For the edge layer, you can use the IBM Switch (8831-S52), which is a Mellanox sourced Ethernet switch. It provides 48x 1 GbE RJ45 + 4x 1/10 GbE SFP+ ports.

For more information about IBM and Mellanox joint solutions, see IBM and Mellanox Solutions.

4.6 InfiniBand network

InfiniBand is the predominant interconnect technology in the HPC market. InfiniBand has many characteristics that make it ideal for HPC:

- Low latency and high throughput
- Remote Direct Memory Access (RDMA)
- A flat Layer 2 that scales out to thousands of end points
- QoS
- Centralized management
- Multi-pathing
- Support for multiple topologies

The following section illustrates how to design an HPC cluster by using a Mellanox InfiniBand interconnect solution.

4.6.1 InfiniBand network topologies

There are several common topologies for an InfiniBand fabric. Here are some of those topologies:

- Fat tree

  A multi-root tree. This is the most popular topology.
- 2D/3D mesh

  Each node is connected to four or six other nodes: positive, negative, X axis, and Y axis.
- 2D/3D torus

  The X, Y, and Z ends of the 2D or 3D mashes are *wrapped around* and connected to the first node.

Other topologies also exist, such as dragonfly and hypercube. This section focuses on fat tree, which has optimized performance and scalability.

4.6.2 Fat tree topology

The most widely used topology in HPC clusters is the fat-tree topology. This topology typically enables the best performance at a large scale when configured as a *non-blocking* network. Where over-subscription of the network is tolerable, it is possible to configure the cluster in a blocking configuration.

A fat-tree cluster typically uses the same bandwidth for all links, and in most cases it uses the same number of ports in all of the switches.