Guide to Myrinet-2000 Switches and Switch Networks

Myricom, Inc.
Revision: 27 August 2001

Safety, Installation, and Use Precautions

Provide power to the switch enclosures only with a 3-wire, IEC line cord from an outlet that includes a safety ground. The input power is 100–127 V~ or 200–240 V~, 50Hz or 60Hz. All switch enclosures are shipped with a power cord appropriate to the unit and to the shipping destination. The power cord should be rated for 10A except for 100–127V~ power cords for the M3-E128, which are rated for 15A.

An M3-E128 switch enclosure fully populated with line cards weighs approximately 31 kg (68 pounds). Two people are required to mount this unit in a rack. One person should not attempt to install an M3-E128 switch unit in a rack. The rack mounting holes conform to the EIA-310 standard for 19-inch racks.

Line cards and the fan tray may be "hot swapped" (inserted or removed with the power on). However,

- Insert and remove line cards gently, according to the instructions that start on page 4.

- Do not operate a switch for extended periods with missing line cards. Use M3-BLANK panels to fill in any empty line-card slots. The front of the enclosure needs to be closed both for the efficiency of the fan cooling and to avoid electromagnetic interference (EMI).

- Each line card includes thermal sensors. Operating the switch without a fan tray for more than one minute may result in line cards turning themselves off.

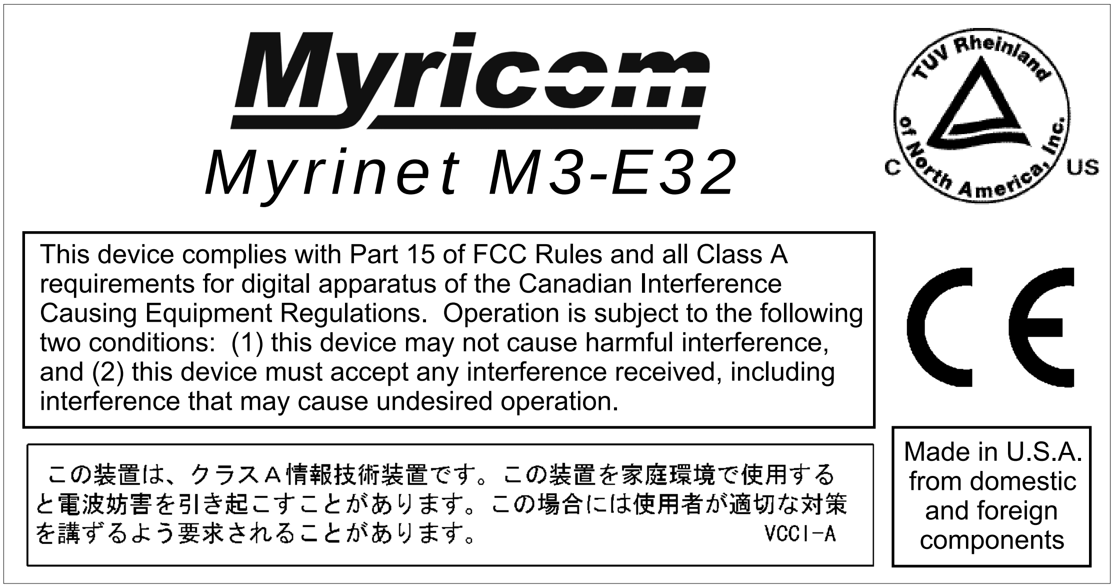
Configurations with optical-fiber Myrinet ports are **Class I Laser Products**. Optical-fiber components with this classification pose no threat of biological damage.

The most recent revision of this document can be downloaded from

http://www.myri.com/myrinet/m3switch/guide/ ,

and supplemental information is linked from this same web page.

Safety and EMC Certifications


Myricom
TUV Rheinland of North America, Inc.
C
US

Myrinet M3-E32

This device complies with Part 15 of FCC Rules and all Class A requirements for digital apparatus of the Canadian Interference Causing Equipment Regulations. Operation is subject to the following two conditions: (1) this device may not cause harmful interference, and (2) this device must accept any interference received, including interference that may cause undesired operation.

この装置は、クラスＡ情報技術装置です。この装置を家庭環境で使用すると電波妨害を引き起こすことがあります。この場合には使用者が適切な対策を講ずるよう要求されることがあります。 VCCI-A

Made in U.S.A.
from domestic
and foreign
components

A label similar to that above appears on the back of M3-E16, M3-E32, M3-E64, and M3-E128 switch enclosures shipped after the corresponding certifications were received.

Safety. Myricom is proud to employ TUV Rheinland of North America, Inc., to perform the independent safety certifications for these Myrinet-2000 switch products. The M3-E16, M3-E32, M3-E64, and M3-E128 were certified in accordance with the IEC System for Conformity Testing and Certification of Electrical Equipment (IECEE), CB Scheme. In addition, TUV certified these products to the US and Canadian safety standards.

