交换机）组成，部署VXLAN网络。DC1的DCI Leaf与DC2的DCI Leaf之间，建立跨承载网（图中绿色底纹所示）的VXLAN隧道，转发跨DC流量。此时需要端到端进行MTU的规划，否则报文在跨DC转发过程中容易丢包。

图 1-1 两个数据中心互连示意图



1.2 DC 内部对报文的处理：对转发报文不分片

CE交换机不会对长度超过接口MTU值的转发报文进行分片，即服务器发出的报文，经过DC内部交换机转发时，尽管报文长度超过了CE交换机的MTU，也不会分片。

CE默认支持的JUMBO帧为9216字节。极端情况下，服务器侧MTU设置为9000字节（一般默认为1500字节），加上VXLAN封装的50字节（20字节IP头+8字节UDP头 + 8字节VXLAN头 + 14字节MAC头）后，CE交换机也能支持正常的报文转发（且不分片）。

说明

CE交换机只对本机发出的**主机报文**（例如协议类报文）进行基于接口MTU值的分片。

1.3 为什么要规划 MTU?

在跨DC的VXLAN互联时，VXLAN报文需要经过第三方承载网络（如上图绿色部分，承载网络仅做Underlay转发）。由于封装VXLAN后的报文比承载网络中某些设备的MTU值大（中间的设备有时不允许调整MTU值，或者有的设备较老，MTU值较小），导致在Underlay转发时，VXLAN报文被分片。根据VXLAN 7348 RFC，VXLAN报文不建议分片，否则在接收端VTEP上会丢弃分片的报文，造成跨DC通信的丢包。

RFC参考：**https://datatracker.ietf.org/doc/rfc7348/?include_text=1**

```
VTEPs MUST NOT fragment VXLAN packets.  Intermediate routers may
fragment encapsulated VXLAN packets due to the larger frame size.
The destination VTEP MAY silently discard such VXLAN fragments.  To
ensure end-to-end traffic delivery without fragmentation, it is
RECOMMENDED that the MTUs (Maximum Transmission Units) across the
physical network infrastructure be set to a value that accommodates
the larger frame size due to the encapsulation.  Other techniques
like Path MTU discovery (see [RFC1191] and [RFC1981]) MAY be used to
address this requirement as well.
```

1.4 规划建议

综上所示，建议在部署前就对MTU进行全局规划，有如下建议，其中推荐第一条。

1. 方式一（推荐）：修改应用层服务器发送报文的长度值，修改后的长度值加上VXLAN封装的50字节后，需保证在整个承载网中，均小于设备的MTU值。使用此方法，修改难度低，需要IT侧配合。
2. 方式二：修改承载网中每一跳网络设备的MTU值，需保证MTU值大于收到的VXLAN报文长度，从而保证不分片（建议承载网中设备设置的MTU值需大于CE交换机的默认JUMBO的大小，即9216字节）。此方式常常受到约束：承载网络中设备众多、分布广泛，且涉及不同厂商，修改难度大；常常没有修改权限（他人资产，不可控）。

