R5300 G6搭配NIC-BCM957412-F-B-10Gb-2P做网络拥塞测试限速

物理服务器 张玉林 2024-11-07 发表

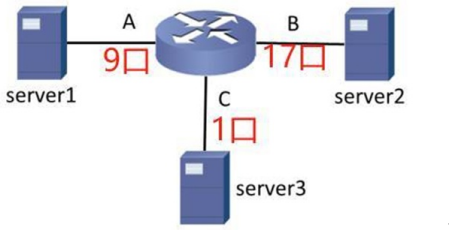
问题描述

2台服务器400G口向一台服务器400G打流，在没开ECN的情况下，2台发送端会自动降速到200G左右

过程分析

测试过程：

（1）按照如图组网搭建测试环境，服务器网卡支持RDMA并使能ECN，交换机接口C在队列N使能ECN并配置ECN水线



（2）服务器server1、server2向server3打对应交换机N队列的RoCE流量，构造交换机接口C拥塞，预期交换机接口C有ECN计数，无丢包计数

（3）停止流量，清除ECN计数，交换机接口C去使能ECN后，再重复步骤2的操作，预期交换机接口C没有ECN计数，有丢包计数

异常点：

（3）步骤清除ECN配置，三个出口ECN标记位报文没有涨，但是服务器速率降速200G

打流方式：

（1） 在C服务器开两个端口接收A和B的流量

ib_send_bw -d mlx5_3 --report_gbits -F --run_infinitely -R -T 128 -p 5000

ib_send_bw -d mlx5_3 --report_gbits -F --run_infinitely -R -T 128 -p 6000

（2） A和B向C开始打流

A：ib_send_bw -d mlx5_0 --report_gbits -F --run_infinitely -R -T 128 -p 5000 IPA

B：ib_send_bw -d mlx5_1 --report_gbits -F --run_infinitely -R -T 128 -p 6000 IPB

解决方法

CX7单400G网卡，使用IB打流（rdma协议），存在buffer确认机制，在服务器进行队列4的2打1流量时，会使网卡自动降速。

备注：ROCEv2使用UDP协议将RDMA数据封装在UDP报文中，ROCE流量在UDP协议后面的RDMA上存在确认机制，该确认是靠RDMA这一层来完成的。网卡侧的buffer确认机制需要等待接收端回复确认后才会继续发送，服务器上的buffer跟交换机无关，当中就算没有交换机或者经过多台交换机，它最终也只考虑两个对端的服务器或者三个对端的服务器。