Electromagnetic Compatibility (EMC). Subject to the limitations listed below, Myrinet-2000 switch products based on the M3-E16, M3-E32, M3-E64, and M3-E128 are within Class A limits for emission of and susceptibility to electromagnetic interference (EMI).

- An M3-BLANK filler panel must be used in any line-card slot without a line card.
- The M3-E16 may use any combination of M3-M, M3-SW16-8F, M3-SPINE-8F, M3-SW16-8S, M3-SPINE-8S, or M3-SW12-4L line cards.
- The M3-E32, M3-E64, and M3-E128 may use any combination of M3-M, M3-SW16-8F, or M3-SPINE-8F line cards.

EMI certification of additional configurations of line cards is in progress as of the date of this revision. As additional EMI certifications are completed, they will be added to this document.

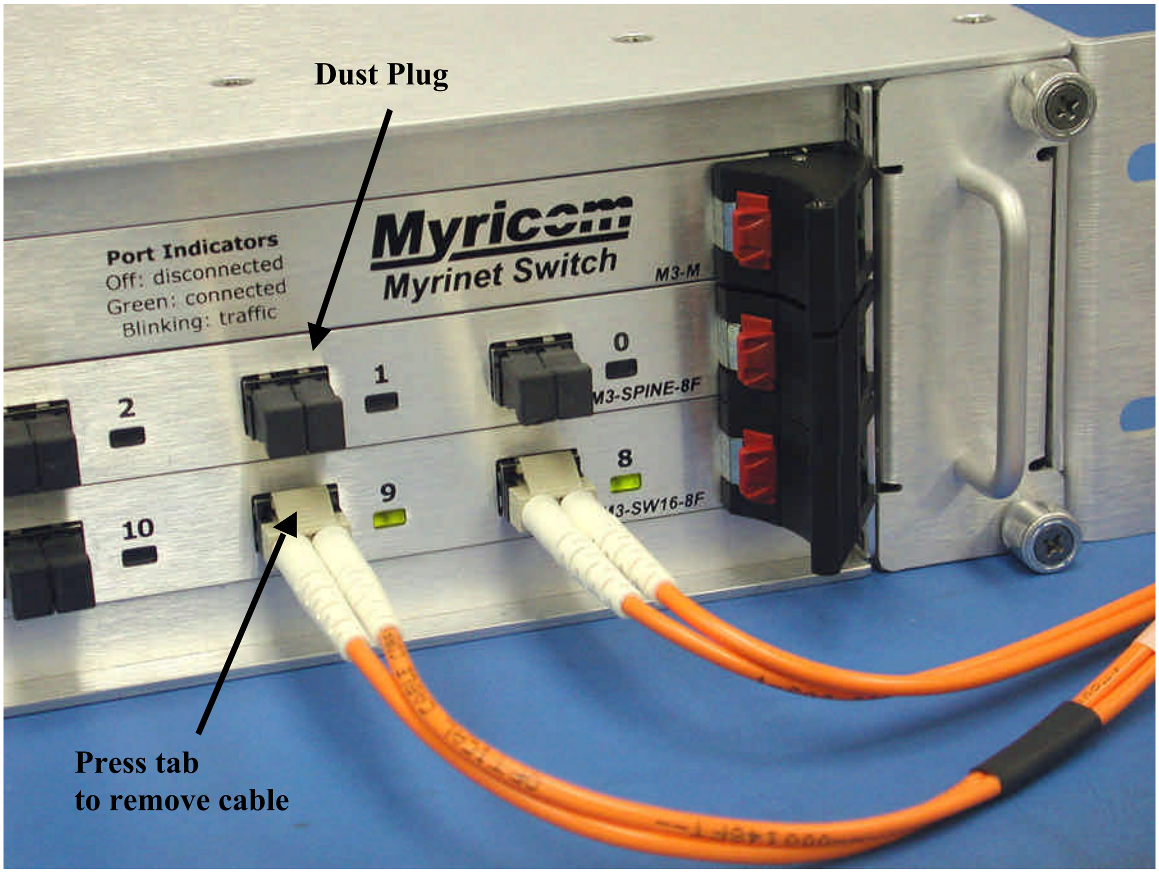
Characteristics and Installation of Myrinet Cables

Although all Myrinet links carry packets in the same format at the Data Link level, Myrinet has five different Physical level (PHY) implementations over four different types of cables. It is important to understand the characteristics of the types of cables used at your site, and how to install them. *All Myrinet cables may be plugged or unplugged with the component power on.*

Fiber is the preferred Myrinet PHY for several reasons:

- The fiber cables are small, light, flexible, and easy to install.
- The fiber cables and connectors exhibit exceptionally high reliability.
- Myrinet-Fiber cables may be up to 200m in length.
- Myrinet-Fiber components and cables operate within Class A limits for the emission of electromagnetic interference (EMI).

