**Work Note of RDMA Study**

Zhuangdi Zhu

zhuangdizhu@yahoo.com

2015-01-25

Target: Trying to make SR-IOV working to enable KVM with IB.

Progress:

Step 1: uninstall OFA\_OFED and install MLNX\_OFED\_LINUX-2.4-1.0.0-rhel6.5-x86\_64

#tar zxvf MLNX\_OFED\_LINUX-2.4-1.0.0-rhel6.5-x86\_64.tgz

#cd MLNX\_OFED\_LINUX-2.4-1.0.0-rhel6.5-x86\_64

#./mlnxofedinstall --enable-sriov

Step 2:

[root@zhuangdizhu1 ~]#ifconfig ib0 172.16.1.2 up

[root@zhuangdizhu2 ~]#ifconfig ib0 172.16.1.3 up

[root@zhuangdizhu1 ~]# ibv\_rc\_pingpong

local address: LID 0x0002, QPN 0x040221, PSN 0x4b63be, GID ::

remote address: LID 0x0001, QPN 0x040224, PSN 0x9f9e6a, GID ::

8192000 bytes in 0.01 seconds = 11558.38 Mbit/sec

1000 iters in 0.01 seconds = 5.67 usec/iter

[root@zhuangdizhu2 ~]# ibv\_rc\_pingpong 172.16.1.2

local address: LID 0x0001, QPN 0x040224, PSN 0x9f9e6a, GID ::

remote address: LID 0x0002, QPN 0x040221, PSN 0x4b63be, GID ::

8192000 bytes in 0.01 seconds = 11774.34 Mbit/sec

1000 iters in 0.01 seconds = 5.57 usec/iter

Step 3:

[root@zhuangdizhu1 ~]# rdma\_server

rdma\_server: start

rdma\_server: end 0

[root@zhuangdizhu2 ~]# rdma\_client -s 172.16.1.2

rdma\_client: start

rdma\_client: end 0

Step 4: Enable “Intel Virtualization Technology” in BIOS(I cannot find “SR-IOV” option in BIOS)

Step 5:

#mst start

Starting MST (Mellanox Software Tools) driver set

[warn] mst\_pci is already loaded, skipping

[warn] mst\_pciconf is already loaded, skipping

Create devices

-W- Missing lsusb command, skipping MTUSB devices detection

Step 6:

#lspci -v

02:00.0 Network controller: Mellanox Technologies MT27500 Family [ConnectX-3]

Subsystem: Mellanox Technologies Device 0017

Flags: bus master, fast devsel, latency 0, IRQ 17

Memory at f7200000 (64-bit, non-prefetchable) [size=1M]

Memory at f2800000 (64-bit, prefetchable) [size=8M]

Expansion ROM at f7100000 [disabled] [size=1M]

Capabilities: [40] Power Management version 3

Capabilities: [48] Vital Product Data

Capabilities: [9c] MSI-X: Enable+ Count=128 Masked-

Capabilities: [60] Express Endpoint, MSI 00

Capabilities: [100] Alternative Routing-ID Interpretation (ARI)

Capabilities: [148] Device Serial Number f4-52-14-03-00-89-b1-b0

**Capabilities: [108] Single Root I/O Virtualization (SR-IOV)**

Capabilities: [154] Advanced Error Reporting

Capabilities: [18c] #19

Kernel driver in use: mlx4\_core

Kernel modules: mlx4\_core

Step 7:update the /boot/grub/grub.conf file

default=0

timeout=5

splashimage=(hd0,0)/grub/splash.xpm.gz

hiddenmenu

title CentOS (2.6.32-431.el6.x86\_64)

root (hd0,0)

kernel /vmlinuz-2.6.32-431.el6.x86\_64 ro root=/dev/mapper/vg\_zhuangdizhu1-lv\_root rd\_LVM\_LV=vg\_zhuangdizhu1/lv\_swap rd\_NO\_MD crashkernel=auto LANG=zh\_CN.UTF-8 rd\_NO\_LUKS KEYBOARDTYPE=pc KEYTABLE=us rd\_NO\_DM rd\_LVM\_LV=vg\_zhuangdizhu1/lv\_root rhgb quiet **intel\_iommu=on**

initrd /initramfs-2.6.32-431.el6.x86\_64.img

Step 8:

#mlxconfig -d /dev/mst/mt4099\_pciconf0 set SRIOV\_EN=1

# flint -d /dev/mst/mt4099\_pciconf0 dc

[HCA]

**num\_pfs = 1**

**total\_vfs = 16**

**sriov\_en = true**

hca\_header\_device\_id = 0x1003

hca\_header\_subsystem\_id = 0x0017

dpdp\_en = true

eth\_xfi\_en = true

mdio\_en\_port1 = 0

pcie\_tx\_polarity = 0x0f

Step 9:

vim /etc/modprobe.d/mlx4\_core.conf

options mlx4\_core port\_type\_array=**1,2** num\_vfs=5 probe\_vf=1

Step 10:Reboot the server.

Step 11:

# lspci | grep Mellanox

02:00.0 Network controller: Mellanox Technologies MT27500 Family [ConnectX-3]

Step 12:

#dmesg | grep mlx

mlx4\_core: Mellanox ConnectX core driver v2.4-1.0.0 (Jan 13 2015)

mlx4\_core: Initializing 0000:02:00.0

mlx4\_core 0000:02:00.0: PCI INT A -> GSI 17 (level, low) -> IRQ 17

mlx4\_core 0000:02:00.0: setting latency timer to 64

mlx4\_core 0000:02:00.0: PCIe link speed is 8.0GT/s, device supports 8.0GT/s

mlx4\_core 0000:02:00.0: PCIe link width is x8, device supports x8

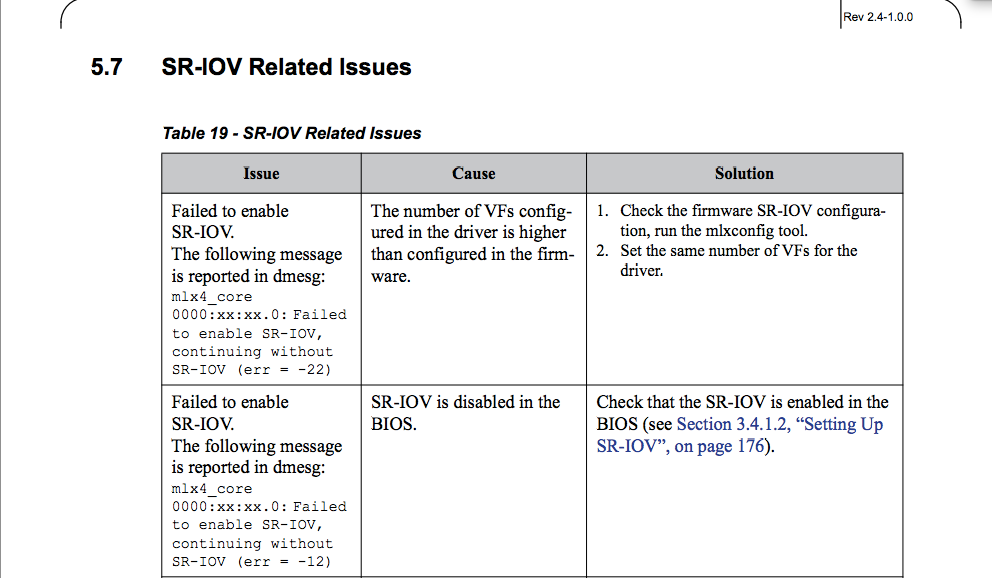
mlx4\_core 0000:02:00.0: Enabling SR-IOV with 5 VFs

**mlx4\_core 0000:02:00.0: not enough MMIO resources for SR-IOV**

**mlx4\_core 0000:02:00.0: Failed to enable SR-IOV, continuing without SR-IOV (err = -12)**

mlx4\_core 0000:02:00.0: irq 43 for MSI/MSI-X

information from Mellanox\_OFED\_Linux\_User\_Manual\_v2.4-1.0.0.pdf:



From the information above, I think the conclusion can be made that SR-IOV is disabled in the BIOS.

Then I tried to find whether the motherboard and its BIOS support SR-IOV or not.

[root@zhuangdizhu2 ~]# dmidecode -t baseboard

# dmidecode 2.12

SMBIOS 2.7 present.

