From Ethernet to InfiniBand

Philipp Czerner

Studienstiftung des deutschen Volkes

Sommerakademie in Leysin, August 2016

Abstract

The interconnect is an important piece of any high-performance computing cluster. This paper presents the basics of networking technologies in general and of Ethernet and InfiniBand in particular. The advantages of the InfiniBand are discussed, with a focus on power-saving potential when used as interconnect.

Contents

1 Introduction

As processors continue to evolve, network technologies need to keep pace, prompting new developments to improve both latency and bandwidth. In recent times the power consumption has started to gain importance as well, as energy costs make up a significant margin of the total cost of a computing cluster. With CPUs becoming more efficient in recent years and computation increasingly parallised, interconnect technologies shape up as a significant factor of the consumed energy and a promising area for improvements.

In this paper I will give a short introduction into network technologies via the OSI model. Then both Ethernet and InfiniBand will be discussed, as the two most common interconnect technologies for high-performance computing (HPC). This includes the improvements made by InfiniBand over Ethernet, especially as related to power consumption.

2 OSI Model

The Open Systems Interconnection model describes and standardises the workings of a communications system. Its main concept is the *layer*, of which there are seven. Each layer uses the functionality of the layer below and provides an interface to the layer above.

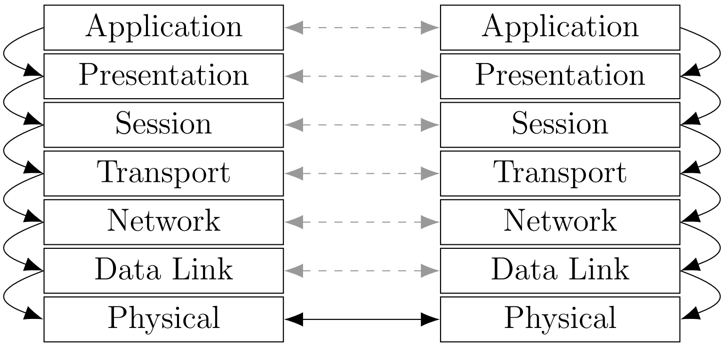

Figure 1: The layers of the OSI model

Direct communication (i.e. the exchange of bits via a physical medium) only happens on layer 1 (Physical Layer), the other layers conduct virtual communication using the facilities of the respective lower layer. The OSI model specifies the following layers:

- Layer 1 (Physical) describes and specifies the physical medium. This includes the type of medium (wireless, copper cable, fiberoptic cable), its properties (insulation, frequencies), plugs (dimensions, pin layout), and timing.

  Examples: Ethernet physical layer, DSL, SONET/SDH

- Layer 2 (Data Link) organises a reliable transfer of data between neighbouring nodes. This includes flow control (ensuring that a faster node does not flood a slower node with data), error checking (via checksum, this detects errors during transmission with high confidence), and collision handling (when using a shared medium).

  Examples: Ethernet, PPP, ATM

- Layer 3 (Network) handles multi-node networking. It provides address translation from physical addresses (MAC address, for example), that are unique and bound to the hardware, to logical addresses (IP address, for example), that may include routing information like subnets. There is also routing (choosing the path between two nodes in the network) and traffic control (prevent the nodes from overloading the network).

Examples: IPv4, IPv4

- Layer 4 (Transport) provides a reliable data transfer between nodes in the network. This includes acknowledging successful packets, checking data integrity using checksums and ensuring correct ordering of packets with sequence numbers.

  Examples: TCP, UDP

- Layer 5 (Session) controls the connection between the nodes. This includes initalization and termination of the connection and its mode of operation (full/half duplex or simplex).

  Examples: Sockets, RPC

- Layer 6 (Presentation) enables independence from data representation. It converts between different formats and encrypts and decrypts the data as needed.

  Examples: TLS, MIME

- Layer 7 (Application) directly communicates with the application process, which itself is outside of the OSI model. Its functionality is application dependant.

  Examples: HTTP, SSH

The internet is based on the TCP/IP stack and does not conform strictly to the specification of the OSI model. It can, however, be used as context to categorize and compare different protocol stacks. As all examples here work with the internet, they do not conform to the specification either and do not necessarily provide all features present in the layer.

3 Ethernet

Ethernet is a family of standards used mostly in local area networks (LANs). The standards 10BASE-T, 100BASE-TX and 1000BASE-T are commonly used in consumer hardware, with almost all motherboards directly providing a port. It is located at both the physical and the data link layer.

10BASE-T, 100BASE-TX and 1000BASE-T are based on shielded twisted pair cabling and have data rates of 10, 100 and 1000 Mbit/s, respectively, but there are other standards for different transmission mediums (fibreoptic cabling, for example) allowing for data rates as high as 100 Gbit/s.

Gigabit Ethernet, as well as 10G (10 Gbit/s) and 40G (40 Gbit/s) Ethernet, are also used as interconnect in HPC, making up 44% of the computer in the Top500 list (as of June 2016) [5]. This makes Ethernet the most used interconnect family.

Ethernet is commonly used together with the protocols IP (Internet Protocol) and TCP (Transmission Control Protocol). IP is located at layer 3 (network) of the OSI model, TCP at layer 4 (transport). As Ethernet does not provide error handling by itself, TCP ensures the correctness of data.