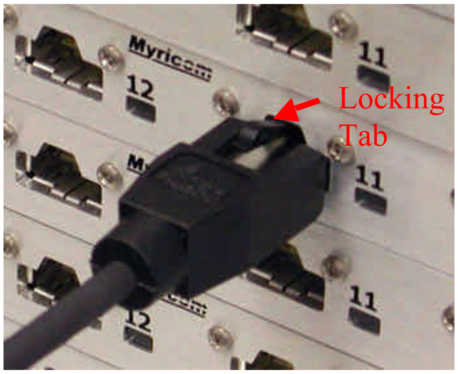
Myrinet-Fiber links carry packet data at a 2+2 Gb/s data rate on industry-standard 50/125 multimode-fiber-pair cables with "LC" connectorization. Myricom ships Fiber components with a dust plug in each port. It is recommended that the dust plug be left in place in unused switch ports. The fiber cable is shipped with dust caps on the ends of the fibers. To install the fiber cable, just remove the dust plug and dust caps, and insert the fiber-end connector.

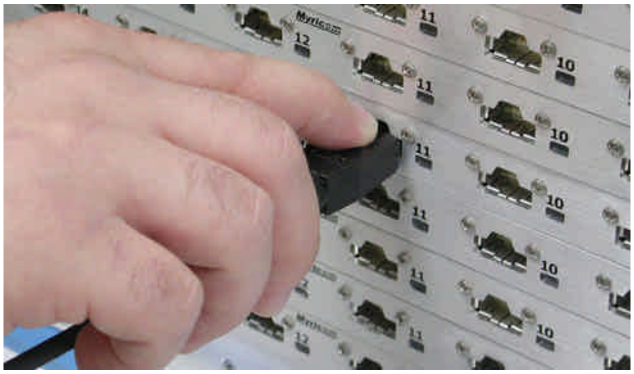


Press the locking tab on the LC cable-end connector to remove the fiber cable.

Serial links carry Myrinet packet data at a 2+2 Gb/s data rate on industry-standard HSSDC (High-Speed Serial Data Connector) cables up to 10m in length. These shielded cables are certified in several small configurations to Class A limits for the emission of EMI[1]. Myricom supplies HSSDC cables from manufacturers whose cables we have qualified through extensive testing. Not all HSSDC cables will function properly as Myrinet Serial cables.

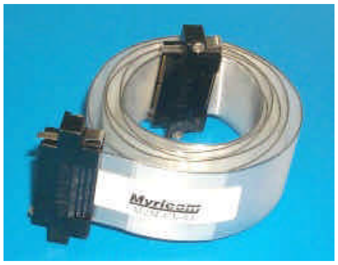
The HSSDC cable-end connector snaps into a Serial port. Please observe the locking tab on the cable-end connector:





In order to remove a cable, it is very important to depress the locking tab. Otherwise, the port connector will be damaged.

SAN (System Area Network) links are carried on unshielded "microstrip" ribbon cables up to 3m in length. In fact, a single cable carries two links, designated as the "A" and "B" links, and the same cables are used for two different Myrinet PHYs, SAN-1280 (1.28+1.28 Gb/s) and SAN-2000 (2.0+2.0 Gb/s). As you can see from the photo to the left, SAN cable-end connectors include a locking mechanism. These locking springs should both be depressed when inserting and when removing a SAN cable-end connector.



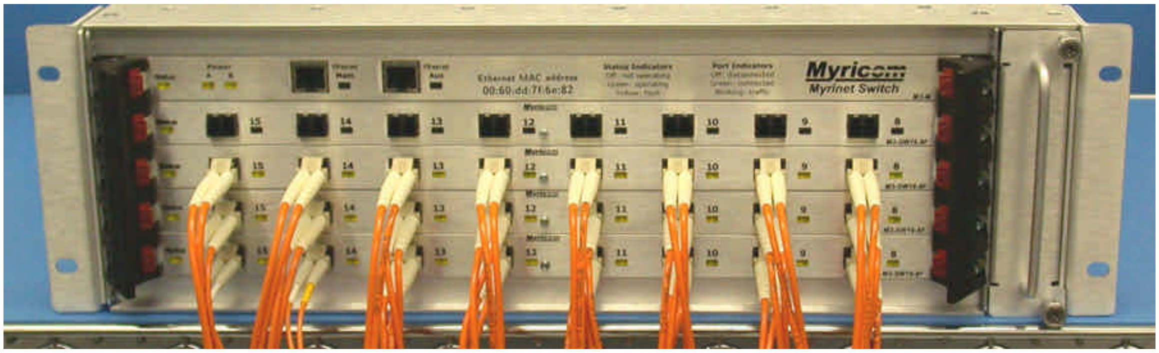
SAN links are for "in-cabinet" applications only. Switch configurations with any SAN line cards (M3-SW16-8M or M3-SW16-4DM) exceed Class A limits for EMI, and should be used only within a shielding enclosure. Also, it is best to restrict the use of SAN cables to within an enclosure because they will withstand only moderate physical abuse.