Handle 0x0002, DMI type 2, 15 bytes

Base Board Information

**Manufacturer: ASUSTeK COMPUTER INC.**

**Product Name: P8Z77-V LK**

Version: Rev X.0x

Serial Number: 130713616602415

Asset Tag: To be filled by O.E.M.

Features:

Board is a hosting board

Board is replaceable

Location In Chassis: To be filled by O.E.M.

Chassis Handle: 0x0003

Type: Motherboard

Contained Object Handles: 0

Handle 0x002A, DMI type 10, 6 bytes

On Board Device Information

Type: Ethernet

Status: Enabled

Description: Onboard Ethernet

Handle 0x005C, DMI type 41, 11 bytes

Onboard Device

Reference Designation: Onboard IGD

Type: Video

Status: Enabled

Type Instance: 1

Bus Address: 0000:00:02.0

Handle 0x005D, DMI type 41, 11 bytes

Onboard Device

Reference Designation: Onboard LAN

Type: Ethernet

Status: Enabled

Type Instance: 1

Bus Address: 0000:00:19.0

Handle 0x005E, DMI type 41, 11 bytes

Onboard Device

Reference Designation: Onboard 1394

Type: Other

Status: Enabled

Type Instance: 1

Bus Address: 0000:03:1c.2

[root@zhuangdizhu1 ~]# dmidecode -t bios

# dmidecode 2.12

SMBIOS 2.7 present.

Handle 0x0000, DMI type 0, 24 bytes

BIOS Information

Vendor: American Megatrends Inc.

**Version: 1104**

Release Date: 08/23/2013

Address: 0xF0000

Runtime Size: 64 kB

ROM Size: 8192 kB

Characteristics:

PCI is supported

BIOS is upgradeable

BIOS shadowing is allowed

Boot from CD is supported

Selectable boot is supported

BIOS ROM is socketed

EDD is supported

5.25"/1.2 MB floppy services are supported (int 13h)

3.5"/720 kB floppy services are supported (int 13h)

3.5"/2.88 MB floppy services are supported (int 13h)

Print screen service is supported (int 5h)

8042 keyboard services are supported (int 9h)

Serial services are supported (int 14h)

Printer services are supported (int 17h)

ACPI is supported

USB legacy is supported

BIOS boot specification is supported

Targeted content distribution is supported

UEFI is supported

BIOS Revision: 4.6

Handle 0x006C, DMI type 13, 22 bytes

BIOS Language Information

Language Description Format: Long

Installable Languages: 8

en|US|iso8859-1

fr|FR|iso8859-1

es|ES|iso8859-1

de|DE|iso8859-1

ru|RU|iso8859-5

ja|JP|unicode

zh|TW|unicode

zh|CN|unicode

Currently Installed Language: en|US|iso8859-1

I searched the motherboard model “ASUS P8Z77-V LK” through google and also downloaded its specification and user manual. None of them give any information about support of SR-IOV. Now the problem for me is to know whether the motherboard supports SR-IOV and how can I enable this in BIOS.

CPU details:

[root@zhuangdizhu1 ~]# cat /proc/cpuinfo

processor : 0

vendor\_id : GenuineIntel

cpu family : 6

model : 58

model name : Intel(R) Core(TM) i7-3770K CPU @ 3.50GHz

stepping : 9

cpu MHz : 3510.352

cache size : 8192 KB

physical id : 0

siblings : 8

core id : 0

cpu cores : 4

apicid : 0

initial apicid : 0

fpu : yes

fpu\_exception : yes

cpuid level : 13

wp : yes

flags : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx rdtscp lm constant\_tsc arch\_perfmon pebs bts rep\_good xtopology nonstop\_tsc aperfmperf pni pclmulqdq dtes64 monitor ds\_cpl vmx est tm2 ssse3 cx16 xtpr pdcm pcid sse4\_1 sse4\_2 popcnt tsc\_deadline\_timer aes xsave avx f16c rdrand lahf\_lm ida arat epb xsaveopt pln pts dts tpr\_shadow vnmi flexpriority ept vpid fsgsbase smep erms

