InfiniBand in the Lab

Erik Bussink

@ErikBussink

www.bussink.ch

www.VSAN.info

Executive Summary

Fast Cheap

Who uses InfiniBand for Storage ?

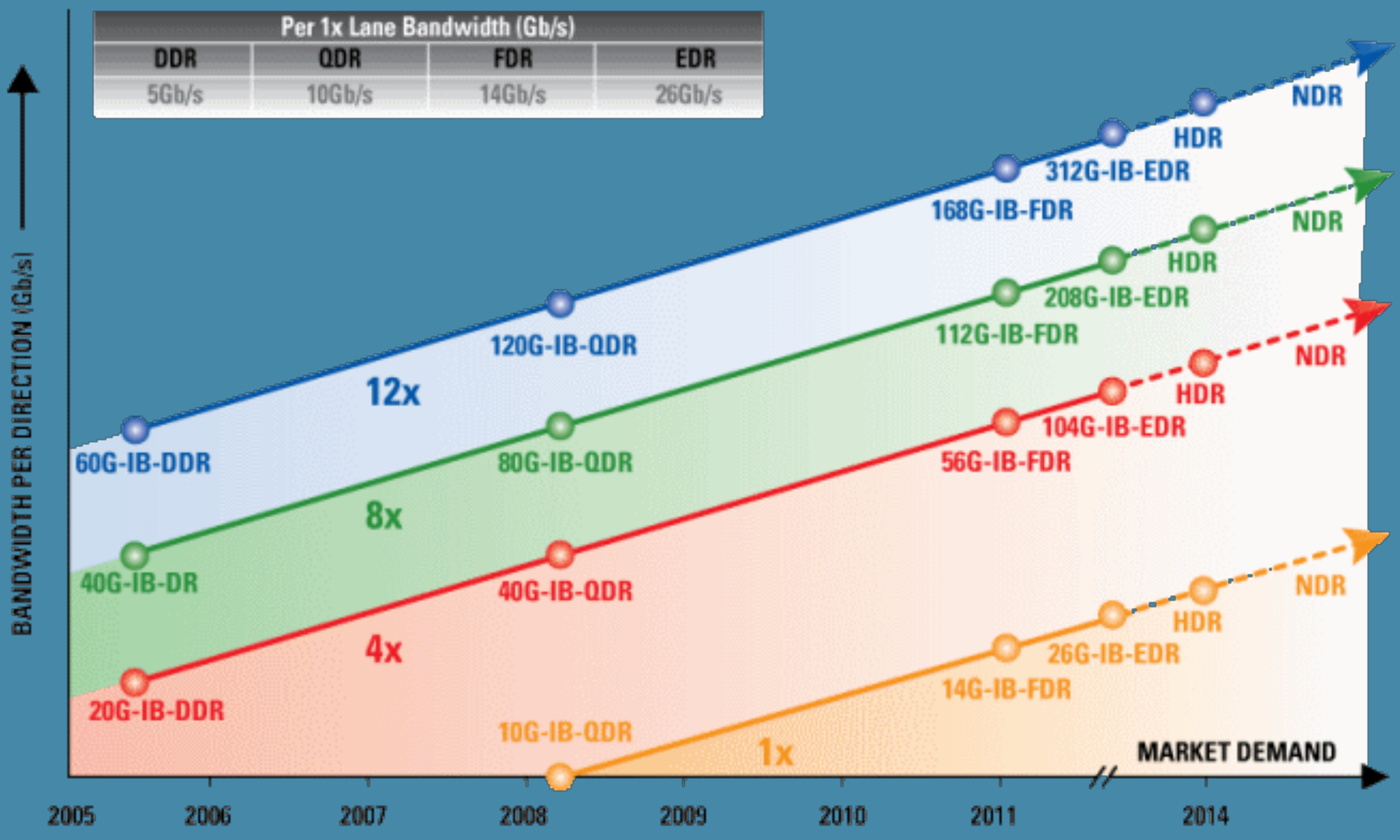
- EMC Isilon
- EMC xtremIO
- PureStorage
- Oracle Exadata
- Nexenta
- TeraData
- Gluster

What InfiniBand Offers

- Lowest Layer for scalable IO interconnect
  - High-Performance
  - Low Latency
  - Reliable Switch Fabric
- Offers higher layers of functionality
  - Application Clustering
  - Fast Inter Process Communications
  - Storage Area Networks

InfiniBand physical layer

- Physical layer based on 802.3z specification operating at 2.5Gb/s same standard as 10G ethernet (3.125Gb/s)
- InfiniBand layer 2 switching uses 16 bit local address (LID), so limited to 2^16 nodes on a subnet.



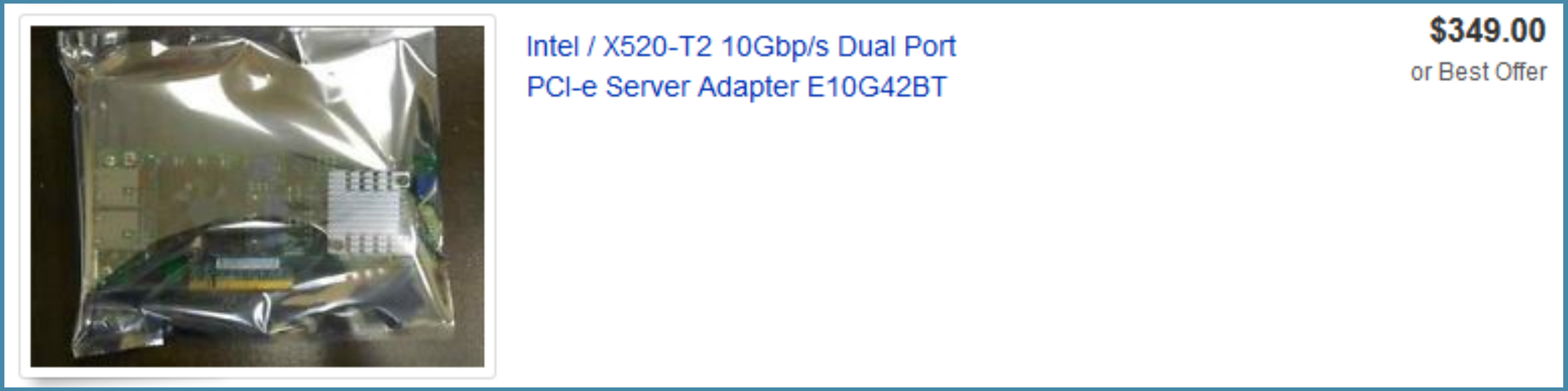
Per 1x Lane Bandwidth (Gb/s)			
DDR	QDR	FDR	EDR
5Gb/s	10Gb/s	14Gb/s	26Gb/s

