一、相关资源

Trex官方网站支持的100G网卡为ConnectX-5, ConnectX-5 驱动下载地址：https://www.mellanox.com/products/infiniband-drivers/linux/mlnx\_ofed

Trex相关文档: https://trex-tgn.cisco.com/trex/doc/

Trex对ConnectX-4/5系列网卡支持相关文档：

<https://trex-tgn.cisco.com/trex/doc/trex_appendix_mellanox.html>

OFED驱动下载: http://www.mellanox.com/page/products\_dyn?product\_family=26&mtag=linux\_sw\_drivers

二、Trex官方对linux支持版本信息

如下2个版本经官方测试并通过：

We tested the following distro with TRex and OFED. Others might work too.

CentOS 7.6 (This is the only verified distro) - up to v2.87

CentOS 7.9 (This is the only verified distro) - v2.88 and up

经过测试但失败的Linux发行版本：

Following distros were tested and did not work for us in the past (with older OFEDs).

Fedora 21 (3.17.4-301.fc21.x86\_64)

Ubuntu 14.04.3 LTS (GNU/Linux 3.19.0-25-generic x86\_64)?—?crash when RSS was enabled MLX RSS issue

三、经过测试在byzoro验证的linux发行版和使用的软硬件性能：

CentOS 7.6， trex v2.87, Intel(R) Xeon(R) Platinum 8160 CPU @ 2.10GHz 双CPU 96核心, 256G 内存，PCIeGen3 X 8 (单向最大65.5G)

CentOS 7.9， trex v2.87, Intel(R) Xeon(R) Gold 6132 CPU @ 2.60GHz双CPU 56核心， 256G内存PCIeGen3 X 16 (单向最大99.6G)

OFED版本：MLNX\_OFED\_LINUX-5.2-2.2.3.0

网卡: ConnectX-5 100G单口或者双口网卡.

四、BIOS 调优

经过测试在byzoro现有的硬件条件下未使用到如下BIOS调优方法，仅在性能达不到要求是谨慎使用.

遵循了解BIOS配置以进行性能调整建议：

1. 应该禁用超线程和虚拟化。 （如果使用VM，则启用虚拟化）

2. 电源管理应集中在最小化系统干预和管理上。 设置为“最高性能配置文件”（如果服务器上可用）

3. 禁用 P-states, (unrestricted) Turbo Mode

4. Disable C-states (or change to C0/C1 preference) and T-states (对于媒体等高带宽应用非常重要)

5. Enabling Turbo mode only on minimum amount of cores is better

6. CPU Frequency

检查CPU的最大可用频率（有用的命令：cpupower frequency-info，lshw，lscpu）

监视CPU的活动，并检查内核的当前频率。

7.提取CPU内核状态的有用命令:

$ cat /proc/cpuinfo | sed -n '/^processor\|^cpu MHz/p'

$ turbostat --interval 1

8.其他OS调优

1>. 禁用不是必需任务所必需的所有服务，例如: cups, gpm, ip6tables, mdmonitor, mdmpd, bluetooth, iptables, irqbalance, sysstat.

2>. cpuspeed, nscd, crond, nt如果可用，应启用以下服务pd, ntp, network, tuned

3>. Set IRQ (interrupt request) affinity, refer to What is IRQ Affinity?

4>.设置系统配置文件，重点关注网络性能/延迟.

$ tuned-adm profile network-throughput

$ cpupower frequency-set --governor performance

5>.为了检查调整后的运行情况并使用正确的策略

$ tuned-adm active

6> 关闭Numa平衡

$ echo 0 > /proc/sys/kernel/numa\_balancing

7>. 配置 tuned.conf

添加到 tuned.conf:

[bootloader]

cmdline = audit=0 idle=poll nosoftlockup mce=ignore\_ce

改变tuned-main.conf:

检查事件之前要睡多长时间（以秒为单位），较高的数字表示较低的开销，但响应时间较长。

sleep\_interval = 1 ===>更改为100

动态调整的更新间隔（以秒为单位）。 它必须是sleep\_interval的倍数。

update\_interval = 10 ===>更改为10000

8>. 减少系统调度的推荐配置:

$ echo 100000000 > /proc/sys/kernel/sched\_min\_granularity\_ns

$ echo 50000000 > /proc/sys/kernel/sched\_migration\_cost\_ns

9>. 其他减少系统调度的推荐配置:

$ echo 0 > /proc/sys/vm/swappiness

$ sysctl -w vm.swappiness=0

$ sysctl -w vm.zone\_reclaim\_mode=0

$ echo never > /sys/kernel/mm/transparent\_hugepage/enabled

10>选择正确的NUMA和核心

在具有两个NUMA的计算机上，重要的是选择最接近所用卡的NUMA。

为了找到最靠近卡的NUMA

$ sudo mst status -v

检查哪个内核位于每个NUMA上:

