Infiniband

Beschreibung

Infiband Karten sind eine günstige Lösung um schnelles Ethernet zu bauen. Eigentlich sind die Karten urspürnglich dazu angedacht SANs bereitzustellen, aber wir können uns dieses auch als 40 GBit Netzwerk zu nutze machen.

Vorraussetzungen:

Infiniband Karten
Infiniband switch
Infiband Kabel

Installation

Insalltallation Infiniband Karte

Installieren der Tools

```
apt-get install infiniband-diags ibutils iperf ethtool
```

Nun laden wir die Module

```
modprobe ib_ipoib
modprobe ib_umad
```

Je nach System heißen die Adapter anders.
Um zu sehen ob die Treiber geladen wurden dann je nach System

```
ip a | grep ib0
ip a | grep ib1


oder
```

```
ip a | grep ibp
```

Sieht die Ausgabe dann so aus (entweder mit ib0 oder ib1 oder halt diese...

```
ip a | grep ibp
13: ibp6s16: <BROADCAST,MULTICAST> mtu 4092 qdisc noop state DOWN group default qlen 256
14: ibp6s16d1: <BROADCAST,MULTICAST> mtu 4092 qdisc noop state DOWN group default qlen 256
```

!!!!Hinweis, sollte kine ib0 oder ipb aufgelistet sein laufen die Karten eventuell im Ethernet modus!!!

```
$ ip a
# InfiniBand Mode
4: ibp129s0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 2044 qdisc mq state UP group default qlen 256
    link/infiniband ...
    inet 192.168.7.100/24 brd 192.168.7.255 scope global ibp129s0
       valid_lft forever preferred_lft forever

# Ethernet Mode
7: ens1f1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq state UP group default qlen 1000
    link/ether ...
    inet 10.200.0.1/24 brd 10.200.0.255 scope global ens1f1
       valid_lft forever preferred_lft forever
```

Diese dann eben der modules Datei hinzufügen, so das diese dann auch beim Systemstart geladen werden.

```
echo ib_umad >> /etc/modules
echo ib_ipoib >> /etc/modules
```

Netzwerk Setup

Die Infinibandkarten werden genauso konfiguriert wie Netzwerkkarten
nur mit dem unterschied das wir noch den Mode angeben.

In meinem Beispiel heißen die Karten (Ports) so:

ibp6s16
ibp6s16d1

```
#/etc/network/interfaces
auto ibp6s16
```

```
iface ibp6s16 inet static
  address 172.30.128.75
  netmask 255.255.240.0
  broadcast 172.30.143.255
  pre-up echo connected > /sys/class/net/ibp6s16/mode
  mtu 65520

auto ibp6s16d1
iface ibp6s16d1 inet static
 address 172.31.128.75
 netmask 255.255.240.0
 broadcast 172.31.143.255
 pre-up echo connected > /sys/class/net/ibp6s16d1/mode
 mtu 65520
```

Nun die Adapter starten

```
ifup ibp6s16
ifup ibp6s16d1
```

Nun pingen wir unsere eigene Karte an.

```
ping 172.30.128.75
```

Ausgabe, wenn das klappt ist die Konfiguration abgeschlossen.

```
ING 172.30.128.75 (172.30.128.75) 56(84) bytes of data.
64 bytes from 172.30.128.75: icmp_seq=1 ttl=64 time=0.017 ms
64 bytes from 172.30.128.75: icmp_seq=2 ttl=64 time=0.021 ms
^C
--- 172.30.128.75 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1002ms
rtt min/avg/max/mdev = 0.017/0.019/0.021/0.002 ms
```

mit dem Befehl ibstat sehen wir den Status der und Eigenschaften der Karte (Ports)

```
ibstat
```

Ausgabe:

```
ibstat
CA 'mlx4_0'
	CA type: MT4099
	Number of ports: 2
	Firmware version: 2.42.5000
	Hardware version: 1
	Node GUID: 0xf4521403009871a0
	System image GUID: 0xf4521403009871a3
	Port 1:
		State: Initializing
		Physical state: LinkUp
		Rate: 40
		Base lid: 0
		LMC: 0
		SM lid: 0
		Capability mask: 0x02514868
		Port GUID: 0xf4521403009871a1
		Link layer: InfiniBand
	Port 2:
		State: Initializing
		Physical state: LinkUp
		Rate: 40
		Base lid: 0
		LMC: 0
		SM lid: 0
		Capability mask: 0x02514868
		Port GUID: 0xf4521403009871a2
		Link layer: InfiniBand
root@vserv0003:~#
```

Wie man sieht bleibt der Status auf initilazing stehen, das liegt daran weil wir noch keinen Subnetmanager haben.
Einige Switche bringen auch einen SM Manager mit.
Ich empfehle aber, diesen zu deaktivieren und auf den Hosts einen SM Manager zu installieren, mit Prioritäten

Installation OPENSM Manager (Ein Manager für Subnetze)

Pakete installieren

```
apt-get install opensm
```

Nachdem der Manager installiert ist, ist der Status der Karten aktiv.

