

Myri-10G Modular Switches

10-Gigabit Ethernet and 10-Gigabit Myrinet



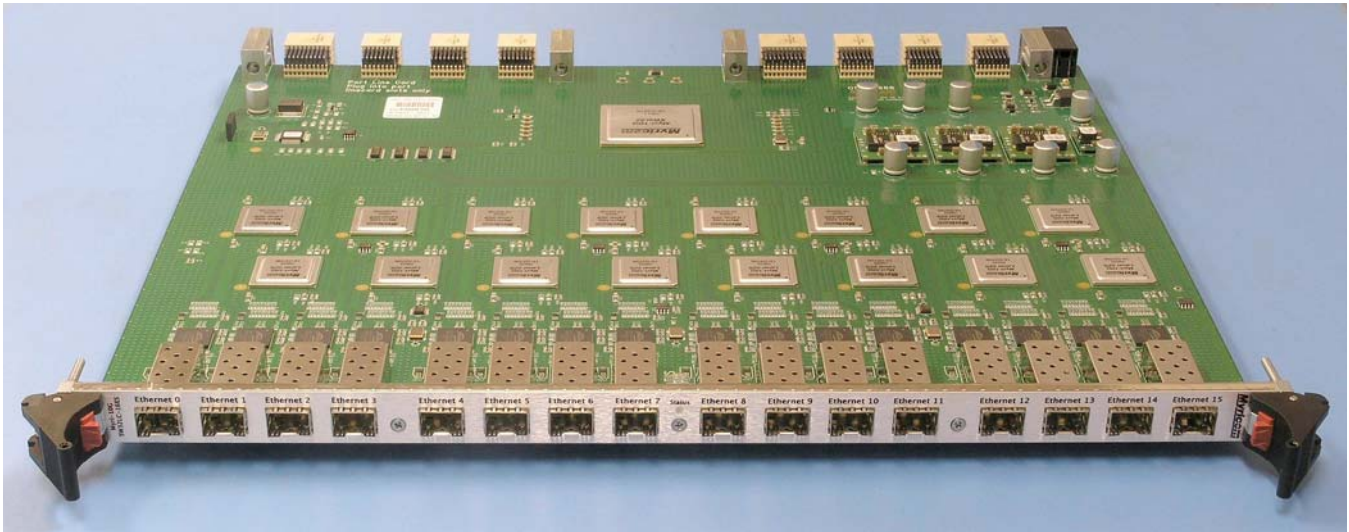
Myricom's Myri-10G switches are based on a 110ns-latency 32-port Myrinet-protocol crossbar-switch chip. The photos above show the current choice of enclosures for these high-port-density, modular switches.

Different mixes of Ethernet or Myrinet external ports, as well as ports with different PHYs, are configured by using different types of line cards. The line cards plug into a backplane that connects the switch chips on the line cards in a diameter-3 full-bisection Clos network. For applications requiring more than 512 host ports, the switch networks are scalable by interconnecting the internal Myrinet fabrics. Myrinet switching is famous for being scalable, and these switches are commonly used with only Myrinet external ports for HPC clusters.

Myricom

www.myri.com

10-Gigabit Ethernet external switch ports are provided by small network processors that perform layer-2 protocol conversion between the internal Myrinet switching fabric (Ethernet encapsulated in Myrinet) and fully interoperable 10-Gigabit Ethernet ports.



10G-SW32LC-16ES line card. There are 16 Myrinet-protocol (XAUI) ports to the backplane and 16 Ethernet-protocol SFP+ ports on the front panel (10GBase-SR or -LR with SFP+ transceivers, or Direct Attach). The chip at the center rear is the 32-port crossbar switch, and the 16 smaller chips are the network processors for Myrinet-Ethernet protocol conversion.

As an example of how Myri-10G networking can draw upon a spectrum of technologies and capabilities, let us look at the design of a very high performance file system with 10-Gigabit Ethernet clients. The highest performance file systems available today are cluster file systems such as Lustre, PVFS (Parallel Virtual File System), IBM's GPFS (Global Parallel File System), or others. Any of these cluster file systems can use TCP/IP over 10-Gigabit Ethernet as the cluster interconnect, but, for example, Lustre and PVFS operate even more efficiently directly over 10-Gigabit Myrinet using the kernel-bypass MX communication layer, which avoids the overhead of the Ethernet protocol stack in the hosts.

By using Myri-10G dual-protocol switches, Myri-10G NICs, and the MX communication layer, applications such as cluster file systems can enjoy the best of both worlds by using efficient Myrinet communication within the cluster and 10-Gigabit Ethernet communication with the clients. Myricom's first experience with this approach was for very high performance file systems for IBM Blue Gene/P supercomputers, which use 10-Gigabit Ethernet as their I/O fabric (see http://www.anl.gov/Media_Center/News/2007/ALCF071109.html), but the technique is more widely applicable to other situations, such as data centers with many 10-Gigabit Ethernet file-system clients.

Many of these modular, Myri-10G switches are supplied with Myrinet-protocol ports for HPC-cluster hosts, but some Ethernet-protocol ports for IP and storage connectivity. For switches with all 10-Gigabit Ethernet ports, these products satisfy a demand for "unmanaged," layer-2 switches that scale economically to large numbers of ports.

Myricom® and Myrinet® are registered trademarks of Myricom, Inc.