Connectors

- For SDR (10Gb/s) and DDR (20Gb/s) use the CX4 connectors

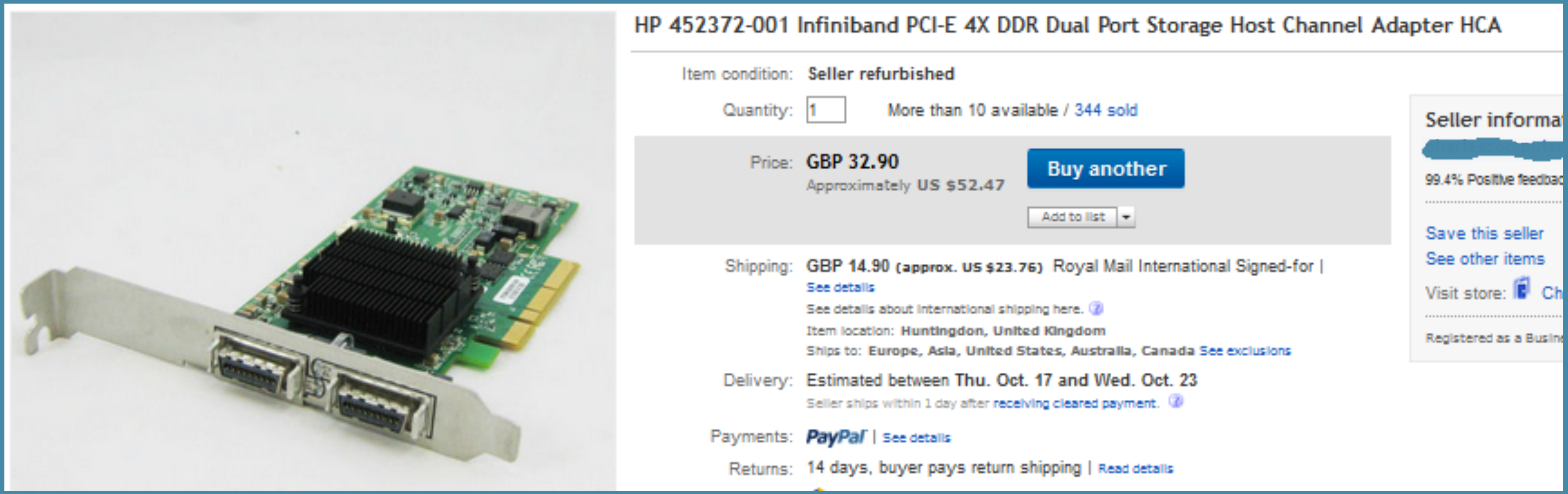
- For QDR (40Gb/s) and FDR (56Gb/s) use QSFP connectors

Price comparison on Adapters



Intel / X520-T2 10Gbp/s Dual Port PCI-e Server Adapter E10G42BT

$349.00
or Best Offer



HP 452372-001 Infiniband PCI-E 4X DDR Dual Port Storage Host Channel Adapter HCA

Item condition: Seller refurbished

Quantity: 1 More than 10 available / 344 sold

Price: GBP 32.90
Approximately US $52.47

Shipping: GBP 14.90 (approx. US $23.76) Royal Mail International Signed-for | See details
See details about international shipping here.
Item location: Huntingdon, United Kingdom
Ships to: Europe, Asia, United States, Australia, Canada See exclusions

Delivery: Estimated between Thu. Oct. 17 and Wed. Oct. 23
Seller ships within 1 day after receiving cleared payment.

Payments: PayPal | See details

Returns: 14 days, buyer pays return shipping | Read details

Seller informat
[redacted]
99.4% Positive feedbac

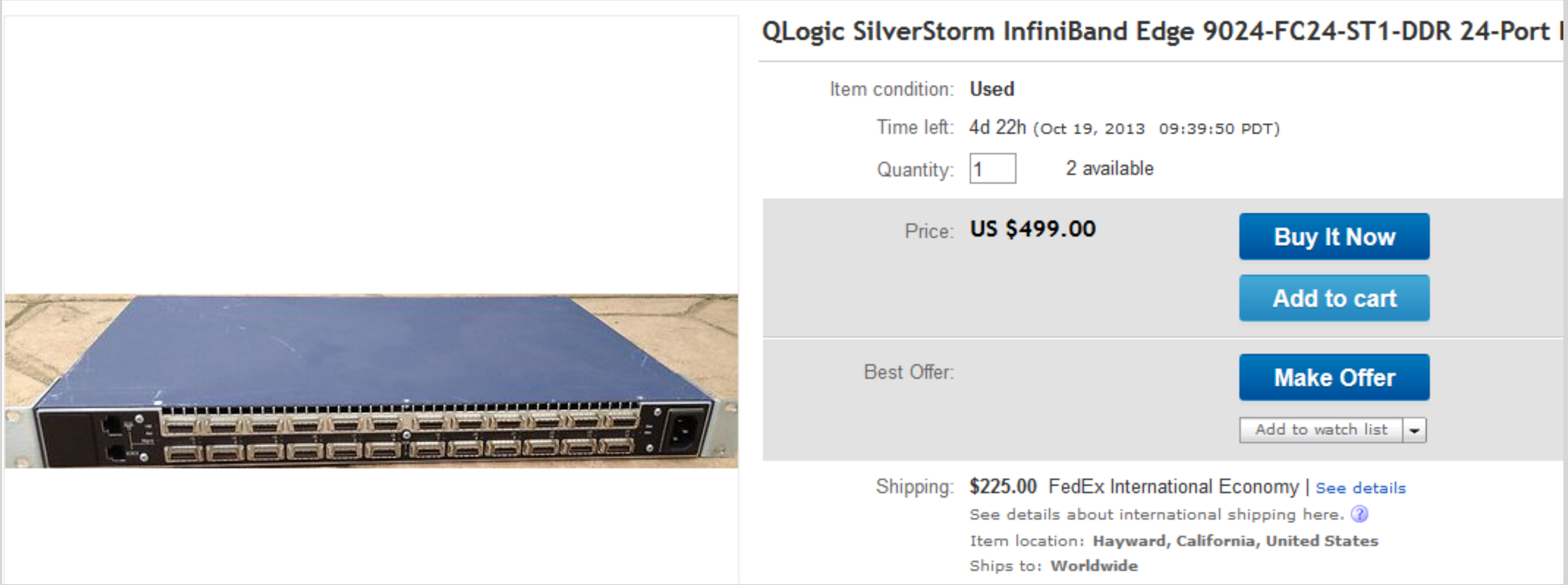
Save this seller
See other items
Visit store: Ch

Registered as a Busin

And switches too...