Ausgabe ibstat

```
ibstat
CA 'mlx4_0'
	CA type: MT4099
	Number of ports: 2
	Firmware version: 2.42.5000
	Hardware version: 1
	Node GUID: 0xf4521403009871a0
	System image GUID: 0xf4521403009871a3
	Port 1:
		State: Active
		Physical state: LinkUp
		Rate: 40
		Base lid: 1
		LMC: 0
		SM lid: 1
		Capability mask: 0x0251486a
		Port GUID: 0xf4521403009871a1
		Link layer: InfiniBand
	Port 2:
		State: Active
		Physical state: LinkUp
		Rate: 40
		Base lid: 3
		LMC: 0
		SM lid: 1
		Capability mask: 0x0251486a
		Port GUID: 0xf4521403009871a2
		Link layer: InfiniBand
```

Mit dem tool ethtool sehen wir dann auch die Ethernetgeschwindigkeit.

```
erhtool ibp6s16

Ausgabe:

Settings for ibp6s16:
	Supported ports: [  ]
```

Supported link modes: Not reported
Supported pause frame use: No
Supports auto-negotiation: No
Supported FEC modes: Not reported
Advertised link modes: Not reported
Advertised pause frame use: No
Advertised auto-negotiation: No
Advertised FEC modes: Not reported
Speed: 40000Mb/s
Duplex: Full
Auto-negotiation: on
Port: Other
PHYAD: 255
Transceiver: internal
Link detected: yes

Wenn man als Manager redunanz haben möchte, sollte auf jedem Node der Infiniband nutzt, opensm installiert sein.
Denn startet ein Host neu, ist das Netzwerk ohne Subnetmanager.
Selbst wenn man nur einen Host hat, kann man diesen schon konfigurieren.

Bei jedem Server muss die Priorität geandert werden.

Die Prioritäten gehen von 0-15
Die höchste übernimmt.
Das heißt den ersten Server konfigurieren wir mit 15 dann 14 usw. Das heit es können maximal 16 Manager in einem Netz sein.
Eine Redunandanz von MAx 16 sollte eigentlich reichen.

Nun die Konfig erstellen auf der ersten Node

```
opensm --create-config /etc/opensm/opensm.conf
```

Ausgabe:

```
-------------------------------------------------
OpenSM 3.3.23
Command Line Arguments:
 Creating config file template '/etc/opensm/opensm.conf'.
 Log File: /var/log/opensm.log
-------------------------------------------------
```

Priorität in der config ändern. Bei Server 1 der Wert 15. Bei Server 2 = 14 und bei Server 3 = 13

```
nano /etc/opensm/opensm.conf
```

Dort in der nähne von Zeile 258 die Prio ändern.

```
# SM priority used for deciding who is the master
# Range goes from 0 (lowest priority) to 15 (highest).
sm_priority 15
```

Nun den Manager neustarten

```
service opensm restart


Möchte man unbedingt die ausgabe sehen dann folgenden Befehl verwenden
opensm -B

Ausgabe:


-------------------------------------------------
OpenSM 3.3.23
 Reading Cached Option File: /etc/opensm/opensm.conf
 Loading Cached Option:sm_priority = 15
Command Line Arguments:
 Daemon mode
 Log File: /var/log/opensm.log
-------------------------------------------------
```

Mit dem Befehl sehen wir folgende.

```
ibdiagnet
```

Ausgabe:

```
oading IBDIAGNET from: /usr/lib/x86_64-linux-gnu/ibdiagnet1.5.7
-W- Topology file is not specified.
    Reports regarding cluster links will use direct routes.
Loading IBDM from: /usr/lib/x86_64-linux-gnu/ibdm1.5.7
-W- A few ports of local device are up.
    Since port-num was not specified (-p option), port 1 of device 1 will be
    used as the local port.
```

```
-I- Discovering ... 3 nodes (1 Switches & 2 CA-s) discovered.


-I---------------------------------------------------
-I- Bad Guids/LIDs Info
-I---------------------------------------------------
-I- No bad Guids were found

-I---------------------------------------------------
-I- Links With Logical State = INIT
-I---------------------------------------------------
-I- No bad Links (with logical state = INIT) were found

-I---------------------------------------------------
-I- General Device Info
-I---------------------------------------------------

-I---------------------------------------------------
-I- PM Counters Info
-I---------------------------------------------------
-I- No illegal PM counters values were found

-I---------------------------------------------------
-I- Fabric Partitions Report (see ibdiagnet.pkey for a full hosts list)
-I---------------------------------------------------
-I-    PKey:0x7fff Hosts:4 full:4 limited:0

-I---------------------------------------------------
-I- IPoIB Subnets Check
-I---------------------------------------------------
-I- Subnet: IPv4 PKey:0x7fff QKey:0x00000b1b MTU:2048Byte rate:10Gbps SL:0x00
-W- Suboptimal rate for group. Lowest member rate:40Gbps > group-rate:10Gbps

-I---------------------------------------------------
-I- Bad Links Info
-I- No bad link were found
-I---------------------------------------------------
----------------------------------------------------------------
-I- Stages Status Report:
   STAGE                                Errors Warnings
```

```
    Bad GUIDs/LIDs Check                 0      0
    Link State Active Check              0      0
    General Devices Info Report          0      0
    Performance Counters Report          0      0
    Partitions Check                     0      0
    IPoIB Subnets Check                  0      1

Please see /var/cache/ibutils/ibdiagnet.log for complete log
----------------------------------------------------------------

-I- Done. Run time was 0 seconds.
```