LAN (Local Area Network) links are a legacy PHY of Myrinet-1280. These multi-conductor, shielded cables are rugged, and are terminated with DB-37 connectors (similar to SCSI DB-25 connectors, but larger). LAN cables and components are certified for EMI to Class A limits. The only installation precaution is to be sure that the locking screws are screwed down **tightly**.

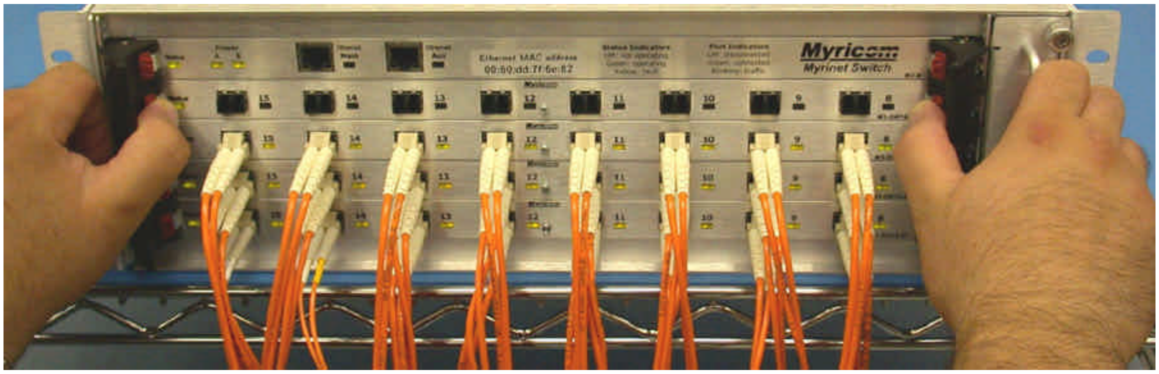
[1] Tests of large configurations of Myrinet switches show that products such as the M3-E128 with all Serial ports show very small margins relative to Class A limits in the 3–6 GHz range. These tests show that this EMI is due to characteristics of the HSSDC connectors. Myricom is continuing engineering efforts to reduce the EMI emissions, including using a different type of HSSDC connector in the most recent production products.

Hot swapping line cards and the fan tray

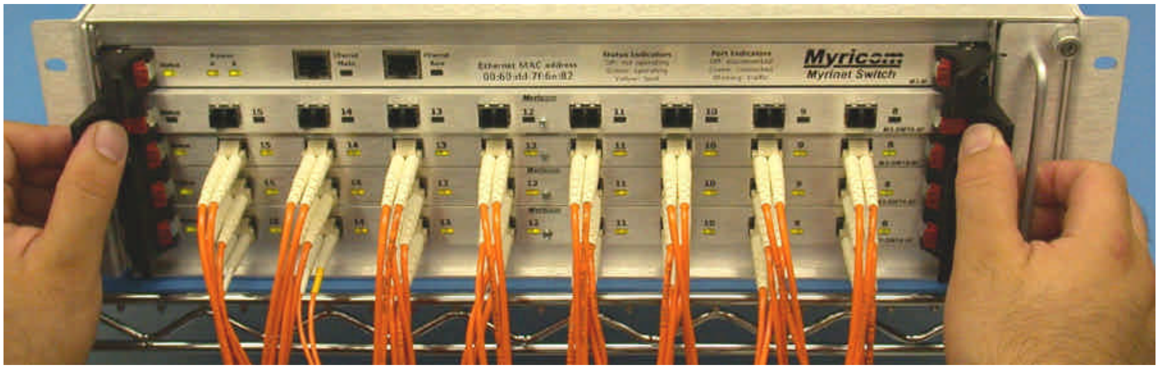
Each Myrinet-2000 switch line card has a pair of "handles" that lock the card in place, and that serve as a lever for inserting or removing the line card. Let's start with a line card that is inserted in a card slot of an enclosure. The handles should be snapped toward the middle, like this:



You'll notice from the illuminated LEDs that this switch is powered. That's OK, and why it's called "**hot** swapping." To remove the line card, first press the red tabs to unlock its handles.