Example... 24 ports DDR (20Gb/s) for $499

Latest firmware (4.2.5) has embedded Subnet Manager



QLogic SilverStorm InfiniBand Edge 9024-FC24-ST1-DDR 24-Port

Item condition: **Used**

Time left: 4d 22h (Oct 19, 2013 09:39:50 PDT)

Quantity: 1 2 available

Price: **US $499.00**

Best Offer:

Shipping: **$225.00** FedEx International Economy | See details

See details about international shipping here.

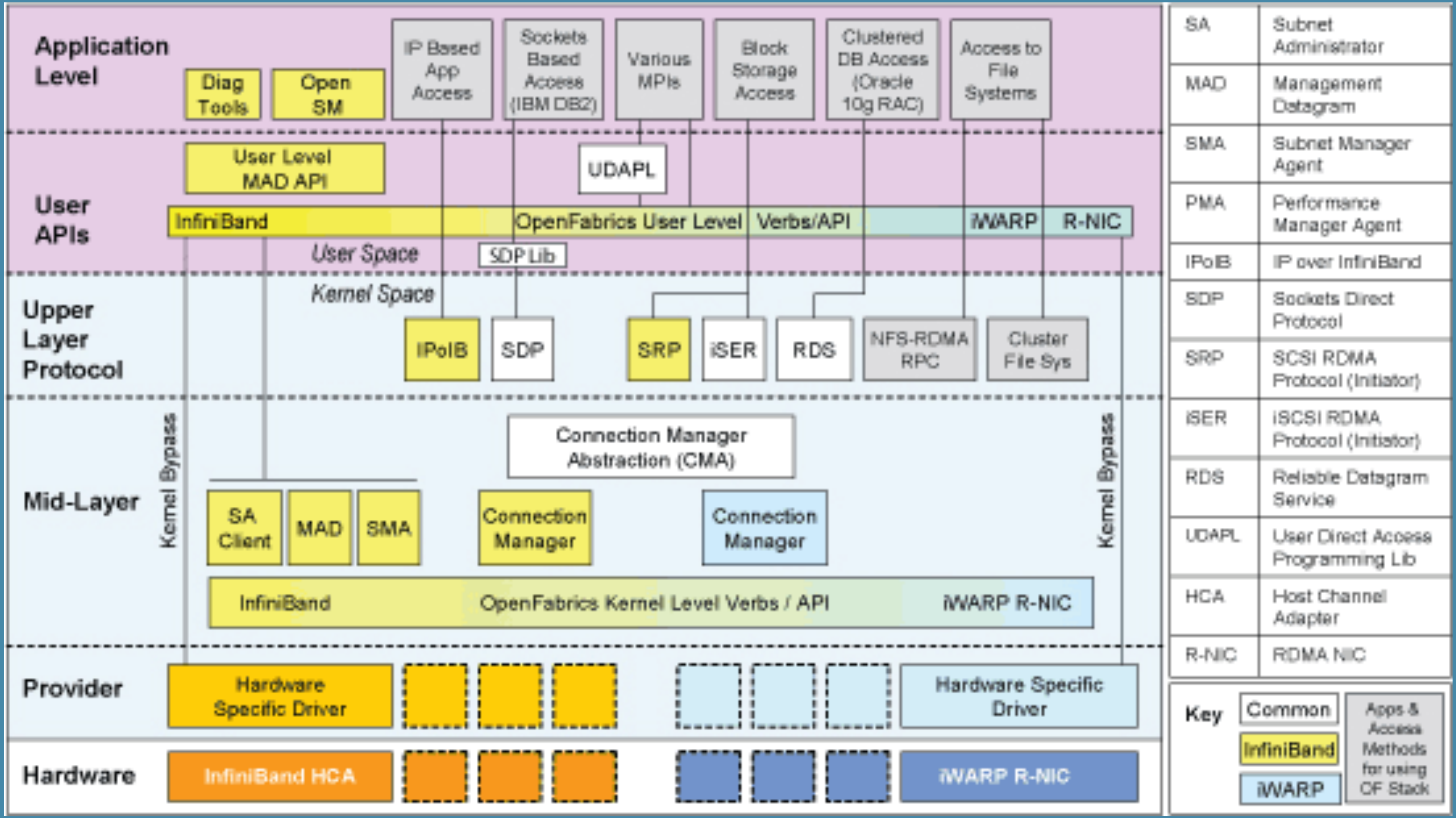
Item location: **Hayward, California, United States**