$ lscpu

11>Huge pages

使用大页面可以减少访问页面表条目所需的系统资源量，从而提高系统性能。

在运行Rivermax之前，请启用大页面

$ echo 1000000000 > /proc/sys/kernel/shmmax

$ echo 800 > /proc/sys/vm/nr\_hugepages

##################

五、安装步骤

1. 硬件确认安装成功，系统识别到ConnectX-5硬件设备

[root@bgiga ~]# lspci | grep Mellanox

3b:00.0 Ethernet controller: Mellanox Technologies MT27800 Family [ConnectX-5]

86:00.0 Ethernet controller: Mellanox Technologies MT27800 Family [ConnectX-5]

确认PCI总线必须使用PCIeGen3 x16否则性能瓶颈在PCI总线硬件性能不能达到100G

[root@bgiga ~]# lspci -vv -s 3b:00.0 |grep PCIeGen

[V0] Vendor specific: PCIeGen3 x16

检查确认PCI端口速度和带宽: Speed 8GT/s, Width x16

[root@bgiga ~]# lspci -vv |grep -i Mell

3b:00.0 Ethernet controller: Mellanox Technologies MT27800 Family [ConnectX-5]

Subsystem: Mellanox Technologies Device 0006

86:00.0 Ethernet controller: Mellanox Technologies MT27800 Family [ConnectX-5]

Subsystem: Mellanox Technologies Device 0006

[root@bgiga ~]# lspci -vv -s 86:00.0|grep Wid

LnkCap: Port #0, Speed 8GT/s, Width x16, ASPM not supported, Exit Latency L0s unlimited, L1 <4us ExtSynch- ClockPM- AutWidDis- BWInt- AutBWInt-

LnkSta: Speed 8GT/s, Width x8, TrErr- Train- SlotClk+ DLActive- BWMgmt- ABWMgmt-

[root@cmob v2.87]# lspci -vv -s 3b:00.0|grep Wid

LnkCap: Port #0, Speed 8GT/s, Width x16, ASPM not supported, Exit Latency L0s unlimited, L1 <4us ExtSynch- ClockPM- AutWidDis- BWInt- AutBWInt-

LnkSta: Speed 8GT/s, Width x8, TrErr- Train- SlotClk+ DLActive- BWMgmt- ABWMgmt-

2.安装OFED驱动

所需安装文件可以访问10.87.30.82:/root获得(MLNX\_OFED\_LINUX-5.2-2.2.3.0-rhel7.6-x86\_64.tgz, trex-v2.87.tar.gz)

下载安装文件MLNX\_OFED\_LINUX-5.2-2.2.3.0-rhel7.6-x86\_64.tgz，解压缩并执行安装命令：

tar zxf MLNX\_OFED\_LINUX-5.2-2.2.3.0-rhel7.6-x86\_64.tgz

cd MLNX\_OFED\_LINUX-5.2-2.2.3.0-rhel7.6-x86\_64

./mlnxofedinstall

安装过程结束后执行如下命令启动相关服务确认安装成功

[root@bgiga v2.87]# /etc/init.d/openibd restart

Unloading HCA driver: [ OK ]

Loading HCA driver and Access Layer: [ OK ]

[root@bgiga v2.87]# mst start

Starting MST (Mellanox Software Tools) driver set

Loading MST PCI module - Success

Loading MST PCI configuration module - Success

Create devices

[root@bgiga v2.87]# ofed\_info

MLNX\_OFED\_LINUX-5.2-2.2.3.0 (OFED-5.2-2.2.3):