4 InfiniBand

InfiniBand (IB) is a networking technology with a focus on high bandwidth and low latency. Similar to Ethernet, it consists of multiple standards. IB is used in HPC as interconnect and was utilized by 41% of systems in the Top500.

The functionality provided by IB spans the first four layers of the OSI model, which makes protocols like TCP or IP unnecessary. They can, however, be mapped onto IB, which enables application to access IB via a more traditional API.

4.1 Architecture

IB's equivalent of a network interface controller (NIC) is called *host channel adapter* (HCA). The HCA is usually installed in the form of a PCI Express card, in contrast to Ethernet, which is often directly integrated into the motherboard.

The HCA provides the functionality of layers 1-4 in hardware. This includes all time-critical functionality (such as sending and receiving data), but does not extend to *all* of IB's features. Ethernet does not provide layer 3 and 4 functionality by itself, but relies on TCP and IP, which are software-based protocols.

IB provides both send/receive and RDMA based communication, both either reliable or unreliable. Send/receive is the traditional mode of network communication, featuring a stream of data that is sent from one end and received at the other. RDMA is described in detail in the next section. Reliable communication gurantees correctness of data, meaning that no data is lost, duplicated or changed during the transfer. Some applications may, however, value the reduced overhead more than error-free data (video streaming, for example), especially in the context of LANs, where errors are rare due to short distances and controlled conditions.

Work Queue Requests (WQR) specify some work for the HCA to do, this may be a send/receive or RDMA request. These are organised into two queues, the send and the receive queue, which form the *Queue Pair* (QP).

4.2 RDMA

Remote Direct Memory Access (RDMA) is a technique to efficiently transfer data over the network, with minimal latency and processing overhead.

Traditionally, an application sends data over the network by calling an appropiate function of the kernel. Then the data is copied into a buffer managed by the kernel and the NIC is notified to send the data over the network to the other node, where this process happens in reverse.

Using RDMA there is no middle step: The application notifies the HCA to write the data over the network, then the HCA reads the data directly and asynchronously off the application's memory and transfers it. The other node's HCA receives the data and writes it into the remote application's memory space.

There is no intermediate copy, making it a zero-copy transfer. Neither the local nor the remote CPU are involved in the process of transmitting the data. When reading data from a remote source, the remote CPU is not involved at all. Additionally the application does not call into the kernel, saving the CPU from doing expensive context switches.

Memory on modern machines is virtualised, meaning that each access from a user-level application goes through an indirection that maps it to the physical address. The kernel manages this mapping. An HCA does not have the same limitations as a user-level program, meaning that additional measures have to be taken to ensure memory safety. To initiate an RDMA request the application needs to register the memory regions first via the kernel, which checks whether the addresses are valid and translates them into physical addresses for the HCA. A key for this memory region is returned, which is needed to start any RDMA request involving this region (both locally and remotely).

4.3 Advantages

The kernel is not involved in any time sensitive operations; this eliminates the need for context switches and improves the latency. When using RDMA, there is minimal to none utilization of the CPU as well, enabling it to spend more time on the calculations as opposed to IO. This is aided by the hardware-based approach of IB, where the networking logic is offloaded onto custom and thus optimised circuits.

4.4 Power Consumption

The power consumption of an HPC cluster is a significant chunk of its operating costs, which in turn make up a large part of the total cost. Efforts to reduce power consumption have been made, but the interconnects have received little attention as of now.

On some HPC systems the interconnect consumes 15-20% of the total idle power, this percentage is expected to scale as the cluster become larger [3].

The main advantage of IB in terms of power consumption is its lower CPU overhead, achieved both by implementing the network stack in hardware and utilizing RDMA.

It is the interconnect of choice for many systems in the Green500 (as of June 2016), where it is used by 48% of the top 100 systems, with 17% using Ethernet [2]. (Comparing the full Green500 does not yield very interesting results, as it is mostly a permutation of the Top500.) For comparision, 41% and 44% of systems in the Top500 use IB and Ethernet, respectively.

5 Further Reading

[3] discusses different options for power saving in HPC interconnects. [4] gives a more in-depth overview of the workings of IB and its potential for improved performance in the context of database query processing. [1] describes techniques for improving the energy efficiency of datacenters via adaptations of the networking technology.

References

[1] Dennis Abts, Michael R Marty, Philip M Wells, Peter Klausler, and Hong Liu. Energy proportional datacenter networks. In *ACM SIGARCH Computer Architecture News*, volume 38(3), pages 338–347. ACM, 2010.

[2] Green500. `https://www.top500.org/green500/`. Accessed: 2016-10-05.

[3] Torsten Hoefler. Software and hardware techniques for power-efficient hpc networking. *Computing in Science & Engineering*, 12(6):30–37, 2010.

[4] Wolf Rödiger, Tobias Mühlbauer, Alfons Kemper, and Thomas Neumann. High-speed query processing over high-speed networks. *Proc. VLDB Endow.*, 9(4):228–239, December 2015.

[5] Top500. `https://www.top500.org/`. Accessed: 2016-10-05.