kvm虚拟机中fastswap配置

一、基本配置

利用SR-IOV技术实现了虚拟机环境下的RDMA配置，能够进行RDMA有关benchmark的运行

1. 进入虚拟机

```
# 1. ssh登录我的n2服务器账号
# -y是为了使用virt-manager GUI，不需要可删掉
ssh -Y -p 9151 linqinluli@202.120.39.14
password: yhz20010101
# 2. 登录虚拟机,虚拟机root密码同登录密码
ssh yanghanzhang@192.168.122.100
password: yanghanzhang
# 3. 登录gpu2，用作rdma server，目前gpu2上版本不对
# 我开了另一台虚拟机用作server实验，
ssh -Y -p 9035 yanghanzhang@202.120.39.14
password: yanghanzhang
```

2. 配置虚拟网卡ip

```
sudo ifconfig enp6s0 20.20.20.100 up
```

3. 查看RDMA配置情况

```
sudo mst start
sudo mst status
```

这里存在找不到设备的情况！！！（并没有什么影响）~~可能是问题所在，网上有人遇到类似问题~~

MST does not load · Issue #21 · Azure/azhpc-images (github.com)

```
yanghanzhang@yanghanzhang:~$ sudo mst start
Starting MST (Mellanox Software Tools) driver set
Loading MST PCI module - Success
Loading MST PCI configuration module - Success
Create devices
Unloading MST PCI module (unused) - Success
Unloading MST PCI configuration module (unused) - Success
yanghanzhang@yanghanzhang:~$ sudo mst status
MST modules:
------------
    MST PCI module is not loaded
    MST PCI configuration module is not loaded

PCI Devices:
------------

        No devices were found.
```

```
ibstat # 查询端口状态
```

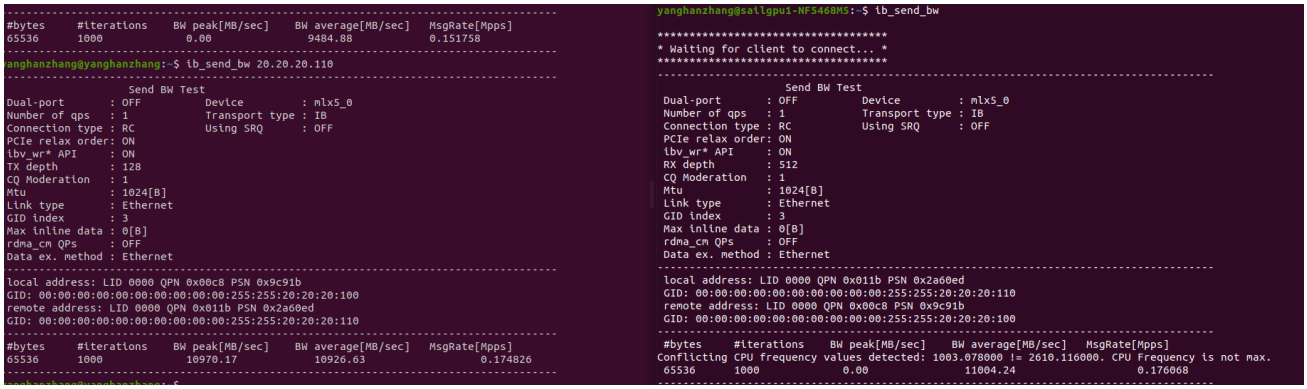
```
yanghanzhang@yanghanzhang:~$ ibstat
CA 'mlx5_0'
        CA type: MT4120
        Number of ports: 1
        Firmware version: 16.33.1048
        Hardware version: 0
        Node GUID: 0x002233fffe445566
        System image GUID: 0x0c42a103007560b4
        Port 1:
                State: Active
                Physical state: LinkUp
                Rate: 100
                Base lid: 0
                LMC: 0
                SM lid: 0
                Capability mask: 0x00010000
                Port GUID: 0x022233fffe445566
                Link layer: Ethernet
```

```
lspci -D | grep Mellanox # PCI状态
sudo ibdev2netdev -v     # 驱动绑定状态
```

```
yanghanzhang@yanghanzhang:~$ ibdev2netdev
mlx5_0 port 1 ==> enp6s0 (Up)
yanghanzhang@yanghanzhang:~$ lspci -D | grep Mellanox
0000:06:00.0 Ethernet controller: Mellanox Technologies MT27800 Family [ConnectX-5 Virtual Function]
yanghanzhang@yanghanzhang:~$ sudo ibdev2netdev -v
0000:06:00.0 mlx5_0 (MT4120 - NA)  fw 16.33.1048 port 1 (ACTIVE) ==> enp6s0 (Up)
```