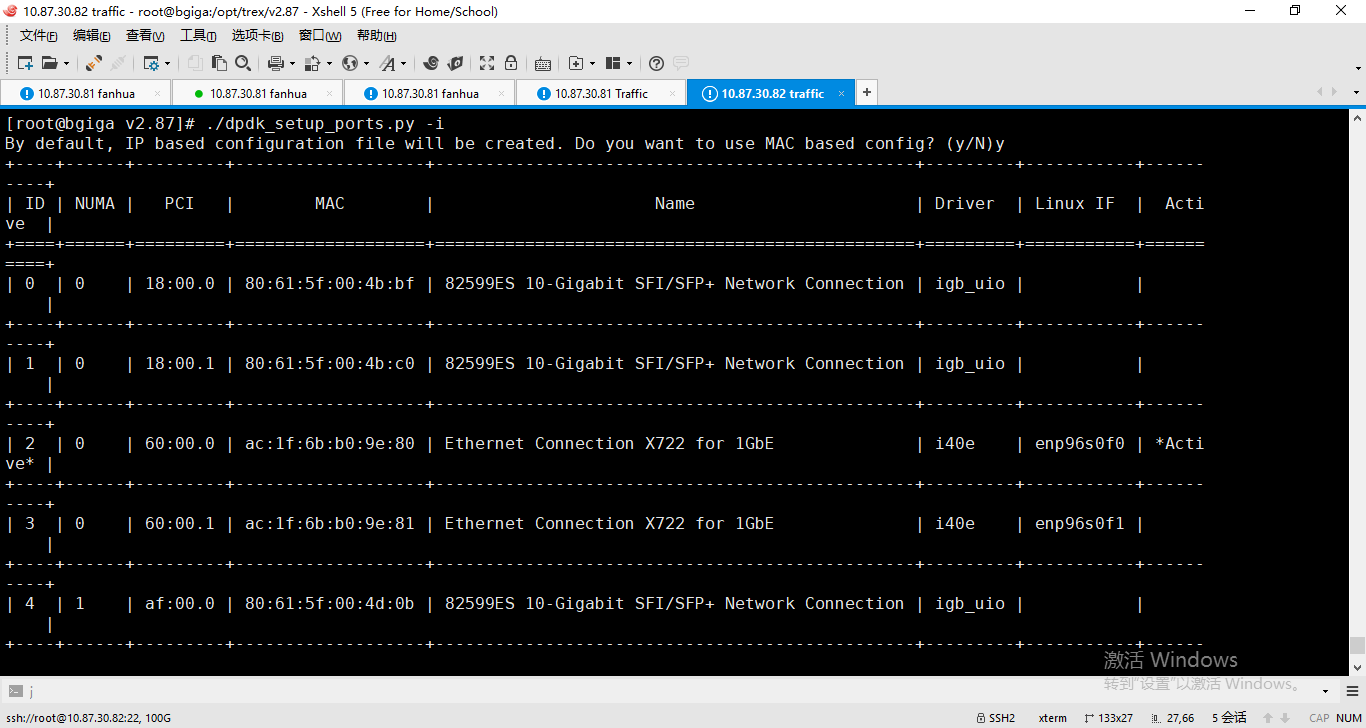
3.安装trex v2.87

下载安装包trex-v2.87.tar.gz，将安装包解压缩到/opt/trex即可

[root@bgiga trex]# mkdir –p /opt/trex

[root@bgiga trex]# tar zxf /root/trex-v2.87.tar.gz

4.运行dpdk\_setup\_ports.py创建配置文件，更详细信息参见TRex使用说明.docx



5. 验证传输性能

1>启动无状态模式服务器端

[root@bgiga v2.87]#./t-rex-64 -i -v 7

使用-c 50参数指定cpu核数50个分配给服务器。

2>在另一个终端上以交互方式启动客户端

[root@bgiga v2.87]# ./trex-console

3>在端口0发送测试数据

trex>start -f stl/tests/single\_cont.py -p 0 --pin -m 100% --force

Removing all streams from port(s) [0.\_]: [SUCCESS]

Attaching 1 streams to port(s) [0.\_]: [SUCCESS]

附：测试日志

centos7.6 + MLNX\_OFED\_LINUX-5.2-2.2.3.0 + trex-2.87

Intel(R) Xeon(R) CPU E5-2665 0 @ 2.40GHz

64b: ./t-rex-64 -i -v 7 -c 15， start -f stl/tests/single\_cont1.py -p 0 --pin -m 100%: 64bit frame L1 bps: 65.3Gbps, cpu 93.0%

128b: ./t-rex-64 -i -v 7 -c 15， start -f stl/tests/single\_cont1.py -p 0 --pin -m 100%: 64bit frame L1 bps: 58.3Gbps, cpu 88%

256b: ./t-rex-64 -i -v 7 -c 15， start -f stl/tests/single\_cont1.py -p 0 --pin -m 100%: 64bit frame L1 bps: 54.3Gbps, cpu 85%

64b: ./t-rex-64 -i -v 7 -c 1， start -f stl/tests/single\_cont1.py -p 0 --pin -m 100%: 64bit frame L1 bps: 18.3Gbps, cpu 100%

128b: ./t-rex-64 -i -v 7 -c 1， start -f stl/tests/single\_cont1.py -p 0 --pin -m 100%: 64bit frame L1 bps: 32.3Gbps, cpu 87%

128b: ./t-rex-64 -i -v 7 -c 1， start -f stl/tests/single\_cont\_256b.py -p 0 --pin -m 100%: 64bit frame L1 bps: 46.3Gbps, cpu 94%

centos7.6, v2.88, ofed-5.2: not work

EAL: Detected 32 lcore(s)

EAL: Detected 2 NUMA nodes

EAL: Static memory layout is selected, amount of reserved memory can be adjusted with -m or --socket-mem

EAL: Multi-process socket /var/run/dpdk/rte/mp\_socket

EAL: Selected IOVA mode 'PA'

EAL: Probing VFIO support...

