Home Infiniband

Why?

- Over the last 18 months I have upgraded all three of my main PCs to modern hardware (PCIe 4.0), including some very fast hard drives.
- Previously, fast hard drive transfer rates were around 80-90 Mbytes per second - less than the 1 Gbit/s of standard Ethernet, even including protocol overheads.
- Now, hard drives can be as fast as 230 Mbyte/s (over 1.8 Gbits/s).
- SATA SSDs are 6 Gbit/s.
- PCIe 4.0 NVMe SSDs are over 5000 Mbyte/s (42 Gbits/s) (and PCIe 5.0 ones are even faster).
- I frequently send large video files from my Windows PC to my MythTV box, between fast hard drives.

Upgraded Ethernet?

- Since I want to transfer files between fast hard drives, I wanted to have any upgraded Ethernet faster than the hard drives. The new next step for Ethernet is 2.5 Gbit/s, which is not fast enough. So even though all three of my new motherboards have 2.5 Gbit Ethernet ports, there is little point in using them as they are not fast enough.
- There is only minimal support for 5 Gbit/s, so the next step for Ethernet is really 10 Gbit/s, which is well supported. But it is quite expensive as it is an enterprise/commercial standard not normally used by home networks (yet).

Ethernet versus Infiniband

- Also used in enterprise/commercial systems is Infiniband. And surprisingly, if you look at the cost of Infiniband cards, they are frequently available at less than or similar cost to 10 Gbit/s Ethernet cards.
- The cheap Infiniband hardware, it turns out, is old Infiniband that is now too slow for the enterprise/commercial environment, and has been sold off second hand. New cards are also available which were never used as they were spares waiting to replace failed cards. This Infiniband level does 40 Gbit/s, or sometimes 56 Gbits/s. For the same price as 10 Gbit/s Ethernet, or cheaper.
- Enterprise Infiniband now does 100 Gbit/s or 200 Gbit/s. A 400 Gbit/s standard is being released later this year. Such cards are mega expensive.

Infiniband

- Infiniband was invented as a faster, better replacement for Ethernet.
- Infiniband is designed to have high bandwidth and also (very importantly) low latency.
- The hardware used is very similar to Ethernet – Infiniband cards can often be used to also do Ethernet (usually with different firmware).
- An Infiniband network is called a “fabric”. They can be very large with hundreds of devices on one subnet, and many subnets.

- The low latency allows Infiniband to be used to tie multiple computers together into a single supercomputer, with minimal delays passing messges between CPUs.
- Fabrics can be so large that one failing device causing network problems can disable all devices on that subnet, so multiple subnets with Infiniband routers between them are used to reduce this problem by having more subnets with fewer devices on them.
- Infiniband routers are also low latency.

RDMA

- RDMA = Remote Direct Memory Access
- Infiniband cards achieve low latency by using RDMA data transfers where the data is received or transmitted directly from a program's memory space.
- RDMA does not use the CPU – data is transferred directly to and from RAM using DMA hardware.
- Data does not need to be copied in and out of kernel buffers to transfer it between the device and a program. Consequently, most Infiniband transfers are "zero copy".

Infiniband Standards and Performance

- /home/stephen/Documents/presentations/Home Infiniband/InfiniBand - Wikipedia.html

QDR QSFP+

- SFP = Small Form Factor Pluggable – 1 Gbit/s
- SFP is used for optical fibre network connections, but for shorter cables, DAC can also be used
- DAC = Direct Access Copper
- SFP+ = 10 Gbit/s
- QSFP+ = Quad SFP+ = 4 x SFP+ = 4 x 10 Gbit/s = 40 Gbit/s
- SDR = Single Data Rate = 10 Gbit/s
- QDR = Quad Data Rate – four times the original SDR = 4 x 10 Gbit/s = 40 Gbit/s

Hardware

- 3 x Mellanox MCX354A-QBCT dual Infiniband 40 Gbit/s QDR QSFP+ PCIe 3.0x8 cards (TradeMe & eBay) $158.50 + $62.72
- Mellanox IS5022 Infiniband switch 40 Gbit/s QDR QSFP+ 8 ports (eBay) $276.98
- 3 x Replacement quiet fans for IS5022 switch (eBay) ($47.55) + bolts ($6.69)
- 5 metre DAC cable 40 Gbit/s QSFP+ (generic) (AliExpress) $43.59
- 1 metre Mellanox MC2207130-001 DAC cable 40 Gbit/s QSFP+ (eBay) $78.52
- 2 x 40 Gbit/s QSFP+ optical tranceivers QSFP-SR4-40G (fs.com) $152.06
- 15 metres OM4 optical fibre with MTP connectors (fs.com) $301.00
- Fibre cable installation (ConTel) $?
- Total: $1087.23

Photos

Which Drivers?