4. RDMA测试，参考该网站，有较完整的RDMA测试benchmark

   How To Enable, Verify and Troubleshoot RDMA (mellanox.com)



二、Fastswap安装

参考fastswap GitHub上的安装流程

clusterfarmem/fastswap: Fastswap, a fast swap system for far memory through RDMA (github.com)

其中最最最最最最最最重要的是，使用和github上一样的4.3**版本的**OFED Driver，后面遇到的所有问题都可以解决了

目前在两台虚拟机上安装的fastswap，需要和之前一样完成虚拟机中RDMA的配置

记住每次虚拟机关机or重启需要重新配置ip

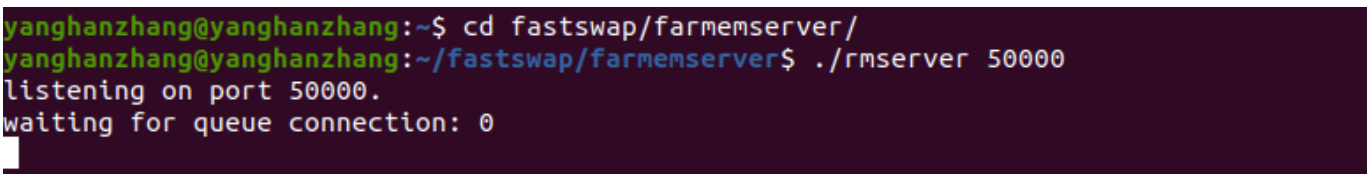
```
sudo ifconfig enp6s0 20.20.20.100 up # client的配置

sudo ifconfig enp6s0 20.20.20.130 up # server的配置
```

之后简单测试能够互相ib_send_bw即可

server:

```
cd fastswap/farmemserver
make
./rmserver 50000
```



client:

```
# 编译
cd drivers
make BACKEND=RDMA
# 安装
sudo insmod fastswap_rdma.ko sport=50000 sip="20.20.20.103" cip="20.20.20.100" nq=4
sudo insmod fastswap.ko
```

