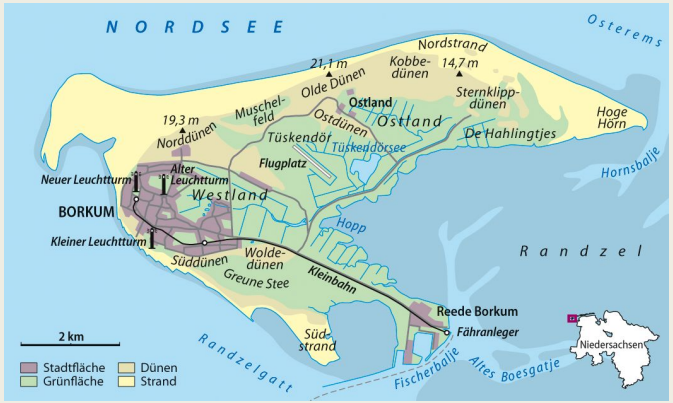
InfiniBand

an overview

/me

2022-now()	Cloud Gardener, HPC/AI Taiga Cloud
...	
2018-2022	Cloud Gardener, OpenStack, k8s, edge computing
...	
2012-2016	1&1, DNS Team, System Admin
pre-2012 Studium	Uni Paderborn, Freelancer

Abgeschlossenes Studium
Mathe/Informatik für Lehramt an Gymnasien



Agenda

- history
- what is infiniband?
  - *speeds, connectors and cables*
  - topologies
  - *OSI Layer*
  - *components*
- security and the cloud

history - why is infiniband?

1999	IB 1.0 Specs
	Future IO & NGIO -> IB 1.0 substitute PCI, Ethernet, FC Supporters Intel, Dell, Sun, MS, …
2001	Mellanox first HW 10 GBit/s
2002	Intel, MS drop interest -> PCIe
2005	Linux 2.6.11 support via OpenIB, OFED is born
2005	Storage Devices
2005	Cisco, acquired competitors, killing vendors

history - what is infiniband and why?

2009 Top 500 - 181/500 IB
259/500 Ethernet

2010 only 2 independent HW vendors left Mellanox, QLogic

2014 -

2016 IB was most used in supercomputers, later replaced by 10GBit/s Ethernet

2019 only 1 independent HW vendor NVidia

speed - what is infiniband?

	Year[20]	Line code		Signaling rate (Gbit/s)	Throughput (Gbit/s)[21]				Adapter latency (µs)[22]
					1x	4x	8x	12x	
SDR	2001, 2003	NRZ	8b/10b[23]	2.5	2	**8**	16	24	5
DDR	2005			5	4	**16**	32	48	2.5
QDR	2007			10	8	**32**	64	96	1.3
FDR10	2011		64b/66b	10.3125[24]	10	**40**	80	120	0.7
FDR	2011			14.0625[25][19]	13.64	**54.54**	109.08	163.64	0.7
EDR	2014[26]			25.78125	25	**100**	200	300	0.5
HDR	2018[26]	PAM4		53.125[27]	50	**200**	400	600	<0.6[28]
NDR	2022[26]		256b/257b[i]	106.25[29]	100	**400**	800	1200	?
XDR	2024[30]	[*to be determined*]	[*to be determined*]	200	200	**800**	1600	2400	[*to be determined*]
GDR	TBA			400	400	**1600**	3200	4800	

https://en.wikipedia.org/wiki/InfiniBand#Performance

why is infiniband?

What we have:

- HPC & AI - huge amounts of data
- memory complex computations
- memory limited computation units (eg A100/H100 80GB/GPU)

How to transfer?

Who does it?

- ~~CPU?~~
- ~~DMA?~~
- RDMA (CA, HCA)
- Latencies?

typical scenario:

- huge amount of data on disk
- “hot” data in RAM

what is infiniband? - but why?