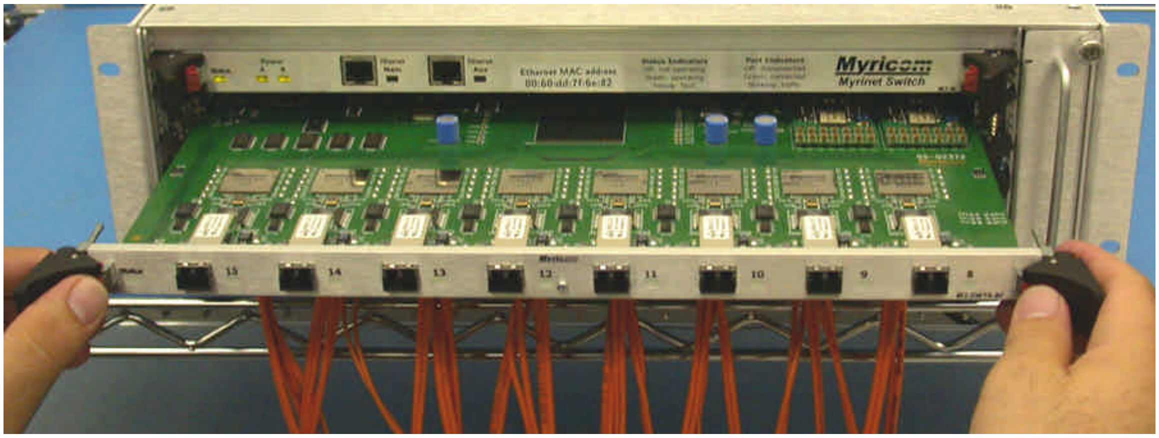


Then, turn both handles outward together. You should be able to see and feel the lever action as the front panel of the line card is ejected toward you.



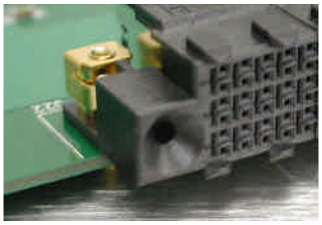
It is normal that the ejection is a little stiff while the line card's front panel is in contact with the front panel or bezels above and below. The bottom of each front panel (except for types with

SAN ports) and the bottom of the top bezel have spring gasketing as an EMI seal. Once the handles reach their outward extreme, the line card should slide easily out of the card cage.



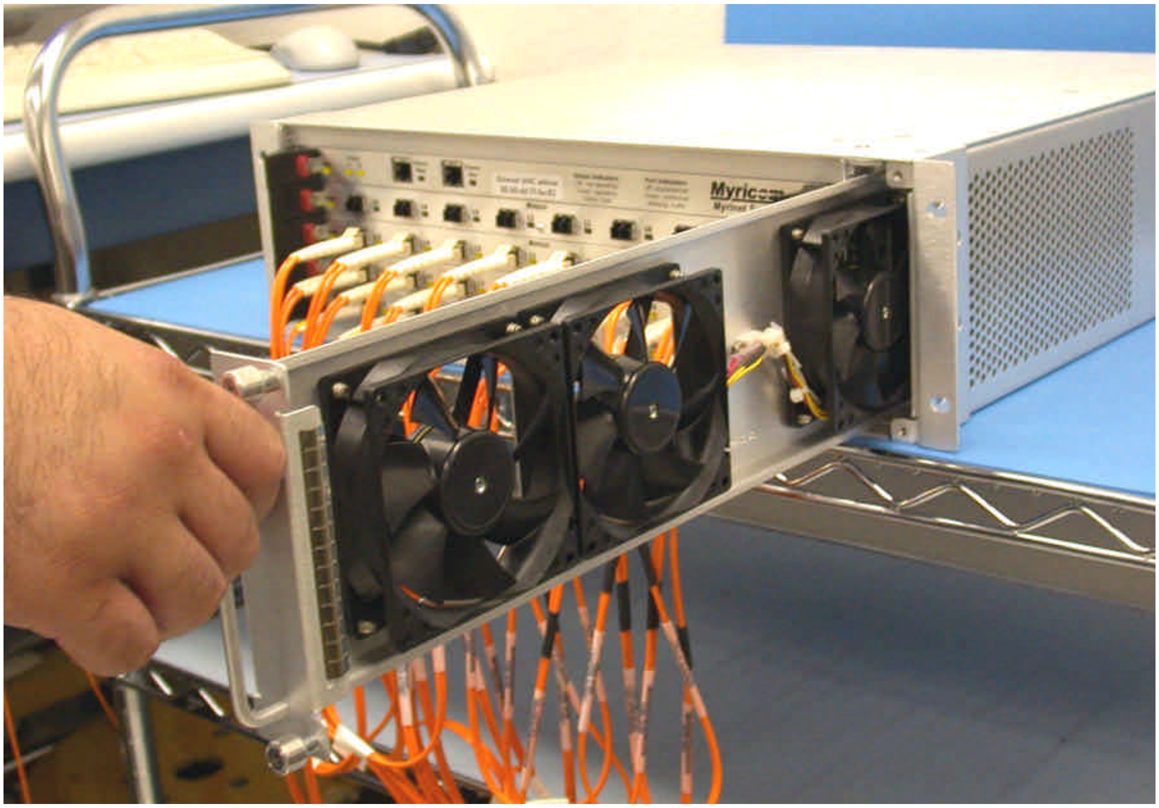
The backplane power connector in the center of the line card protrudes downward below other components on the bottom of the circuit board. Hence, just before the card is completely free, you may need to lift the card slightly so that the power connector will clear the top of the front panel or bezel below the line card you are removing.