二、Fastswap安装

参考fastswap GitHub上的安装流程

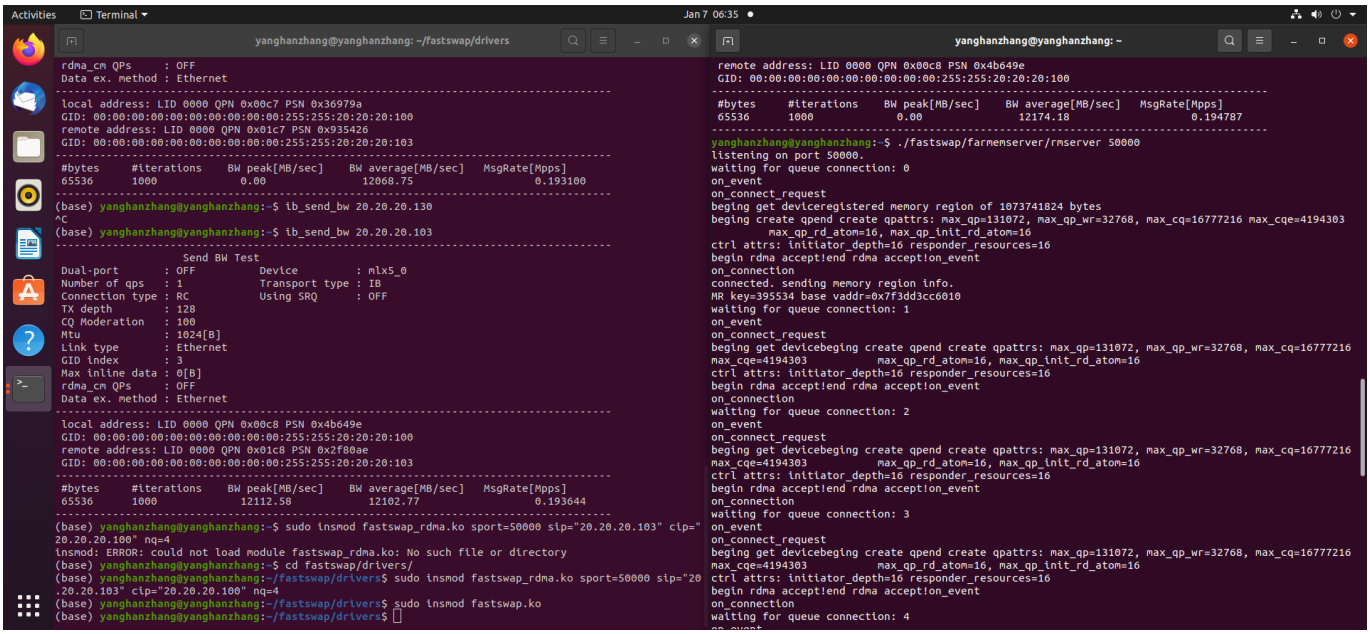
clusterfarmem/fastswap: Fastswap, a fast swap system for far memory through RDMA (github.com)

rdma server端运行：

```
./fastswap/farmemserver/rmserver 50000
```

rdma client端（即虚拟机内）运行

```
sudo insmod fastswap_rdma.ko sport=50000 sip="20.20.20.103" cip="20.20.20.100" nq=4
```