Ships to: **Worldwide**

Subnet Manager

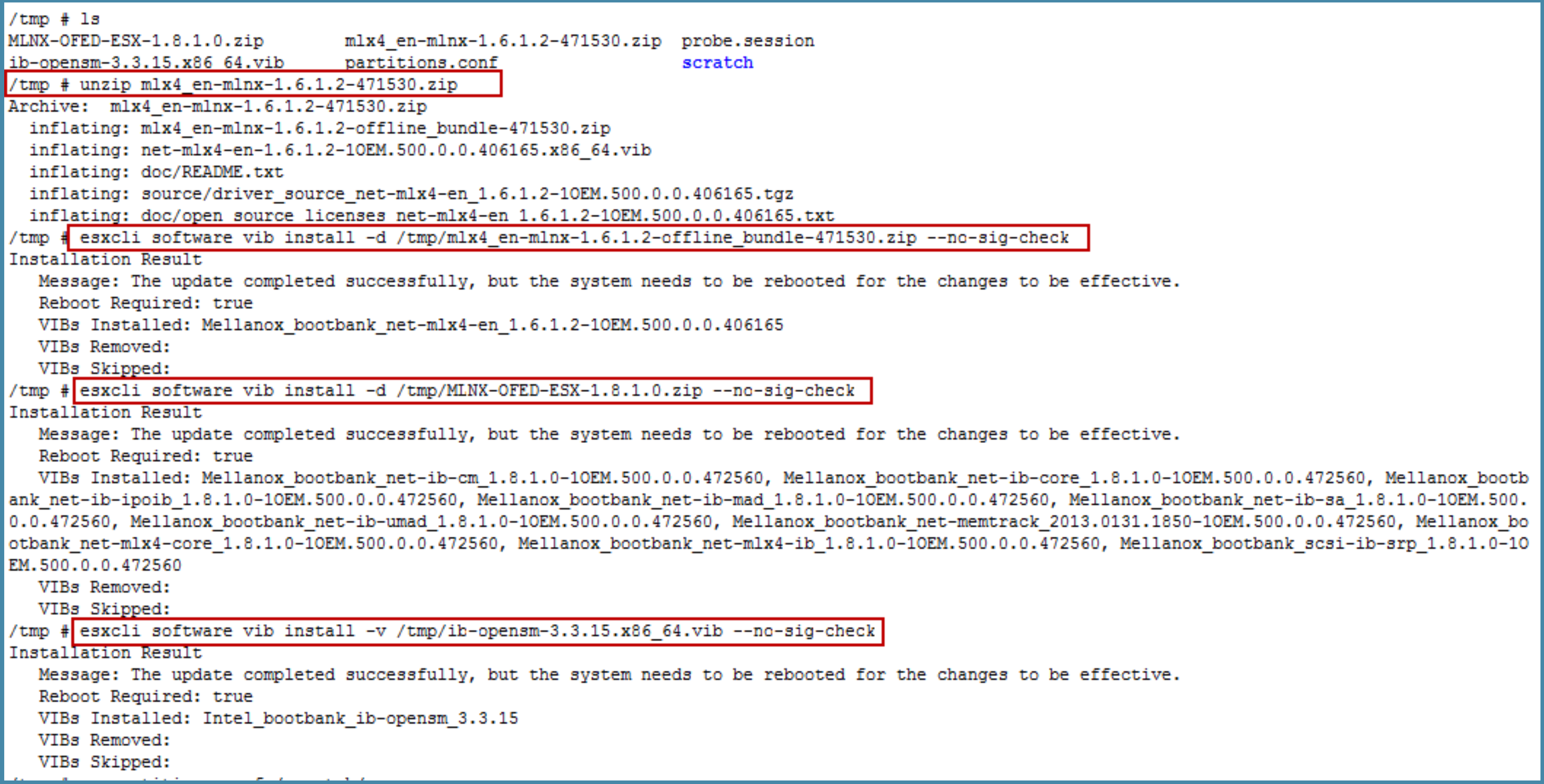
- The Subnet Manager assigns Local IDentifiers (LIDs) to each port in the IB fabric, and develops a routing table based off the assigned LIDs
- Hardware based Subnet Manager are located in the Switches
- Software based Subnet Manager is located on a host connected to the IB fabric
- The OpenSM from the OpenFabrics Alliance works on Windows and Linux

OpenFabrics Enterprise Distribution (OFED)



SA	Subnet Administrator
MAD	Management Datagram
SMA	Subnet Manager Agent
PMA	Performance Manager Agent
IPoIB	IP over InfiniBand
SDP	Sockets Direct Protocol
SRP	SCSI RDMA Protocol (Initiator)
iSER	iSCSI RDMA Protocol (Initiator)
RDS	Reliable Datagram Service
UDAPL	User Direct Access Programming Lib
HCA	Host Channel Adapter
R-NIC	RDMA NIC