Hot insertion of a line card just reverses the steps above. Be sure to start with the handles in their out position, and guide the line card carefully into the card guides. The signal pins and ground blades on the high-density connectors on the line card align themselves by means of alignment/grounding posts on the backplane. Conical depressions on the line-card connectors center themselves on the posts to align the connectors with great precision. Nevertheless, it is best to insert the line cards slowly and gently. When you use the lever handles to complete the insertion of the line card, you'll hear the locks snap into place, and see the red tabs pop out.



When the connectors start to seat, you may see the Status LED momentarily showing yellow before it becomes green. The yellow is a fault indication, and is due to some of the pins being not yet connected. As the line card becomes fully inserted, the Status LED should show green within ~0.5s. The green Status LED indicates that the line card has passed all of its self tests, and is operating.

The fan tray can be removed by loosening the two locking screws and pulling the fan tray out with the handle.



When a fan try is removed, it should be replaced within approximately one minute, or else line cards may power themselves off in response to an over-temperature condition.

Principles of Operation

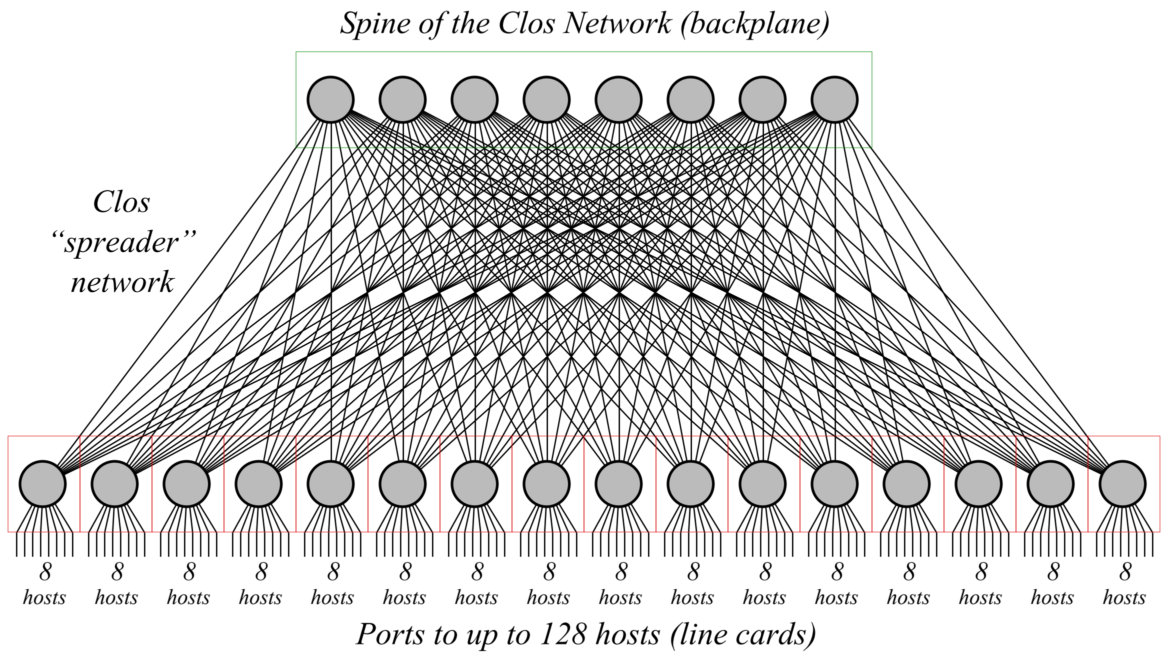
Guide to Myrinet-2000 Switches and Switch Networks

Principles of Operation

1. Introduction

The basic building block of this family of switch products is a 16-port Myrinet crossbar switch, which is implemented on a single chip designated as the XBar16. The XBar16 is pictured in the block diagrams below as a circle, and the Myrinet links as lines. A *Myrinet link* is a full-duplex pair of Myrinet channels.