EAL: Could not find space for memseg. Please increase CONFIG\_RTE\_MAX\_MEMSEG\_PER\_TYPE and/or CONFIG\_RTE\_MAX\_MEM\_PER\_TYPE in configuration.

EAL: Couldn't remap hugepage files into memseg lists

EAL: FATAL: Cannot init memory

EAL: Cannot init memory

You might need to run ./trex-cfg once

EAL: Error - exiting with code: 1

Cause: Invalid EAL arguments

Killing Scapy server...

centos7.6, MLNX\_OFED\_LINUX-4.6-1.0.1.1-rhel7.6-x86\_64, trex-v2.86: not work

EAL: Static memory layout is selected, amount of reserved memory can be adjusted with -m or --socket-mem

EAL: so/x86\_64/libmlx5-64.so: undefined symbol: mlx5dv\_devx\_umem\_reg

EAL: FATAL: Cannot init plugins

EAL: Cannot init plugins

You might need to run ./trex-cfg once

EAL: Error - exiting with code: 1

Cause: Invalid EAL arguments

Killing Scapy server... Scapy server is killed

[root@localhost v2.86]#

brcm-shell

libdoc LibByz.py LibByz.html

./dpdk\_setup\_ports.py -t

./dpdk\_setup\_ports.py -i

./t-rex-64 -f cap2/dns.yaml -c 4 -m 1 -d 10

ConnectX-5 on centos7.9:

driver:

https://www.mellanox.com/products/infiniband-drivers/linux/mlnx\_ofed

MLNX\_OFED\_LINUX-5.2-1.0.4.0-rhel7.9-x86\_64.tgz

https://community.mellanox.com/s/article/getting-started-with-connectx-5-100gb-s-adapters-for-linux

https://www.mellanox.com/related-docs/prod\_software/Performance\_Tuning\_Guide\_for\_Mellanox\_Network\_Adapters.pdf

To load the new driver, run:

/etc/init.d/openibd restart

mlxfwreset -d mlx5\_0 --yes r

mst start

service firewalld stop

systemctl disable firewalld

service iptables stop

ibv\_devinfo | grep vendor\_part\_id

mlxconfig -d /dev/mst/mt4121\_pciconf0 q

ifconfig enp7s0f0 192.168.1.1/24 up

ifconfig enp7s0f1 192.168.2.1/24 up

ifconfig enp7s0f1 mtu 9000

ifconfig enp7s0f0 mtu 9000

ibv\_devinfo

ofed\_info

ofed\_info -s

10.87.30.82：

[root@cmob v2.87]# ibdev2netdev

mlx5\_0 port 1 ==> enp59s0 (Up)

mlx5\_1 port 1 ==> enp134s0 (Up)

ibdev2netdev

mlx5\_0 port 1 ==> enp7s0f0 (Down)

mlx5\_1 port 1 ==> enp7s0f1 (Down)

lspci | grep Mellanox

mii-tool

[bash]>sudo lspci | grep Mellanox

3b:00.0 Ethernet controller: Mellanox Technologies MT27800 Family [ConnectX-5]

86:00.0 Ethernet controller: Mellanox Technologies MT27800 Family [ConnectX-5]

[root@cmob v2.87]#

[bash]>sudo lspci -vv -s 3b:00.0

LnkSta: Speed 8GT/s, Width x8, TrErr- Train- SlotClk+ DLActive- BWMgmt- ABWMgmt-

[bash]>sudo lspci -vt | grep Mellanox

http://www.h3c.com/cn/d\_202007/1317229\_30005\_0.htm

mlxlink -d /dev/mst/mt4121\_pciconf0 -a UP

ibstat

ibdiagnet

connectx\_port\_config

ca\_self\_test.ofed

ibv\_devices

ib\_write\_bw

使用numactl来查看node0和node1上认领的cpu核数,以及内存资源

numactl --hardware

cat /sys/devices/system/node/node\*/meminfo | fgrep Huge

内存：

free -m

EAL: 128 hugepages of size 1073741824 reserved, but no mounted hugetlbfs found for that size

echo 8192 > /sys/devices/system/node/node0/hugepages/hugepages-2048kB/nr\_hugepages

echo 8192 > /sys/devices/system/node/node1/hugepages/hugepages-2048kB/nr\_hugepages

echo 16 > /sys/devices/system/node/node0/hugepages/hugepages-1048576kB/nr\_hugepages

echo 16 > /sys/devices/system/node/node1/hugepages/hugepages-1048576kB/nr\_hugepages

mkdir -p /mnt/huge2m

mkdir -p /mnt/huge1g

umount /mnt/huge1g

umount /mnt/huge2m

mount -t hugetlbfs none /mnt/huge1g -o pagesize=1GB

mount -t hugetlbfs none /mnt/huge2m -o pagesize=2MB