- low latencies - guaranteed media access times
  - *“access token” per P2P links*
  - depending on the topology: guaranteed latency
- guaranteed delivery and ack
- RDMA
  remote direct memory access
- at the time of development: higher speeds than competitors

	Year[20]	Line code		Signaling rate (Gbit/s)	Throughput (Gbit/s)[21]				Adapter latency (µs)[22]
					1x	4x	8x	12x	
SDR	2001, 2003	NRZ	8b/10b[23]	2.5	2	**8**	16	24	5
DDR	2005			5	4	**16**	32	48	2.5
QDR	2007			10	8	**32**	64	96	1.3
FDR10	2011		64b/66b	10.3125[24]	10	**40**	80	120	0.7
FDR	2011			14.0625[25][19]	13.64	**54.54**	109.08	163.64	0.7
EDR	2014[26]			25.78125	25	**100**	200	300	0.5
HDR	2018[26]	PAM4		53.125[27]	50	**200**	400	600	<0.6[28]
NDR	2022[26]		256b/257b[i]	106.25[29]	100	**400**	800	1200	?
XDR	2024[30]	[*to be determined*]	[*to be determined*]	200	200	**800**	1600	2400	[*to be determined*]
GDR	TBA			400	400	**1600**	3200	4800	

https://en.wikipedia.org/wiki/InfiniBand#Performance

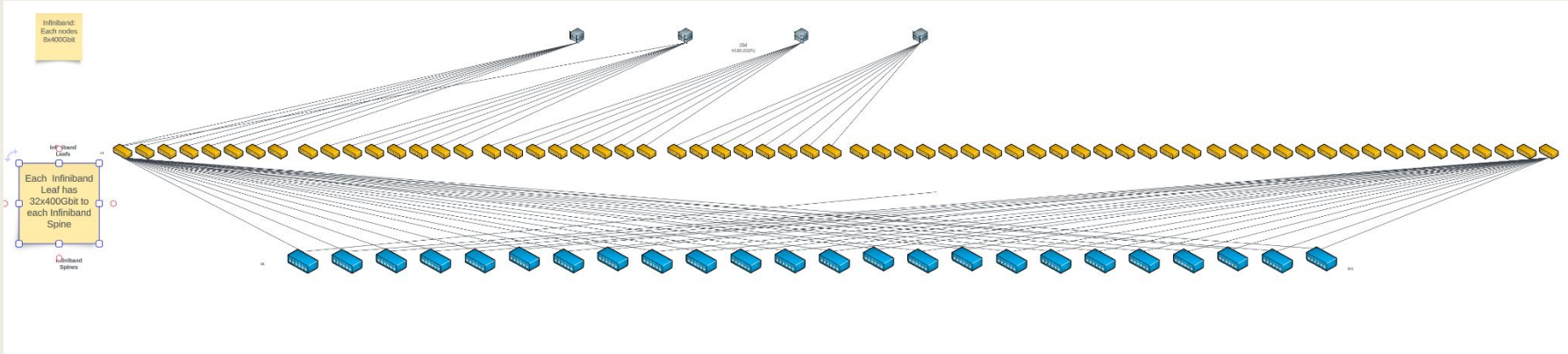
what is infiniband? - use case examples

- IB + GPU
  - *physics simulations*
  - *particle systems*
  - *weather modells*
  - *AI*
  - *logistics*

- only IB
  - *automated stock exchange trade - low latencies*

IB typical topologies

- P2P links - small l2 failure domains
- Clos network
- other:
  - *torus networks, butterfly networks*
- accumulation of path capacity
- For HPC/AI Non-Blocking xfer



L1 & connectors

QSFP

OSFP

Multimode

Singlemode

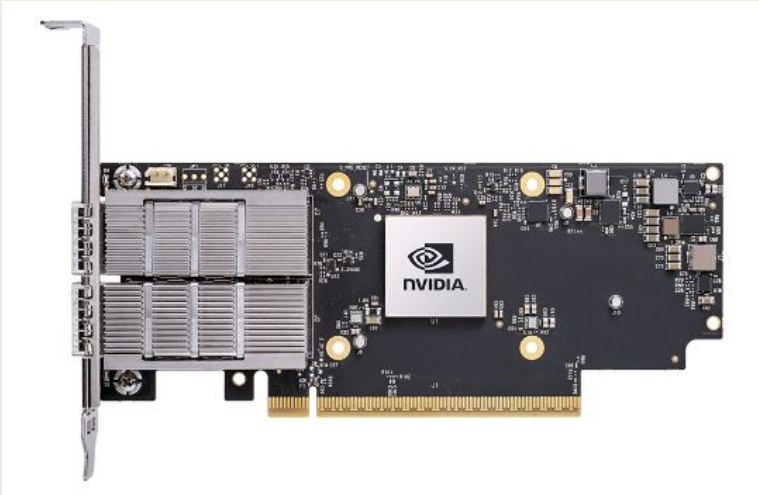
DAC - Cooper

https://docs.nvidia.com/networking/display/cablemanagfaq/cable+and+connector+definitions