到此fastswap已经安装完成，即初步的并行远内存系统后端搭建完成，下一步在该系统运行不同程序，分析程序特征

问题排查

一、fastswap安装版本问题

问题描述

其中为了编译通过对fastswap_rdma.c文件进行了修改，修改之前报错如下：

```
yanghanzhang@yanghanzhang:~/fastswap/drivers$ make BACKEND=RDMA
make -C /lib/modules/`uname -r`/build M=$PWD
make[1]: Entering directory '/usr/src/linux-headers-4.11.0-fastswap'
  CC [M]  /home/yanghanzhang/fastswap/drivers/fastswap.o
  CC [M]  /home/yanghanzhang/fastswap/drivers/fastswap_rdma.o
/home/yanghanzhang/fastswap/drivers/fastswap_rdma.c:42:13: error: initialization from incompatible pointer type [-Werror=incompatible-pointer-types]
   .add    = sswap_rdma_addone,
             ^
/home/yanghanzhang/fastswap/drivers/fastswap_rdma.c:42:13: note: (near initialization for 'sswap_rdma_ib_client.add')
/home/yanghanzhang/fastswap/drivers/fastswap_rdma.c: In function 'sswap_rdma_create_qp':
/home/yanghanzhang/fastswap/drivers/fastswap_rdma.c:110:28: error: 'IB_QP_EXP_CREATE_ATOMIC_BE_REPLY' undeclared (first use in this function)
   init_attr.create_flags = IB_QP_EXP_CREATE_ATOMIC_BE_REPLY & 0;
                            ^
/home/yanghanzhang/fastswap/drivers/fastswap_rdma.c:110:28: note: each undeclared identifier is reported only once for each function it appears in
/home/yanghanzhang/fastswap/drivers/fastswap_rdma.c: In function 'sswap_rdma_post_rdma':
/home/yanghanzhang/fastswap/drivers/fastswap_rdma.c:492:42: error: passing argument 3 of 'ib_post_send' from incompatible pointer type [-Werror=incompatible-pointer-types]
   ret = ib_post_send(q->qp, &rdma_wr.wr, &bad_wr);
                                          ^
In file included from /home/yanghanzhang/fastswap/drivers/fastswap_rdma.h:4:0,
                 from /home/yanghanzhang/fastswap/drivers/fastswap_rdma.c:3:
/usr/src/ofa_kernel/4.11.0-fastswap/include/rdma/ib_verbs.h:3969:19: note: expected 'const struct ib_send_wr **' but argument is of type 'struct ib_send_wr **'
 static inline int ib_post_send(struct ib_qp *qp,
                   ^
/home/yanghanzhang/fastswap/drivers/fastswap_rdma.c: In function 'sswap_rdma_post_recv':
/home/yanghanzhang/fastswap/drivers/fastswap_rdma.c:536:34: error: passing argument 3 of 'ib_post_recv' from incompatible pointer type [-Werror=incompatible-pointer-types]
   ret = ib_post_recv(q->qp, &wr, &bad_wr);
                                  ^
In file included from /home/yanghanzhang/fastswap/drivers/fastswap_rdma.h:4:0,
                 from /home/yanghanzhang/fastswap/drivers/fastswap_rdma.c:3:
/usr/src/ofa_kernel/4.11.0-fastswap/include/rdma/ib_verbs.h:3986:19: note: expected 'const struct ib_recv_wr **' but argument is of type 'struct ib_recv_wr **'
 static inline int ib_post_recv(struct ib_qp *qp,
                   ^
cc1: some warnings being treated as errors
scripts/Makefile.build:300: recipe for target '/home/yanghanzhang/fastswap/drivers/fastswap_rdma.o' failed
make[2]: *** [/home/yanghanzhang/fastswap/drivers/fastswap_rdma.o] Error 1
Makefile:1492: recipe for target '_module_/home/yanghanzhang/fastswap/drivers' failed
make[1]: *** [_module_/home/yanghanzhang/fastswap/drivers] Error 2
make[1]: Leaving directory '/usr/src/linux-headers-4.11.0-fastswap'
Makefile:19: recipe for target 'default' failed
make: *** [default] Error 2
```

针对报错修改如下：

a. IB_QP_EXP_CREATE_ATOMIC_BE_REPLY用来看编译所有头文件对不对，但现在这个版本似乎没有这个变量，直接删掉

b. static void sswap_rdma_addone(struct ib_device *dev)函数改为

static int sswap_rdma_addone(struct ib_device *dev)

c. struct ib_send_wr *bad_wr和struct ib_recv_wr *bad_wr分别添加const限定

之后编译通过，继续按照安装步骤进行安装

到了安装fastswap_rdma.ko文件的时候遇到了运行错误

```
sudo insmod fastswap_rdma.ko sport=50000 sip="20.20.20.21" cip="20.20.20.100" nq=4
```