bogomips : 7020.70

clflush size : 64

cache\_alignment : 64

address sizes : 36 bits physical, 48 bits virtual

power management:

processor : 1

vendor\_id : GenuineIntel

cpu family : 6

model : 58

model name : Intel(R) Core(TM) i7-3770K CPU @ 3.50GHz

stepping : 9

cpu MHz : 3510.352

cache size : 8192 KB

physical id : 0

siblings : 8

core id : 1

cpu cores : 4

apicid : 2

initial apicid : 2

fpu : yes

fpu\_exception : yes

cpuid level : 13

wp : yes

flags : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx rdtscp lm constant\_tsc arch\_perfmon pebs bts rep\_good xtopology nonstop\_tsc aperfmperf pni pclmulqdq dtes64 monitor ds\_cpl vmx est tm2 ssse3 cx16 xtpr pdcm pcid sse4\_1 sse4\_2 popcnt tsc\_deadline\_timer aes xsave avx f16c rdrand lahf\_lm ida arat epb xsaveopt pln pts dts tpr\_shadow vnmi flexpriority ept vpid fsgsbase smep erms

bogomips : 7020.70

clflush size : 64

cache\_alignment : 64

address sizes : 36 bits physical, 48 bits virtual

power management:

processor : 2

vendor\_id : GenuineIntel

cpu family : 6

model : 58

model name : Intel(R) Core(TM) i7-3770K CPU @ 3.50GHz

stepping : 9

cpu MHz : 3510.352

cache size : 8192 KB

physical id : 0

siblings : 8

core id : 2

cpu cores : 4

apicid : 4

initial apicid : 4

fpu : yes

fpu\_exception : yes

cpuid level : 13

wp : yes

flags : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx rdtscp lm constant\_tsc arch\_perfmon pebs bts rep\_good xtopology nonstop\_tsc aperfmperf pni pclmulqdq dtes64 monitor ds\_cpl vmx est tm2 ssse3 cx16 xtpr pdcm pcid sse4\_1 sse4\_2 popcnt tsc\_deadline\_timer aes xsave avx f16c rdrand lahf\_lm ida arat epb xsaveopt pln pts dts tpr\_shadow vnmi flexpriority ept vpid fsgsbase smep erms

bogomips : 7020.70

clflush size : 64

cache\_alignment : 64

address sizes : 36 bits physical, 48 bits virtual

power management:

processor : 3

vendor\_id : GenuineIntel

cpu family : 6

model : 58

model name : Intel(R) Core(TM) i7-3770K CPU @ 3.50GHz

stepping : 9

cpu MHz : 3510.352

cache size : 8192 KB

physical id : 0

siblings : 8

core id : 3

cpu cores : 4

apicid : 6

initial apicid : 6

fpu : yes

fpu\_exception : yes

cpuid level : 13

wp : yes

flags : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx rdtscp lm constant\_tsc arch\_perfmon pebs bts rep\_good xtopology nonstop\_tsc aperfmperf pni pclmulqdq dtes64 monitor ds\_cpl vmx est tm2 ssse3 cx16 xtpr pdcm pcid sse4\_1 sse4\_2 popcnt tsc\_deadline\_timer aes xsave avx f16c rdrand lahf\_lm ida arat epb xsaveopt pln pts dts tpr\_shadow vnmi flexpriority ept vpid fsgsbase smep erms

bogomips : 7020.70

clflush size : 64

cache\_alignment : 64

address sizes : 36 bits physical, 48 bits virtual

power management:

processor : 4

vendor\_id : GenuineIntel

cpu family : 6

model : 58

model name : Intel(R) Core(TM) i7-3770K CPU @ 3.50GHz

stepping : 9

cpu MHz : 3510.352

cache size : 8192 KB

physical id : 0

siblings : 8

core id : 0

cpu cores : 4

apicid : 1

initial apicid : 1

fpu : yes

fpu\_exception : yes

cpuid level : 13

wp : yes

flags : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx rdtscp lm constant\_tsc arch\_perfmon pebs bts rep\_good xtopology nonstop\_tsc aperfmperf pni pclmulqdq dtes64 monitor ds\_cpl vmx est tm2 ssse3 cx16 xtpr pdcm pcid sse4\_1 sse4\_2 popcnt tsc\_deadline\_timer aes xsave avx f16c rdrand lahf\_lm ida arat epb xsaveopt pln pts dts tpr\_shadow vnmi flexpriority ept vpid fsgsbase smep erms