Installing InfiniBand on vSphere 5.1

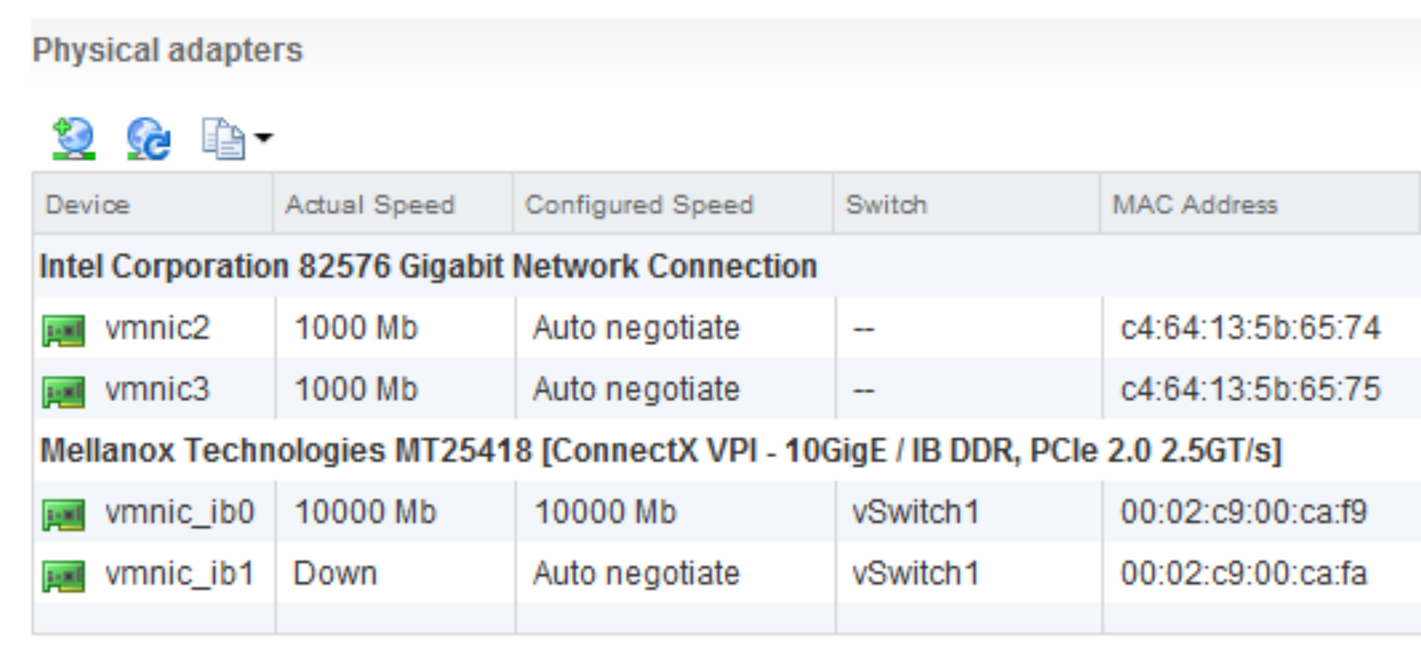


```
/tmp # ls
MLNX-OFED-ESX-1.8.1.0.zip        mlx4_en-mlnx-1.6.1.2-471530.zip  probe.session
ib-opensm-3.3.15.x86_64.vib      partitions.conf                  scratch
/tmp # unzip mlx4_en-mlnx-1.6.1.2-471530.zip
Archive:  mlx4_en-mlnx-1.6.1.2-471530.zip
  inflating: mlx4_en-mlnx-1.6.1.2-offline_bundle-471530.zip
  inflating: net-mlx4-en-1.6.1.2-1OEM.500.0.0.406165.x86_64.vib
  inflating: doc/README.txt
  inflating: source/driver_source_net-mlx4-en_1.6.1.2-1OEM.500.0.0.406165.tgz
  inflating: doc/open_source_licenses_net-mlx4-en_1.6.1.2-1OEM.500.0.0.406165.txt
/tmp # esxcli software vib install -d /tmp/mlx4_en-mlnx-1.6.1.2-offline_bundle-471530.zip --no-sig-check
Installation Result
   Message: The update completed successfully, but the system needs to be rebooted for the changes to be effective.
   Reboot Required: true
   VIBs Installed: Mellanox_bootbank_net-mlx4-en_1.6.1.2-1OEM.500.0.0.406165
   VIBs Removed:
   VIBs Skipped:
/tmp # esxcli software vib install -d /tmp/MLNX-OFED-ESX-1.8.1.0.zip --no-sig-check
Installation Result
   Message: The update completed successfully, but the system needs to be rebooted for the changes to be effective.
   Reboot Required: true
   VIBs Installed: Mellanox_bootbank_net-ib-cm_1.8.1.0-1OEM.500.0.0.472560, Mellanox_bootbank_net-ib-core_1.8.1.0-1OEM.500.0.0.472560, Mellanox_bootb
ank_net-ib-ipoib_1.8.1.0-1OEM.500.0.0.472560, Mellanox_bootbank_net-ib-mad_1.8.1.0-1OEM.500.0.0.472560, Mellanox_bootbank_net-ib-sa_1.8.1.0-1OEM.500.
0.0.472560, Mellanox_bootbank_net-ib-umad_1.8.1.0-1OEM.500.0.0.472560, Mellanox_bootbank_net-memtrack_2013.0131.1850-1OEM.500.0.0.472560, Mellanox_bo
otbank_net-mlx4-core_1.8.1.0-1OEM.500.0.0.472560, Mellanox_bootbank_net-mlx4-ib_1.8.1.0-1OEM.500.0.0.472560, Mellanox_bootbank_scsi-ib-srp_1.8.1.0-1O
EM.500.0.0.472560
   VIBs Removed:
   VIBs Skipped:
/tmp # esxcli software vib install -v /tmp/ib-opensm-3.3.15.x86_64.vib --no-sig-check
Installation Result
   Message: The update completed successfully, but the system needs to be rebooted for the changes to be effective.
   Reboot Required: true
   VIBs Installed: Intel_bootbank_ib-opensm_3.3.15
   VIBs Removed:
   VIBs Skipped:
```

Configure MTU and OpenSM

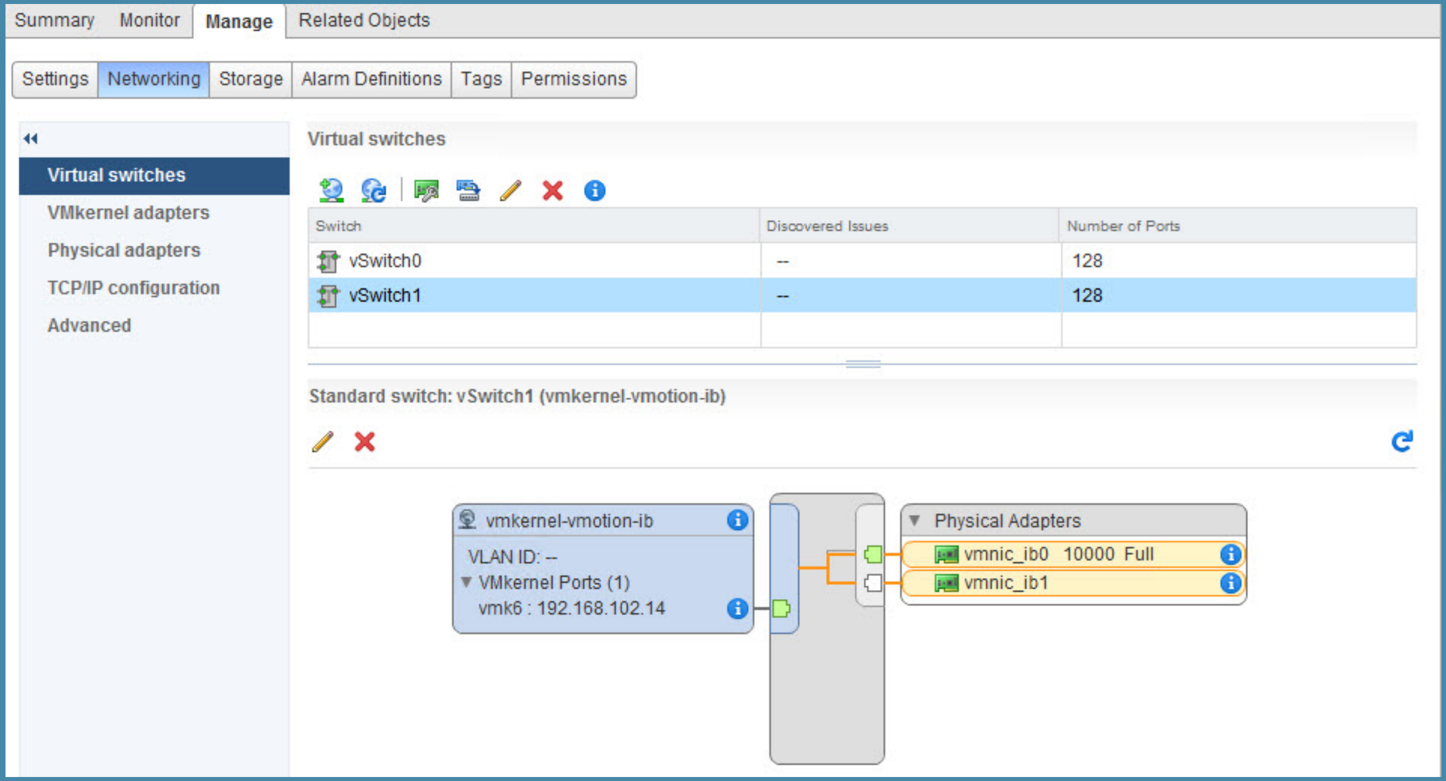
- esxcli system module paramters set y-m=mlx4_core -p=mtu_4k=1
- On Old switches (No 4K support) MTU is 2044 (2048-4) for IPoIB