- Drivers for most Infiniband cards are builtin to the Linux kernel as modules.
- Drivers can also be downloaded from the Nvidia/Mellanox site with all the associated software and installed using their installer. Windows drivers are installed this way.
- As the cards I am using are end-of-life, I chose to use the builtin Linux drivers as they are being maintained by the Linux maintainers.
- For newer Infiniband cards the Nvidia drivers may be better as they will be getting full support and maintenance with fixes coming faster than the kernel drivers.

Installing Ubuntu Software

- Arch Linux has a good page with install instructions here:

  https://wiki.archlinux.org/title/InfiniBand

  There are some minor differences from Ubuntu, but mostly those instructions for Arch Linux will work the same.
- *apt install infiniband-diags rdma-core*
- Edit the */etc/rdma/modules/rdma.conf* and */etc/rdma/modules/infiniband.conf* files (and any other .conf files) to enable the modules you want to use.
- Load the IP over Infiniband module: *modprobe ib_ipoib*

Rename the Infiniband IP ports (1)

Find the hardware addresses of the Infiniband ports: *ip show link*

The results should show your Infiniband ports:

8: ibp14s0: <BROADCAST,MULTICAST> mtu 4092 qdisc noop state DOWN mode DEFAULT group default qlen 256

link/infiniband 80:00:02:08:fe:80:00:00:00:00:00:00:f4:52:14:03:00:68:3b:31 brd 00:ff:ff:ff:ff:12:40:1b:ff:ff:00:00:00:00:00:00:ff:ff:ff:ff

9: ibp14s0d1: <BROADCAST,MULTICAST> mtu 4092 qdisc noop state DOWN mode DEFAULT group default qlen 256

link/infiniband 80:00:02:09:fe:80:00:00:00:00:00:00:f4:52:14:03:00:68:3b:32 brd 00:ff:ff:ff:ff:12:40:1b:ff:ff:00:00:00:00:00:00:ff:ff:ff:ff

Rename the Infiniband IP ports (2)

The hardware address used in the udev rules is the last 8 bytes of the *link/infiniband* value. Use that to create udev rules such as these in your */etc/udev/rules.d/70-persistent-ipoib.rules* file:

ACTION=="add", SUBSYSTEM=="net", DRIVERS=="?", ATTR{type}=="32", ATTR{address}=="?*f4:52:14:03:00:68:3b:31", NAME="ib0"*

ACTION=="add", SUBSYSTEM=="net", DRIVERS=="?", ATTR{type}=="32", ATTR{address}=="?*f4:52:14:03:00:68:3b:32", NAME="ib1"*

- Update the Infiniband card firmware. For Nvidia/Mellanox cards, use mstflint (instructions on that Arch Linux page):

  apt install mstflint
- Configure your new Infiniband IP ports in the usual way (eg via GUI, in */etc/network/interfaces*, ...)
- Check your */etc/samba/smb.conf* file. If you are specifying the interfaces to bind to in an *interfaces* line, add the new names for your infiniband ports (eg *ib0,ib1*) or use their IP addresses
- Reboot

Subnet Manager

- Each Infiniband subnet can need a subnet manager to do address assignment and manage switches.
- Multiple subnet managers can be used, with the highest priority one controlling the subnet and failover to the next highest priority subnet manager if that is shut down or fails.
- Some Infiniband switches come with a builtin subnet manager.
- The usual subnet manager is OpenSM:

  apt install opensm
- Configure OpenSM in */etc/opensm/opensm.conf*

Windows Support

Windows support is available for Mellanox cards by downloading the matching software version from the Nvidia/Mellanox site. It comes with all the associated software usually installed on Linux and is configured in a similar manner.

Protocols

There are a lot of protocols that run over Infiniband, which can be used as well as the base Infiniband Verbs API. Some examples:

- IPoIB – Internet Protocol over IB. This allows full TCP/IP over an IB connection, but can not use RDMA due to how the TCP/IP stack works.
- SMB Direct – Using SMB v3 protocol, an SMB connection is initially made using IPoIB, and then the connection moves to direct IB using RDMA and runs without IPoIB overheads. This is supported by Windows as well as SAMBA.
- iSCSI – When IB/RDMA support is installed, iSCSI makes an initial connection using IPoIB and then moves to a direct IB connection using RDMA. This allows using a hard drive on another PC as though it was local, at full hard drive speed. Several hard drives can be used this way at the same time with a 40 Gbit/s connection.