The structure of these highly modular switches is best understood by starting with the maximal configuration within a single enclosure, a 128-host Clos network, which includes 24 XBar16s:

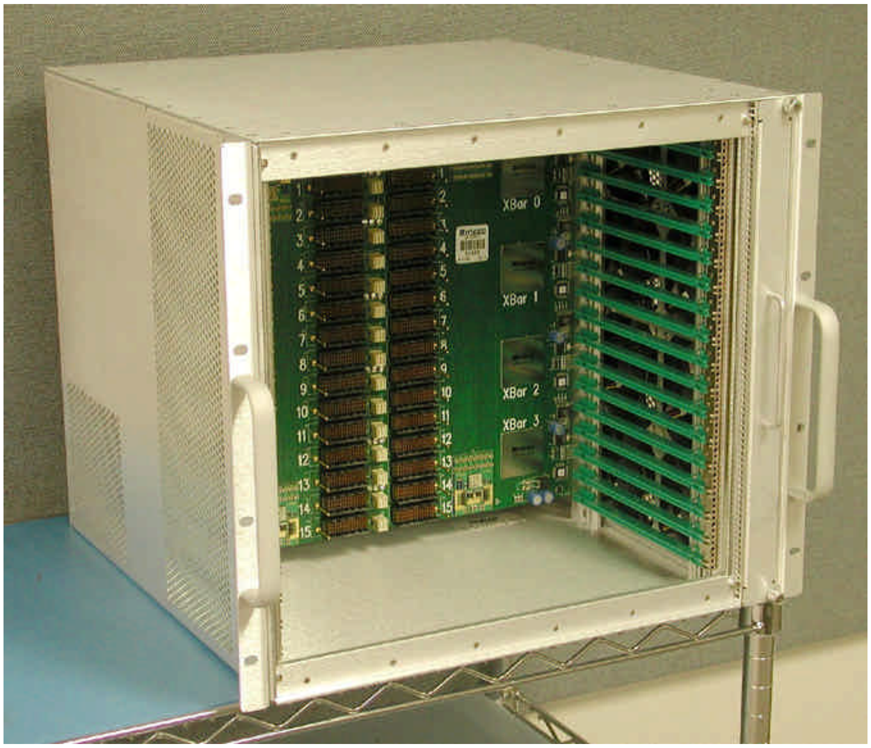


The network pictured above provides routes from any host to any other host.

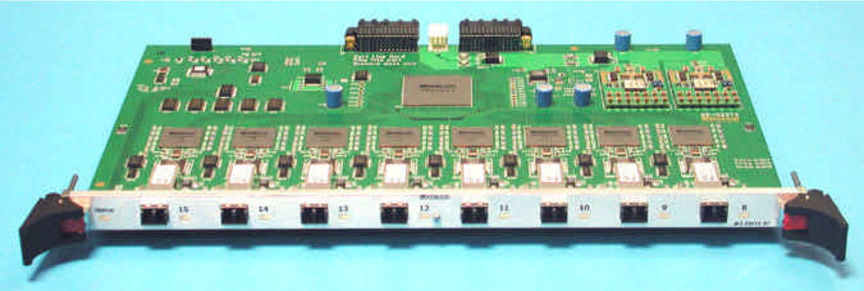
- There is a unique shortest route between hosts connected to the same XBar16.
- The eight minimal routes between hosts connected to different XBar16s traverse three XBar16 switches.

This topology provides so many paths between hosts that the *minimum bisection* of this network – its traffic-handling capacity – is as large as possible. This *full-bisection* and other properties of the Clos network topology will be discussed in greater detail starting in section 7.

The upper row of 8 XBar16s is the Clos network *spine*, which is packaged as an active backplane (green box) built into an enclosure that also provides power and cooling. The spine is connected through a Clos *spreader network* inside of the backplane to the lower row of 16 *leaf* XBar16s, which are packaged on *port line cards* (red boxes). These port line cards have up to 8 Myrinet ports through their front panel, and 8 Myrinet-SAN ports at connectors to the backplane. Myricom supplies different port line cards in which the front-panel ports are Myrinet Fiber, Serial, SAN (switchable between SAN-2000 and SAN-1280), or (legacy) LAN.



An M3-E128 enclosure with no line cards, allowing a view of part of the backplane



*A port line card with connectors to the backplane on the back,
and connectors to external links on the front panel*