- vi partitions.conf
- Add single line "Default=ox7fff,ipoib,mtu=5:ALL=full;"
- copy partitions.conf /scratch/opensm/adapter_1_hca/
- copy partitions.conf /scratch/opensm/adapter_2_hca/



Physical adapters

Device	Actual Speed	Configured Speed	Switch	MAC Address
Intel Corporation 82576 Gigabit Network Connection				
vmnic2	1000 Mb	Auto negotiate	--	c4:64:13:5b:65:74
vmnic3	1000 Mb	Auto negotiate	--	c4:64:13:5b:65:75
Mellanox Technologies MT25418 [ConnectX VPI - 10GigE / IB DDR, PCIe 2.0 2.5GT/s]				
vmnic_ib0	10000 Mb	10000 Mb	vSwitch1	00:02:c9:00:ca:f9
vmnic_ib1	Down	Auto negotiate	vSwitch1	00:02:c9:00:ca:fa



Virtual switches

Switch	Discovered Issues	Number of Ports
vSwitch0	--	128
vSwitch1	--	128

Standard switch: vSwitch1 (vmkernel-vmotion-ib)

vmkernel-vmotion-ib

VLAN ID: --

VMkernel Ports (1)

vmk6 : 192.168.102.14

Physical Adapters

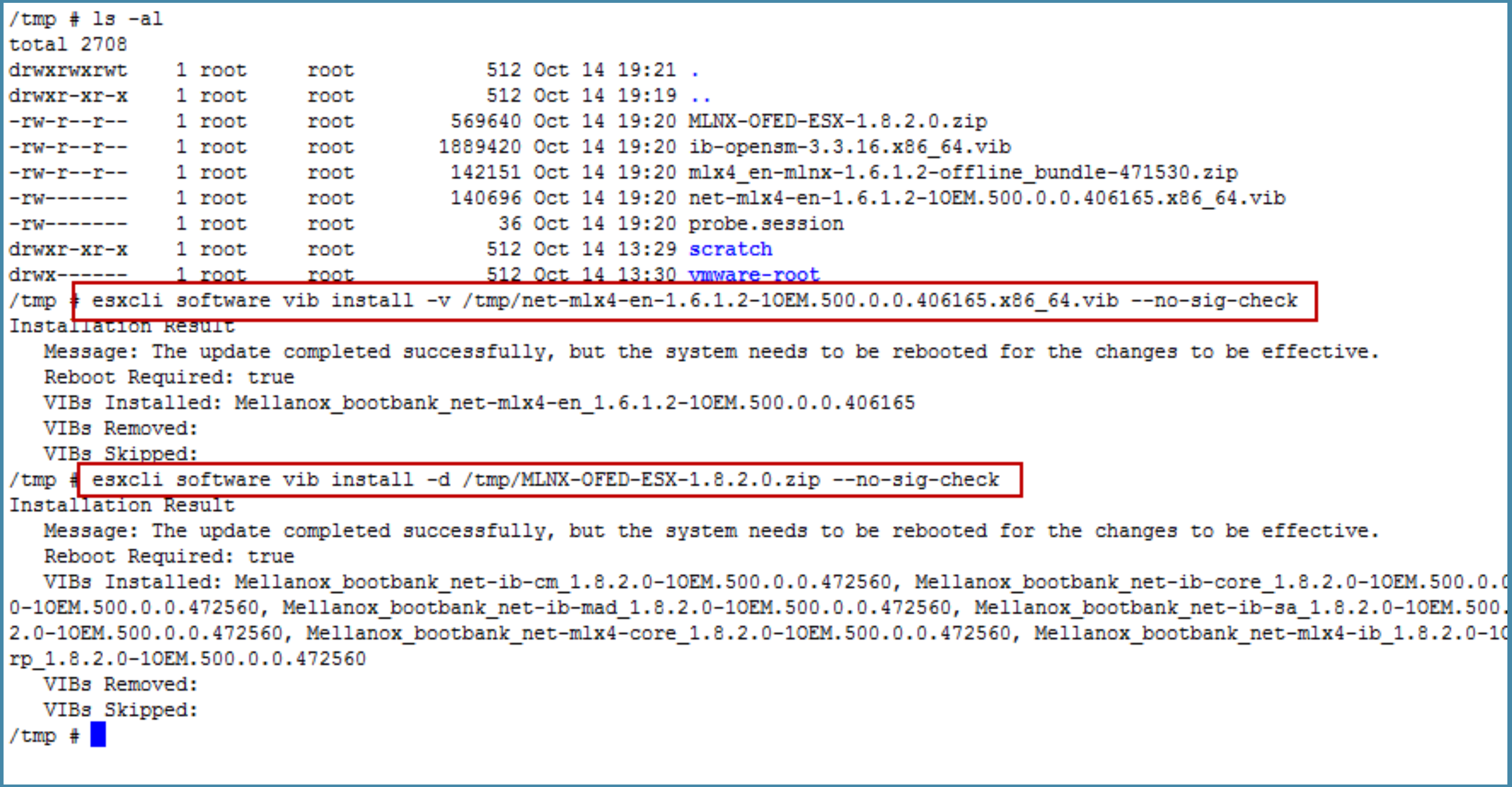
vmnic_ib0 10000 Full

vmnic_ib1

Installing InfiniBand on vSphere 5.5

- Mellanox drivers are in vSphere 5.5, and they support 40GbE inbox when using ConnectX-3and SwitchX products.
- If not same HCA you need to uninstall the Mellanox drivers
- esxcli software vib remove -n=net-mlx4-en -n=net-mlx4-core
- And reboot
- Then install the Mellanox OFED 1.8.2
- esxcli software vib install -d /tmp/MLX
- And reboot

