This section presents high-speed interconnects for cloud computing. A cloud, sometimes referred to as

a warehouse-scale computer (WSC), has an infrastructure consisting of a very large number of servers

interconnected by high-speed networks. This infrastructure is homogeneous in terms of the hardware

and software running on individual servers.

Although processor and memory technology have followedMoore’s law (i.e., that computer processors

double in complexity every two years), the interconnection networks have evolved at a slower pace

and have become a major factor in determining the overall performance and cost of the system. The

speed of the Ethernet has increased from 1 Gbps in 1997 to 100 Gbps in 2010; this increase is slightly

slower than Moore’s law for traffic [251], which would require 1 Tbps Ethernet by 2013.

InfiniBand is another interconnection network used by supercomputers as well as computer clouds.

InfiniBand has a switched fabric topology designed to be scalable. It supports several signaling rates,

and the energy consumption depends on the throughput. Links can be bonded together for additional

throughput; InfiniBand ’s architectural specification defines multiple operational data rates: single data

rate (SDR), double data rate (DDR), quad data rate (QDR), fourteen data rate (FDR), and enhanced data

rated (EDR). The signaling rates are: 2.5 Gbps in each direction per connection for an SDR connection;

5 Gbps for DDR; 10Gbps for QDR; 14.0625 Gbps for FDR; and 25.78125 Gbps per lane for EDR. SDR,

DDR, and QDR link encoding is 8 B/10 B, every 10 bits sent carry 8 bits of data. Thus single, double,

and quad data rates carry 2, 4, or 8 Gbps useful data, respectively. The effective data transmission rate

is four-fifths of the raw rate.

InfiniBand allows links to be configured for a specified speed and width; the reactivation time of

the link can vary from several nanoseconds to several microseconds. Exadata and Exalogic systems

from Oracle implement the InfiniBand QDR with 40 Gbps (32 Gbps effective) using Sun switches; the

InifiniBand fabric is used to connect compute nodes, compute nodes with storage servers, and Exadata

and Exalogic systems.

InfiniBand has high throughput and low latency and supports QoS guarantees and failover, the

capability to switch to a redundant or standby system. It offers point-to-point bidirectional serial links

intended for the connection of processors with high-speed peripherals, such as disks, as well as multicast

operations.

The networking infrastructure of a cloud must satisfy several requirements, including scalability,

cost, and performance. The network should allow low-latency, high-speed communication and, at the

same time, provide location transparent communication between servers; in other words, every server

should be able to communicate with every other server with similar speed and latency. This requirement

ensures that applications need not be location aware and, at the same time, it reduces the complexity

of the system management.

Important elements of the interconnection fabric are routers and switches. Routers are switches with

a very specific function: joining multiple networks, LANs, and WANs. They receive IP packets, look

inside each packet to identify the source and target IP addresses, then forward these packets as needed

to ensure that data reaches its final destination.

Typically, the networking infrastructure is organized hierarchically. The servers are packed into racks

and interconnected by a top-of-the-rack router; then rack routers are connected to cluster routers, which

in turn are interconnected by a local communication fabric. Finally, interdata center networks connect

multiple WSCs [197]. The switching fabric must have sufficient bidirectional bandwidth for cloud

computing. Clearly, in a hierarchical organization, true location transparency is not feasible and cost

considerations ultimately decide the actual organization and performance of the communication fabric.

The cost of routers and the number of cables interconnecting the routers are major components of

the overall cost of the interconnection network. We should note that the wire density has scaled up at

a slower rate than processor speed, and the wire delay has remained constant over time; thus, better performance and lower costs can only be achieved with innovative router architecture. This motivates

us to take a closer look at the actual design of routers.

The number of ports of a router distinguishes low-radix routers, with a small number of ports, from

high-radix routers, with a large number of ports. High-radix chips divide the bandwidth into a larger

number of narrow ports; low-radix chips divide the bandwidth into a smaller number of wide ports.

The number of intermediate routers in high-radix networks is greatly reduced, and such networks

enjoy a lower latency and reduced power consumption. As a result of the increase in the signaling

rate and in the number of signals, the pin bandwidth of the chips used for switching has increased by

approximately an order of magnitude every five years during the past two decades.

The topology of an interconnection network determines the network diameter5 and its bisection

bandwidth6, as well as the cost and power consumption [193]. First, we introduce informally the Clos

and the flattened butterfly topologies. The name butterfly comes from the pattern of inverted triangles

created by the interconnections, which look like butterfly wings. A butterfly network transfers the data

using the most efficient route, but it is blocking, so it cannot handle a conflict between two packets

attempting to reach the same port at the same time.

A Clos network is a multistage nonblocking network with an odd number of stages (see

Figure 7.11(a)). The network consists of two butterfly networks, and the last stage of the input is

fused with the first stage of the output. In a Clos network all packets overshoot their destination and

then hop back to it. Most of the time the overshoot is not necessary and increases the latency, meaning

that a packet takes twice as many hops as it really needs. In a folded Clos topology the input and output

networks share switch modules Figure 7.11(b). Such networks are sometimes called fat tree; many

commercial high-performance interconnects such as Myrinet, InfiniBand, and Quadrics implement a

fat-tree topology. Some folded Clos networks use low-radix routers (e.g., the Cray XD1 uses radix-24

routers). The latency and the cost of the network can be lowered using high-radix routers.

The Black Widow topology extends the folded Clos topology and has a lower cost and latency. It

adds side links, which permit a statical partitioning of the global bandwidth among peer subtrees [321].

The Black Widow topology is used in Cray computers.

The flattened butterfly topology [192] is similar to the generalized hypercube that was proposed in

the early 1980s, but the wiring complexity is reduced and this topology is able to exploit high-radix

routers. When constructing a flattened butterfly, we start with a conventional butterfly and combine the

switches in each row into a single, higher-radix one; each router is linked to more processors, and this

halves the number of router-to-router connections.

The latency is reduced because data from one processor can reach another processor with fewer

hops, though the physical path may be longer. For example, in Figure 7.12(b) we see a 2-ary 4-fly

butterfly. We combine the four switches S0, S1, S2, and S3 in the first row into a single switch S

0. The

flattened butterfly adaptively senses congestion and overshoots only when it needs to. On adversarial

traffic patterns, the flattened butterfly has similar performance to that of the folded Clos but provides

over an order of magnitude increase in performance compared to the conventional butterfly.

The authors of [193] argue that the cost of the networks in storage area networks (SANs) and computer

clusters can be reduced by a factor of two when high-radix routers (radix-64 or higher) and the flattened

butterfly topology are used. The flattened butterfly does not reduce the number of local cables, (e.g.,

backplane wires from the processors to routers), but it reduces the number of global cables. The cost of

the cables represents as much as 80% of the total network cost (e.g., for a 4 K system the cost savings

of the flattened butterfly exceed 50%).