dmesg看到的输出如下，其中用输出debug法定位到问题在ret = rdma_connect(q->cm_id, &param)这个函数，运行到这儿直接阻塞住，进而触发后续的wait失败

```
[   47.070270] fastswap_rdma: * RDMA BACKEND *
[   47.070290] fastswap_rdma: sswap_rdma_addone() = mlx5_0
[   47.070293] fastswap_rdma: will try to connect to 20.20.20.110:50000
[   47.070295] fastswap_rdma: start: sswap_rdma_parse_ipaddr
[   47.070297] fastswap_rdma: start: sswap_rdma_parse_ipaddr
[   47.070298] fastswap_rdma: start: sswap_rdma_init_queue
[   47.070712] fastswap_rdma: cm_handler msg: address resolved (0) status 0 id ffff9628b9785400
[   47.070715] fastswap_rdma: start: sswap_rdma_addr_resolved
[   47.070717] fastswap_rdma: selecting device mlx5_0
[   47.071736] fastswap_rdma: start: sswap_rdma_create_queue_ib
[   47.073926] fastswap_rdma: start: sswap_rdma_create_qp
[   47.079307] fastswap_rdma: cm_handler msg: route resolved  (2) status 0 id ffff9628b9785400
[   47.079310] fastswap_rdma: max_qp_rd_atom=16 max_qp_init_rd_atom=16
[   47.079311] fastswap_rdma: begin rdma_connect
[  107.712758] fastswap_rdma: sswap_rdma_wait_for_cm failed
[  242.881338] INFO: task kworker/u8:4:206 blocked for more than 120 seconds.
[  242.881413]       Tainted: G           OE   4.11.0-fastswap #6
[  242.881457] "echo 0 > /proc/sys/kernel/hung_task_timeout_secs" disables this message.
[  242.881515] kworker/u8:4    D    0   206      2 0x00000000
[  242.881543] Workqueue: rdma_cm cma_work_handler [rdma_cm]
[  242.881546] Call Trace:
[  242.881576]  __schedule+0x3b9/0x8f0
[  242.881580]  schedule+0x36/0x80
[  242.881583]  schedule_preempt_disabled+0xe/0x10
[  242.881586]  __mutex_lock.isra.5+0x271/0x4e0
[  242.881591]  __mutex_lock_slowpath+0x13/0x20
[  242.881594]  ? __mutex_lock_slowpath+0x13/0x20
[  242.881597]  mutex_lock+0x2f/0x40
[  242.881606]  rdma_connect+0x23/0x50 [rdma_cm]
[  242.881614]  sswap_rdma_cm_handler+0xe2/0x3b0 [fastswap_rdma]
[  242.881623]  cma_work_handler+0xa3/0xe0 [rdma_cm]
[  242.881627]  process_one_work+0x16b/0x4a0
[  242.881631]  worker_thread+0x4b/0x500
[  242.881635]  kthread+0x109/0x140
[  242.881638]  ? process_one_work+0x4a0/0x4a0
[  242.881642]  ? kthread_create_on_node+0x70/0x70
[  242.881647]  ret_from_fork+0x2c/0x40
[  242.881670] INFO: task insmod:1870 blocked for more than 120 seconds.
```

rdma_wait_for_cm_failed：

直接原因：

ret = sswap_rdma_wait_for_cm(queue)失败

该函数调用

```
wait_for_completion_interruptible_timeout(&queue->cm_done,
msecs_to_jiffies(CONNECTION_TIMEOUT_MS) + 1)
//CONNECTION_TIMEOUT_MS 设置的60s应该不会是这个设置太短的原因
```

而queue->cm_done信号量一直没有被释放，运行超时，看cm_handler函数里面对cm_done进行释放的事件即RDMA_CM_EVENT_ESTABLISHED，因此可以看出没有成功创建连接。

通过pr_info定位到程序阻塞在了rdma_connect()该函数中

```
pr_info("max_qp_rd_atom=%d max_qp_init_rd_atom=%d\n",
        q->ctrl->rdev->dev->attrs.max_qp_rd_atom,
        q->ctrl->rdev->dev->attrs.max_qp_init_rd_atom);
pr_info("begin rdma_connect");
ret = rdma_connect(q->cm_id, &param);
pr_info("end rdma_connect");
pr_info("ret");
if (ret)
{
```

问题解决：

a. 我尝试过调整参数，或者直接设为NULL还是会导致程序一直挂在这儿

b. 尝试从后两句报错入手修改，也不是问题所在，还是会卡在rdma_connect这一步

c. 发现使用的ofed内核版本不对，不是4.11的版本，有点怀疑之前装内核这一步错了orz

d. rdma通信两端都改成4.3版本的OFED Driver，问题解决！