Installing InfiniBand on vSphere 5.5

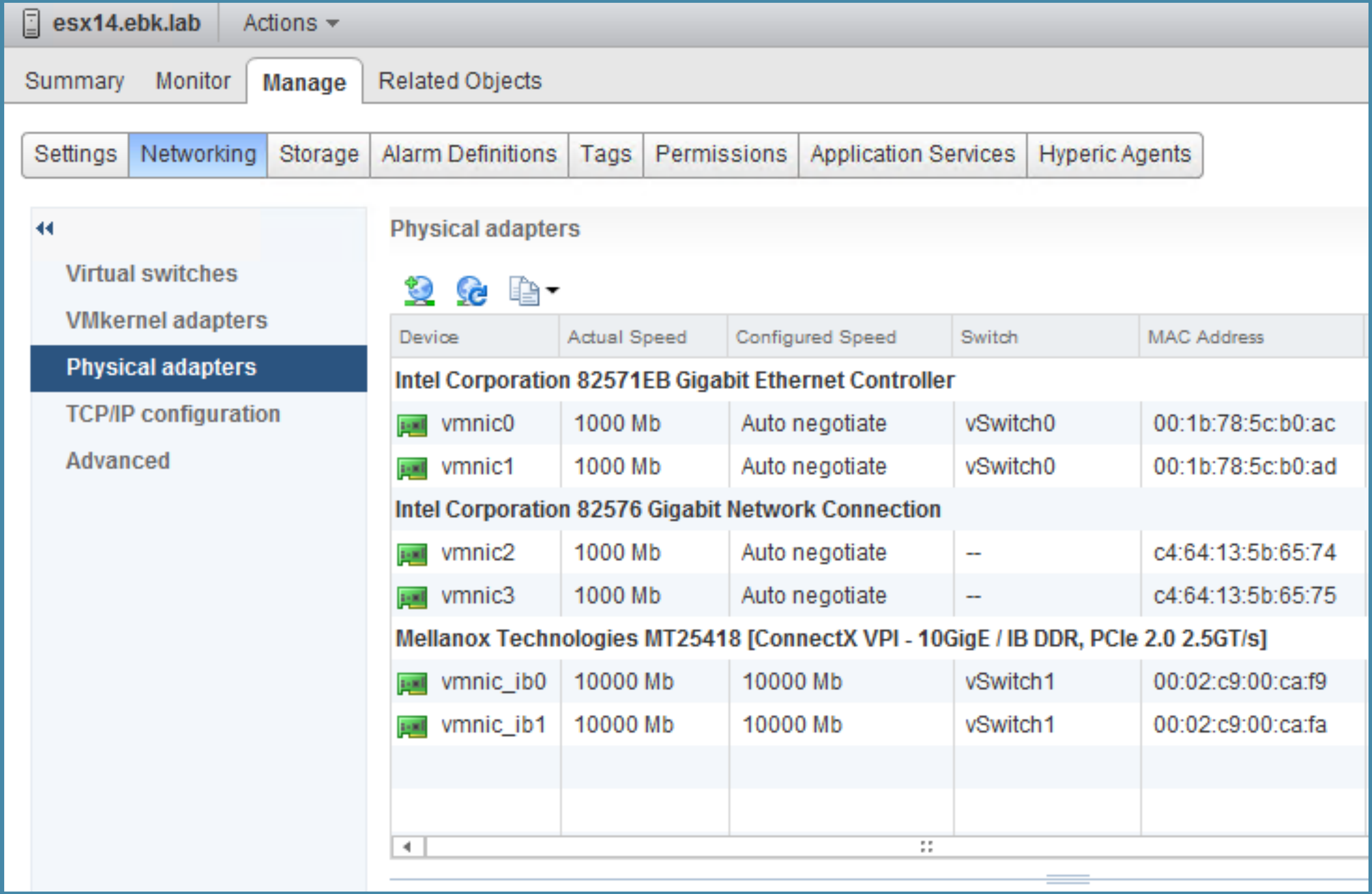


```
/tmp # ls -al
total 2708
drwxrwxrwt    1 root     root           512 Oct 14 19:21 .
drwxr-xr-x    1 root     root           512 Oct 14 19:19 ..
-rw-r--r--    1 root     root        569640 Oct 14 19:20 MLNX-OFED-ESX-1.8.2.0.zip
-rw-r--r--    1 root     root       1889420 Oct 14 19:20 ib-opensm-3.3.16.x86_64.vib
-rw-r--r--    1 root     root        142151 Oct 14 19:20 mlx4_en-mlnx-1.6.1.2-offline_bundle-471530.zip
-rw-------    1 root     root        140696 Oct 14 19:20 net-mlx4-en-1.6.1.2-1OEM.500.0.0.406165.x86_64.vib
-rw-------    1 root     root            36 Oct 14 19:20 probe.session
drwxr-xr-x    1 root     root           512 Oct 14 13:29 scratch
drwx------    1 root     root           512 Oct 14 13:30 vmware-root
/tmp # esxcli software vib install -v /tmp/net-mlx4-en-1.6.1.2-1OEM.500.0.0.406165.x86_64.vib --no-sig-check
Installation Result
   Message: The update completed successfully, but the system needs to be rebooted for the changes to be effective.
   Reboot Required: true
   VIBs Installed: Mellanox_bootbank_net-mlx4-en_1.6.1.2-1OEM.500.0.0.406165
   VIBs Removed:
   VIBs Skipped:
/tmp # esxcli software vib install -d /tmp/MLNX-OFED-ESX-1.8.2.0.zip --no-sig-check
Installation Result
   Message: The update completed successfully, but the system needs to be rebooted for the changes to be effective.
   Reboot Required: true
   VIBs Installed: Mellanox_bootbank_net-ib-cm_1.8.2.0-1OEM.500.0.0.472560, Mellanox_bootbank_net-ib-core_1.8.2.0-1OEM.500.0.0
0-1OEM.500.0.0.472560, Mellanox_bootbank_net-ib-mad_1.8.2.0-1OEM.500.0.0.472560, Mellanox_bootbank_net-ib-sa_1.8.2.0-1OEM.500.
2.0-1OEM.500.0.0.472560, Mellanox_bootbank_net-mlx4-core_1.8.2.0-1OEM.500.0.0.472560, Mellanox_bootbank_net-mlx4-ib_1.8.2.0-1O
rp_1.8.2.0-1OEM.500.0.0.472560
   VIBs Removed:
   VIBs Skipped:
/tmp #
```