L2

GUID

- HCA
- CA
- Switch Ports



NVIDIA.COM Connect X7 IB HCA

unconfigured L2 link, can only send datagrams

eg. MAD

L2 & L3 are closely entangled

what is infiniband? - L3 - Subnet Manager

Subnet Manager

- assigns LIDs
- configures ports HCAs, “switches”
- sets and programs routing into ports
- manages multicast groups
- manages partitions

what is infiniband? - L3 - Subnet Manager

Cold start/Big Sweep

- BFS/DFS, scan for connected ports

Standby/Small Sweep

- scans/listens for topology changes

L2 and L3 are closely entangled

Configuration done via MAD - Management Datagrams

what is infiniband? - L4 - QPairs

- Transport ~~Protocol~~ Engine in HW
  Queue Model (see Token based Media Access)

- Dest LID (implies Src LID)
- Modus (reliable vs unreliable, connection vs datagram)
- ~~Packet~~ Message sizes
- per peer DMA addresses, sizes

Usage?

typically use-case special libraries

not POSIX

Infiniband vs Ethernet

- Layer 2 - Infiniband
  - *GUID*
- Layer 3
  - *LID*
  - *routing - in HW*
  - *use case/topology optimized routing algo*
- Layer 4 - QPairs

- Layer 2 - Ethernet
  - *MAC - vendor assigned*
- Layer 3 - IP
  - *IP address - DHCP/static*
  - *routing - IP stack of OS*
  - *prefix routing*
- Layer 4 - TCP

IB terminology recap

- Devices
  - *HCA*
  - *switches*
  - *(routers)*
  - *Ports*
- Adresses
  - *GUIDs*
  - *LIDS*
- Subnet Manager
- Datagrams
  - *MADs (Management Datagrams)*
- QPairs

Subnet Manager

configuresDevices, Adresses, Routes per Device per Dest per Source

→Link-State DB

- Partitions
  - *32767 = 2^15 0000-7fff*
    - special 0000, 7fff - Management
  - *Membership MSB - full vs limited*
  - *Pkey-Table per port/CA*
- Keys - per subnet optionally per host, Subnetmanager configures this
  - *MKey*
  - *SMKey*
  - *SAKey*
  - *PKey*
  - *VSKey*

Subnet Manager

- detects devices - GUIDs → listens on plug events, link change events
  - *sweeps, full vs small sweeps*
- assigns LIDs (persistent)
- management datagrams
  - *MKey !*
  - *SMKey !*
  - *pushes routing/forwarding information*
  - *pushes Pkey tables*
- act as management hosts
  - *switch config - VSKey (TBD)*
  - *firmware updates - VSKey (TBD)*

Subnet Manager

configures Adresses, Routes per Device per Dest per Source

→ Everything!

- What if it dies?
- HA ? - active/passive priorities
  - *net split? I don't care, optimized for speed ;)*

What we have: Infiniband → L3 fabric

- P2P links - small l2 failure domains
  - *addresses assigned via SubnetManager*
- l3 routed
  - *bandwidth sum of all links*
    → scales with links and leaves connected
  - *limited by number of ports of the NIC*
  - *routing information distributed by MAD/Subnetmanager*
    - routing in ASIC (switches/routers/bluefields) - hardware
    - ~~routing in software~~

- no prefix routing!
- Use-Case Cloud: tenant network isolation
  → partitions

IP o IB - Subnet manager

- possible :) no IPv6
- per partition
- no ARP
  - → *Subnet Manager handles/simulates this*
    - IB Multicast groups
    - What if the subnet manager dies? - Broken

Software Stack

- run the network
  - *OFED / WinOFED*
  - *Subnet Manager: opensm vs UFM (NVidia)*

- compute
  - *Transfer Engine abstraction libs*
  - *scheduler*

    → MPI / openMPI

    (nearly) no one works with QPairs directly!

observations: Software Stack - HPC

MPI

- abstracts message passing
- … abstracts Unicast vs Multicast
- … abstracts transfer of data structures
- … uses transport library

→ Transport library (no POSIX)