bogomips : 7020.70

clflush size : 64

cache\_alignment : 64

address sizes : 36 bits physical, 48 bits virtual

power management:

processor : 5

vendor\_id : GenuineIntel

cpu family : 6

model : 58

model name : Intel(R) Core(TM) i7-3770K CPU @ 3.50GHz

stepping : 9

cpu MHz : 3510.352

cache size : 8192 KB

physical id : 0

siblings : 8

core id : 1

cpu cores : 4

apicid : 3

initial apicid : 3

fpu : yes

fpu\_exception : yes

cpuid level : 13

wp : yes

flags : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx rdtscp lm constant\_tsc arch\_perfmon pebs bts rep\_good xtopology nonstop\_tsc aperfmperf pni pclmulqdq dtes64 monitor ds\_cpl vmx est tm2 ssse3 cx16 xtpr pdcm pcid sse4\_1 sse4\_2 popcnt tsc\_deadline\_timer aes xsave avx f16c rdrand lahf\_lm ida arat epb xsaveopt pln pts dts tpr\_shadow vnmi flexpriority ept vpid fsgsbase smep erms

bogomips : 7020.70

clflush size : 64

cache\_alignment : 64

address sizes : 36 bits physical, 48 bits virtual

power management:

processor : 6

vendor\_id : GenuineIntel

cpu family : 6

model : 58

model name : Intel(R) Core(TM) i7-3770K CPU @ 3.50GHz

stepping : 9

cpu MHz : 3510.352

cache size : 8192 KB

physical id : 0

siblings : 8

core id : 2

cpu cores : 4

apicid : 5

initial apicid : 5

fpu : yes

fpu\_exception : yes

cpuid level : 13

wp : yes

flags : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx rdtscp lm constant\_tsc arch\_perfmon pebs bts rep\_good xtopology nonstop\_tsc aperfmperf pni pclmulqdq dtes64 monitor ds\_cpl vmx est tm2 ssse3 cx16 xtpr pdcm pcid sse4\_1 sse4\_2 popcnt tsc\_deadline\_timer aes xsave avx f16c rdrand lahf\_lm ida arat epb xsaveopt pln pts dts tpr\_shadow vnmi flexpriority ept vpid fsgsbase smep erms

bogomips : 7020.70

clflush size : 64

cache\_alignment : 64

address sizes : 36 bits physical, 48 bits virtual

power management:

processor : 7

vendor\_id : GenuineIntel

cpu family : 6

model : 58

model name : Intel(R) Core(TM) i7-3770K CPU @ 3.50GHz

stepping : 9

cpu MHz : 3510.352

cache size : 8192 KB

physical id : 0

siblings : 8

core id : 3

cpu cores : 4

apicid : 7

initial apicid : 7

fpu : yes

fpu\_exception : yes

cpuid level : 13

wp : yes

flags : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx rdtscp lm constant\_tsc arch\_perfmon pebs bts rep\_good xtopology nonstop\_tsc aperfmperf pni pclmulqdq dtes64 monitor ds\_cpl vmx est tm2 ssse3 cx16 xtpr pdcm pcid sse4\_1 sse4\_2 popcnt tsc\_deadline\_timer aes xsave avx f16c rdrand lahf\_lm ida arat epb xsaveopt pln pts dts tpr\_shadow vnmi flexpriority ept vpid fsgsbase smep erms

bogomips : 7020.70

clflush size : 64

cache\_alignment : 64

address sizes : 36 bits physical, 48 bits virtual

power management:

[root@zhuangdizhu1 ~]#

[root@zhuangdizhu2 ~]# dmesg |grep -E 'IOMMU'

Intel-IOMMU: enabled

Reference:

1. <http://www.mellanox.com/pdf/MFT/MFT_user_manual.pdf>
2. <https://community.mellanox.com/docs/DOC-1317>
3. <http://www.mellanox.com/related-docs/prod_software/Mellanox_OFED_Linux_User_Manual_v2.4-1.0.0.pdf>