Um den IPoB Subnets Check zu beheben erstellen / editieren wir folgende Datei

```
nano /etc/opensm/partitions.conf
```

Und fügen folgende Zeile ein :

```
 Default=0x7fff, ipoib, mtu=5, rate=7, defmember=full : ALL=full, ALL_SWITCHES=full,SELF=full;
```

Nun sieht die Ausgabe von ibdiagnet so aus

```
Loading IBDIAGNET from: /usr/lib/x86_64-linux-gnu/ibdiagnet1.5.7
-W- Topology file is not specified.
    Reports regarding cluster links will use direct routes.
Loading IBDM from: /usr/lib/x86_64-linux-gnu/ibdm1.5.7
-W- A few ports of local device are up.
    Since port-num was not specified (-p option), port 1 of device 1 will be
    used as the local port.
-I- Discovering ... 3 nodes (1 Switches & 2 CA-s) discovered.



-I---------------------------------------------------
-I- Bad Guids/LIDs Info
-I---------------------------------------------------
-I- No bad Guids were found

-I---------------------------------------------------
-I- Links With Logical State = INIT
-I---------------------------------------------------
-I- No bad Links (with logical state = INIT) were found
```

```
-I---------------------------------------------------
-I- General Device Info
-I---------------------------------------------------

-I---------------------------------------------------
-I- PM Counters Info
-I---------------------------------------------------
-I- No illegal PM counters values were found

-I---------------------------------------------------
-I- Fabric Partitions Report (see ibdiagnet.pkey for a full hosts list)
-I---------------------------------------------------
-I-    PKey:0x7fff Hosts:4 full:4 limited:0

-I---------------------------------------------------
-I- IPoIB Subnets Check
-I---------------------------------------------------
-I- Subnet: IPv4 PKey:0x7fff QKey:0x00000b1b MTU:4096Byte rate:40Gbps SL:0x00

-I---------------------------------------------------
-I- Bad Links Info
-I- No bad link were found
-I---------------------------------------------------
----------------------------------------------------------------
-I- Stages Status Report:
    STAGE                                Errors Warnings
    Bad GUIDs/LIDs Check                   0      0
    Link State Active Check                0      0
    General Devices Info Report            0      0
    Performance Counters Report            0      0
    Partitions Check                       0      0
    IPoIB Subnets Check                    0      0

Please see /var/cache/ibutils/ibdiagnet.log for complete log
----------------------------------------------------------------

-I- Done. Run time was 0 seconds.
```

Fertig.

Test der Geschwindigkeit

Mittels iperf testen wir die Kopiergeschwindigkeit im Netzwerk

Wir haben hier in unserem beispiel zwei IP-Adressen

172.30.128.75
172.30.128.76

wir brauchen dazu zwei SSH Terminalsitzungen.
Einmal auf Server A und auf Server B
in der einen starten wir den iperf Server (der mit der 75)

```
iperf -s
```

und in dem anderen (der mit 76) den client zum verbinden
Der Parameter -t geben an wie lange der Test laufen soll. Hier 20 Sekunden
Der Parameter -c gibt die IP Adressen des iperf Servers an

```
iperf -t 20 -c 172.30.128.75
------------------------------------------------------------
Client connecting to 172.30.128.75, TCP port 5001
TCP window size: 2.50 MByte (default)
------------------------------------------------------------
[  3] local 172.30.128.75 port 54800 connected with 172.30.128.75 port 5001
[ ID] Interval       Transfer     Bandwidth
[  3] 0.0000-20.0000 sec   125 GBytes  53.5 Gbits/sec
```

Version #12
Erstellt: 23 Mai 2023 08:36:45 von Admin
Zuletzt aktualisiert: 2 August 2023 10:03:17 von Admin