From Ethernet to InfiniBand

Sommerakademie in Leysin
AG 2 – Effizientes Rechnen

Philipp Czerner

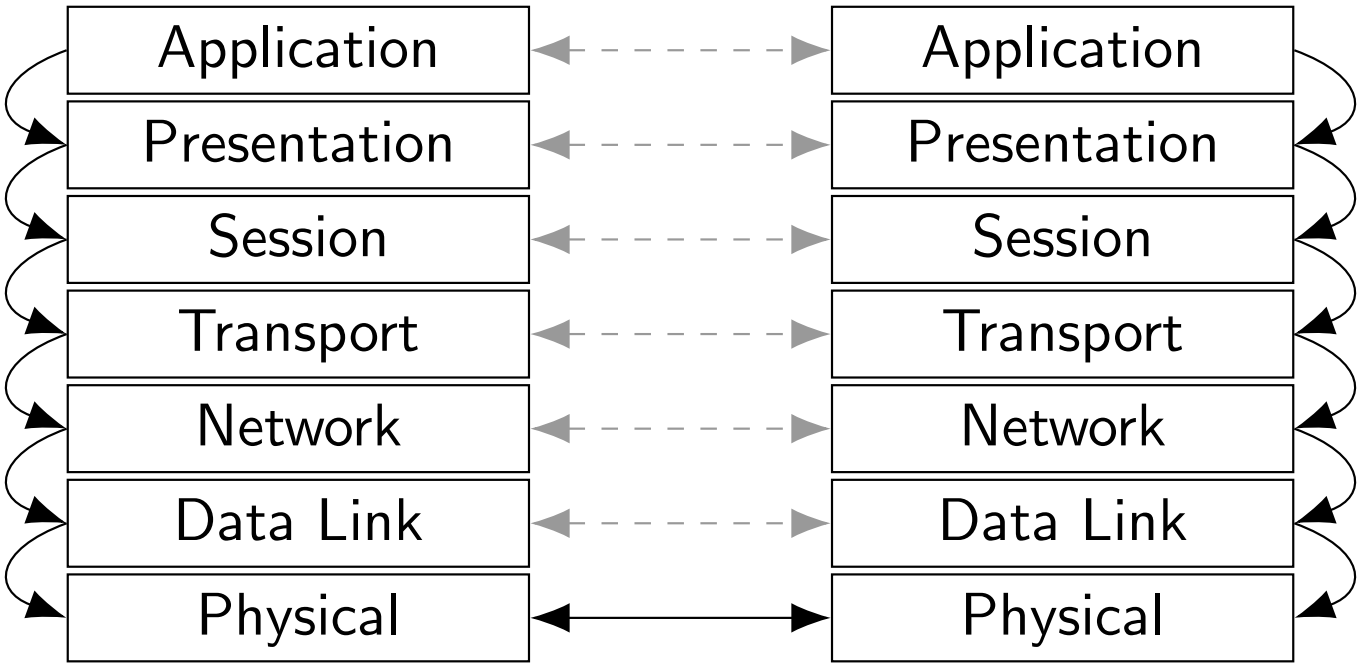
TU Clausthal

August 2016

Outline

OSI Model

- Open Systems Interconnection model
- Describes and standardises the workings of a communications system
- Partitions into layers 1-7
  - Each layer uses the functionionality of the layer below and provides an interface for the one above
  - Direct communication only happens on layer 1 (Physical layer)



OSI Model – Layers 1 to 4

- Layer 1 (Physical) deals with the specification of the physical medium (pin layout, timing, half/full duplex)
  - Examples: Ethernet physical layer, DSL, SONET/SDH
- Layer 2 (Data Link) allows to transfer data reliably between neighbouring nodes (flow control, error checking, connections)
  - Examples: Ethernet, PPP, ATM
- Layer 3 (Network) handles multi-node networking (address translation, routing, traffic control)
  - Examples: IPv4, IPv6
- Layer 4 (Transport) provides a reliable data transfer between nodes in the network (acknowledging, error control, sequence numbers)
  - Examples: TCP, UDP

Ethernet

- Ethernet is a family of standards, mostly for LANs
- 10BASE-T, 100BASE-TX and 1000BASE-T are commonly used in consumer hardware, using shielded twisted pair cabling
- Ethernet is located at both the physical layer and the data link layer
  - It does not provide all functions the OSI model specifies, for example there is no error recovery
- Recent standards have raised the possible data rate to about 100 Gbit/s

Ethernet – 100BASE-TX

- Most common form of LAN technology in consumer hardware
- Uses a category 5 (or better) twisted pair cable and an 8P8C plug (sometimes called RJ45)
  - Only two of the four pairs of a standard Cat-5 cable are used
- Bit rate of 100 Mbit/s

Ethernet – Gigabit and beyond

- Gigabit, 10G and 40G Ethernet continue to evolve the Ethernet family
- As of June 2016 44% of the Top500 computers use an Ethernet type interconnect (41% use Infiniband), making it the most used interconnect

InfiniBand

- InfiniBand (IB) is a networking technology with a focus an high bandwidth and low latency
- Like Ethernet there are multiple standards
- Used in High-Performance-Computing
- Resides at layers 1-4 of the OSI model
- Standard protocols (like TCP/IP) can be mapped
- Throughoutput of up to 97 Gbit/s for a 4X link
  - The earliest version (in 2001) already had 8 Gbit/s

RDMA

- Traditionally the OS copies application data into buffers prior to sending
- Remote Direct Memory Access (RDMA) enables a zero-copy transfer
- The HCA of the remote node directly accesses the application memory
  - Does not involve the remote CPU, OS or caches
  - No context switches on either end
- The application must register the memory ranges with the HCA beforehand via the Kernel
  - The registration returns a key to the memory needed to access it remotely, for security reasons

InfiniBand – Architecture

- A host channel adapter (HCA) connects a CPU to the IB network over a PCI Express interface
- The HCA provides functionality of layers 1-4 in hardware
- Provides both send/recieve and RDMA based communication
- Uses a queue based model, consisting of a *Queue Pair* (QP) with a send and a recieve queue
  - *Work Queue Requests* (WQR) are placed in either of these
  - WQR can specify a send/recieve or RDMA request
- Both reliable and unreliable communication is available

InfiniBand – Key Advantages

- Transmitting data bypasses the kernel in order to reduce latency
  - The kernel is involved in things like registering memory or initializing the HCA, which are not time sensitive
- Remote Direct Memory Access (RDMA) allows one-sided communication

InfiniBand – Power Consumption

- Implementing the network stack in hardware reduces CPU overhead
- RDMA also lowers CPU involvement
- 48% of the top 100 of the Green500 (June 2016) use Infiniband, 17% use Ethernet (vs. 41% and 44% for the Top500, respectively)
- On some HPC systems the interconnect consumes 15-20% of the total power
  - The percentage increases when the system is not at full load