OpenSM and vSphere 5.5

- Currently ib-opensm-3.3.16 installs on vSphere 5.5 but doesn't see the IB ports

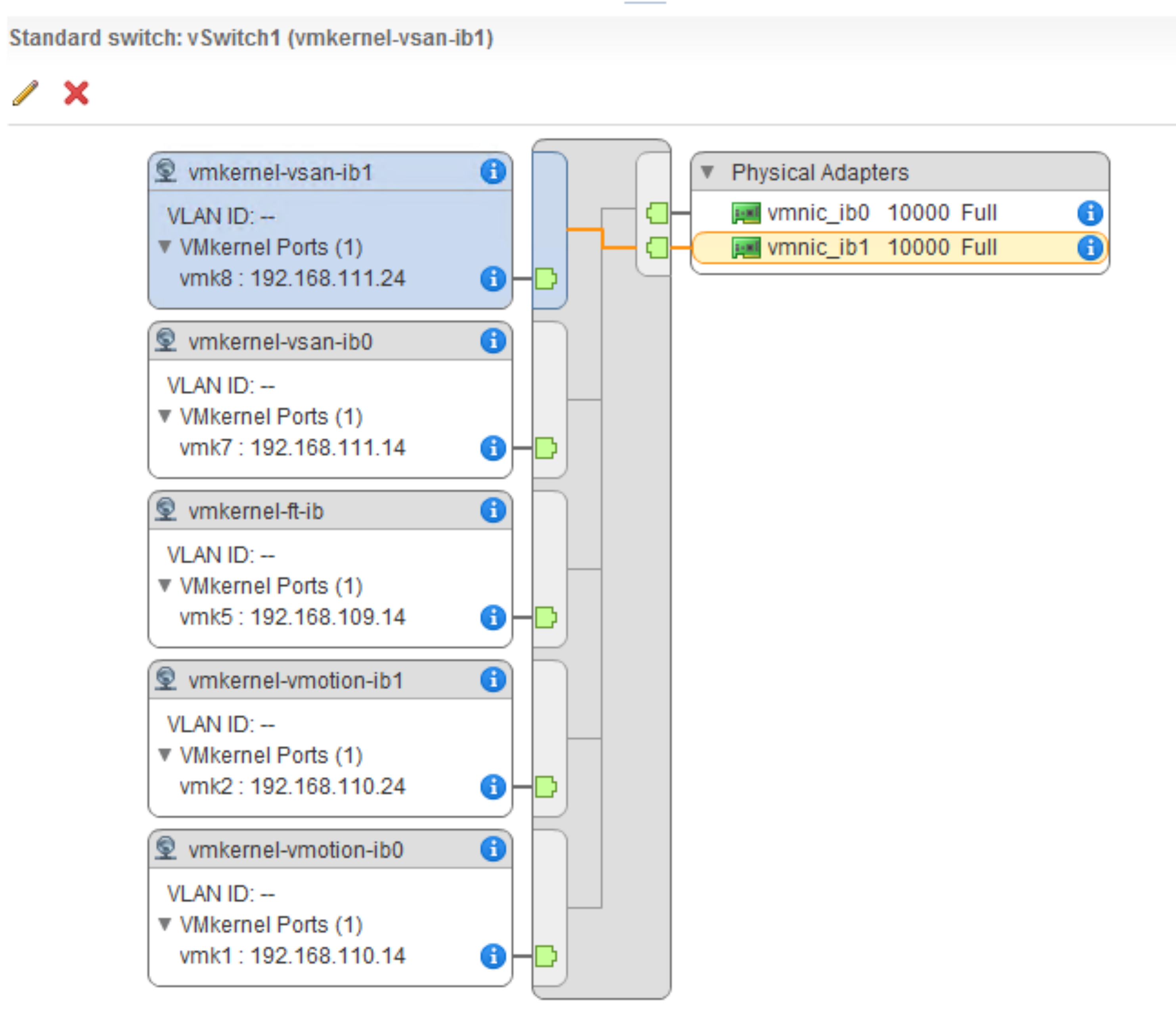
- Case 1 have a Switch with Subnet Manager
- Case 2 add a host (CentOS) with an IB HCA and run OpenSM on it
- Case 3 waiting for Raphael Schitz (@hypervizor_fr) to pull a magic rabbit out of his hat before this presentation !



Physical adapters

Device	Actual Speed	Configured Speed	Switch	MAC Address
Intel Corporation 82571EB Gigabit Ethernet Controller				
vmnic0	1000 Mb	Auto negotiate	vSwitch0	00:1b:78:5c:b0:ac
vmnic1	1000 Mb	Auto negotiate	vSwitch0	00:1b:78:5c:b0:ad
Intel Corporation 82576 Gigabit Network Connection				
vmnic2	1000 Mb	Auto negotiate	--	c4:64:13:5b:65:74
vmnic3	1000 Mb	Auto negotiate	--	c4:64:13:5b:65:75
Mellanox Technologies MT25418 [ConnectX VPI - 10GigE / IB DDR, PCIe 2.0 2.5GT/s]				
vmnic_ib0	10000 Mb	10000 Mb	vSwitch1	00:02:c9:00:ca:f9
vmnic_ib1	10000 Mb	10000 Mb	vSwitch1	00:02:c9:00:ca:fa

InfiniBand IPoIB backbone for VSAN



Standard switch: vSwitch1 (vmkernel-vsan-ib1)

vmkernel-vsan-ib1

VLAN ID: --

▼ VMkernel Ports (1)

vmk8 : 192.168.111.24

vmkernel-vsan-ib0

VLAN ID: --

▼ VMkernel Ports (1)

vmk7 : 192.168.111.14

vmkernel-ft-ib

VLAN ID: --

▼ VMkernel Ports (1)

vmk5 : 192.168.109.14

vmkernel-vmotion-ib1

VLAN ID: --

▼ VMkernel Ports (1)

vmk2 : 192.168.110.24

vmkernel-vmotion-ib0

VLAN ID: --

▼ VMkernel Ports (1)

vmk1 : 192.168.110.14

▼ Physical Adapters

vmnic_ib0 10000 Full

vmnic_ib1 10000 Full

InfiniBand in the Lab

Fast & Cheap

Thanks to Raphael Schitz,William Lam,
Vladan Seget, Gregory Roche

Erik Bussink

@ErikBussink

www.bussink.ch

www.VSAN.info