- part of vendor driver bundle OFED
- abstracts QPairs and RDMA

observations: Software Stack - AI

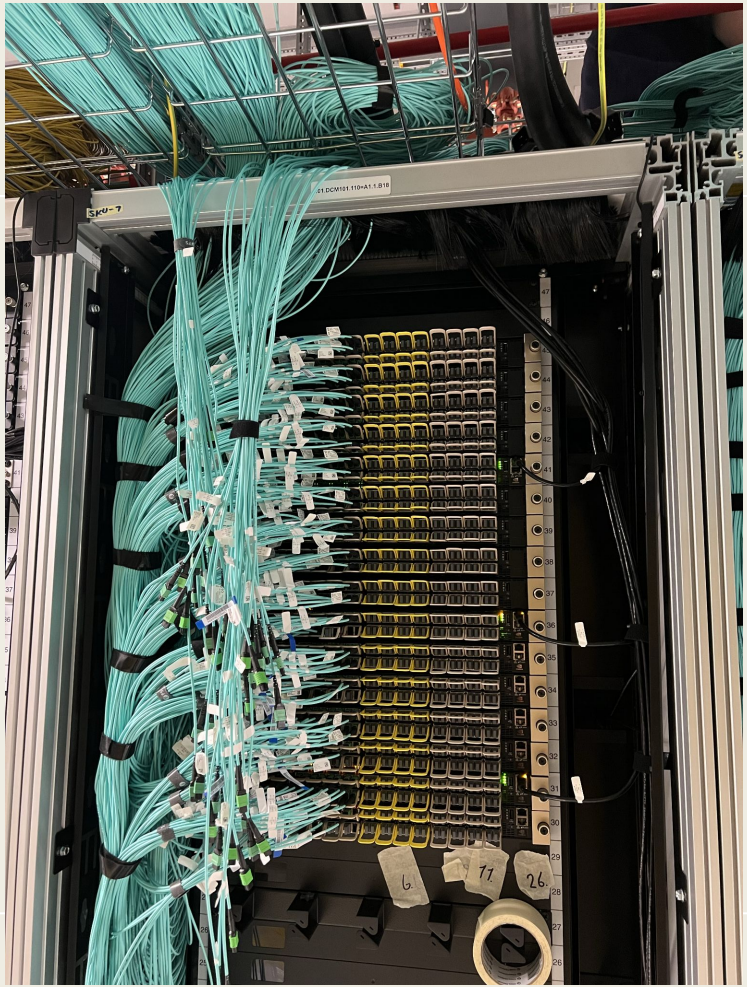
- on top of Docker or k8s
- typical additions layers
  - *PyTorch → NCCL → MPI*
- Slurm which uses
- MPI
- which uses IB via OFED libs

Deployments

- Host Side, 32x 400GB

Deployments

- Leaf - Spine, 32x 400GB



Use-Case Cloud-alike

- OpenStack - possible but …
  - *Nova - very specific PCI Passthrough*
  - *Ironic?*
    - BMC integration
  - *overhead hell, use-case a bit different*
    - lower churn
  - *no IB support*
- InfiniBand Tenant Network Isolation? →partitions

Learnings - security nightmares?

- Subnet Manager
  - *rogue subnet managers? → keys*
    - HA / priority
  - *rogue “network administrators” and port config*
    - if a SM is in place →
      - *MKey protect port config config changes*
      - *SMKey → address and routing changes*
      - *…*
      - *VSKey! and passive environments*
        - MKey leaks
        - Who and which CAs are allowed to send MADs?

Learnings - security nightmares?

- InfiniBand Tenant Network Isolation
  - *who controls the PKeyTable per port per CA*

- Switch & HCA firmwares, signed
  - *permissions handled by*
    - port config via subnet manager
    - firmware API
    - driver in kernel

  - *bug free?*
  - NVMEoF & RDMA security - https://arxiv.org/abs/2202.08080
  - *trust anchor?*

Learnings - Operations

- Subnet manager - extremely stable
  - *error messages are cryptic but often contain paths to erroneous/misbehaving devices*
    - eg … 0,1,13,44,6 …

Learnings - Operations

- Debug, Stats and Monitoring?
  - *nice toolchain*
  - *iblinkinfo - parseable output of complete network topology*
  - *ibhosts*
  - *ibswitches*
  - *ibtracert*
  - *ibping!*
  - *ib_send_bw/ib_read_bw*
  - *opensm - telemetry plugins*
  - *prom exporters*
    - https://github.com/treydock/infiniband_exporter
    - https://github.com/guilbaults/infiniband-exporter
    - https://gitlab.cern.ch/lhcb-online/cables-info-exporter#

Sources

- InfiniBand Network Architecture, Addison-Wesley 2002
- https://tin6150.github.io/psg/infiniband.html

Nice to know

- NVMEoF & RDMA security - https://arxiv.org/abs/2202.08080
- https://imbue.com/research/70b-infrastructure/

?