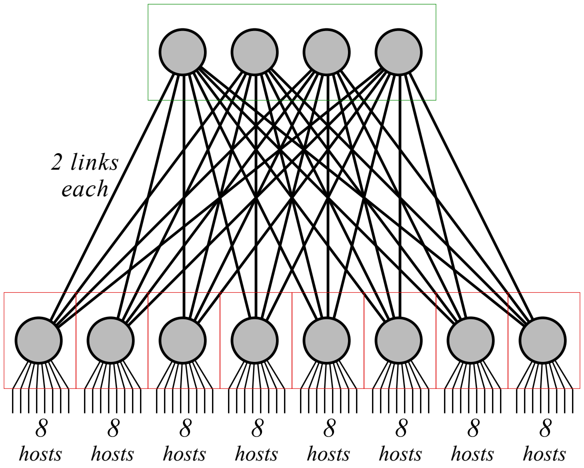
The topology is a full-bisection Clos network with any combination of port line cards inserted or omitted. For example, you could plug 10 port line cards into the M3-E128 enclosure to support up to 80 hosts. Ten of the ports on each of the 8 spine XBar16s would be used, and the other 6 ports on each XBar16 would be unused. The 8 10-port spine switches would have exactly the capacity required to carry traffic between 80 hosts. Even with only two line cards, the network is a full-bisection Clos topology. The spine would be 8 2-port switches, which can carry all of the traffic between the two line cards.

2. The Family of Enclosures

Myricom offers a series of enclosures, so that you can choose the one that best fits your needs.

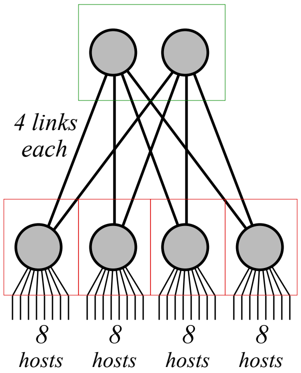
Enclosure product code	*M3-E16*	*M3-E32*	*M3-E64*	*M3-E128*
Maximum # of port line cards	2	4	8	16
Maximum # of front-panel ports	16	32	64	128
Physical height (1U = 1.75 in)	2U	3U	5U	9U

As you may note from the last paragraph of the preceding section, a Clos network for 64 or fewer hosts (8 or fewer leaf switches) would require only 8-port switches for the spine. An XBar16 can serve the purpose of two 8-port switches (with some additional communication provided "free"). Thus, a suitable topology for a 64-host Clos network of 16-port switches is:

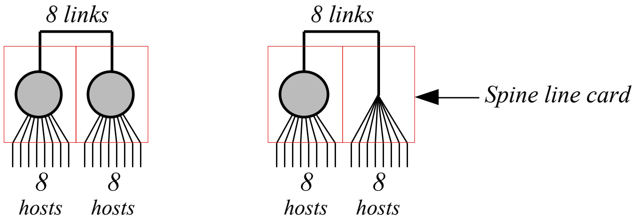


in which each of the thicker lines represents two links. (If we wanted to get "theoretical," we'd point out that the paired links connect to each of the two hypothetical 8-port switches within each 16-port spine switch.) This is the topology of the backplane and port line cards of the M3-E64.

Similarly, here is the topology of the backplane and port line cards of the M3-E32:



The M3-E16 enclosure, which has 2 port-line-card slots, is a special case. The direct extrapolation of the networks above would employ one XBar16 on the spine, but it is not needed. Instead, the backplane of the M3-E16 enclosure simply connects like ports of the two line cards:



With regular port line cards with switches, even the form on the left has an unnecessary level of switching. However, the Fiber and Serial versions of the Myrinet-2000 switch line cards are available in an alternative “spine” version that connects the 8 front panel ports through Physical-level conversion circuitry directly to the backplane spine, as illustrated on the right. As we shall see later, the spine line cards have other applications.

3. Other Features of the Enclosures and Line Cards

The Myrinet-2000 enclosures and line cards are designed so that the line cards or the fan tray can be “hot-swapped,” *i.e.,* inserted or removed from an enclosure that is powered. See the hot-swap instructions starting on page 5. There is, in fact, no power switch on the enclosure, lest a switch network accidentally be powered off.

Other features of the enclosure products include:

- a slot for an optional line card for monitoring and control of the switch. This monitoring line card (product code M3-M) is the same size as the port line cards, but has different backplane connectors, such that it can be inserted only into the top line-card slot. The M3-M includes a “big” microcontroller that communicates externally via either of two Ethernet ports, and internally via serial communication links. These serial communication links connect through the backplane with a “small” microcontroller (µC) on each line card, with a µC associated with each XBar16 on the backplane, and with a µC that monitors the rotation rate of the fans. The switch network operates without intervention of the monitoring line card, so that the monitoring line card is optional and hot-swappable.

- either a single, built-in power supply, or dual-redundant hot-pluggable power supplies[2]. These power supplies produce +12V from an IEC input of 100-127V~ or 200-240V~, either 50Hz or 60Hz. Dual +12V power-distribution networks convey the +12V power to the backplane, to each line card, and to the fan tray. At each of these locations, the two sources of +12V are combined with diodes, and the voltages required by the circuitry on the line cards and backplanes (typically +3.3V, +2.5V, and/or +1.25V) are generated

[2] Expected to be available for the M3-E128 in 4Q01.

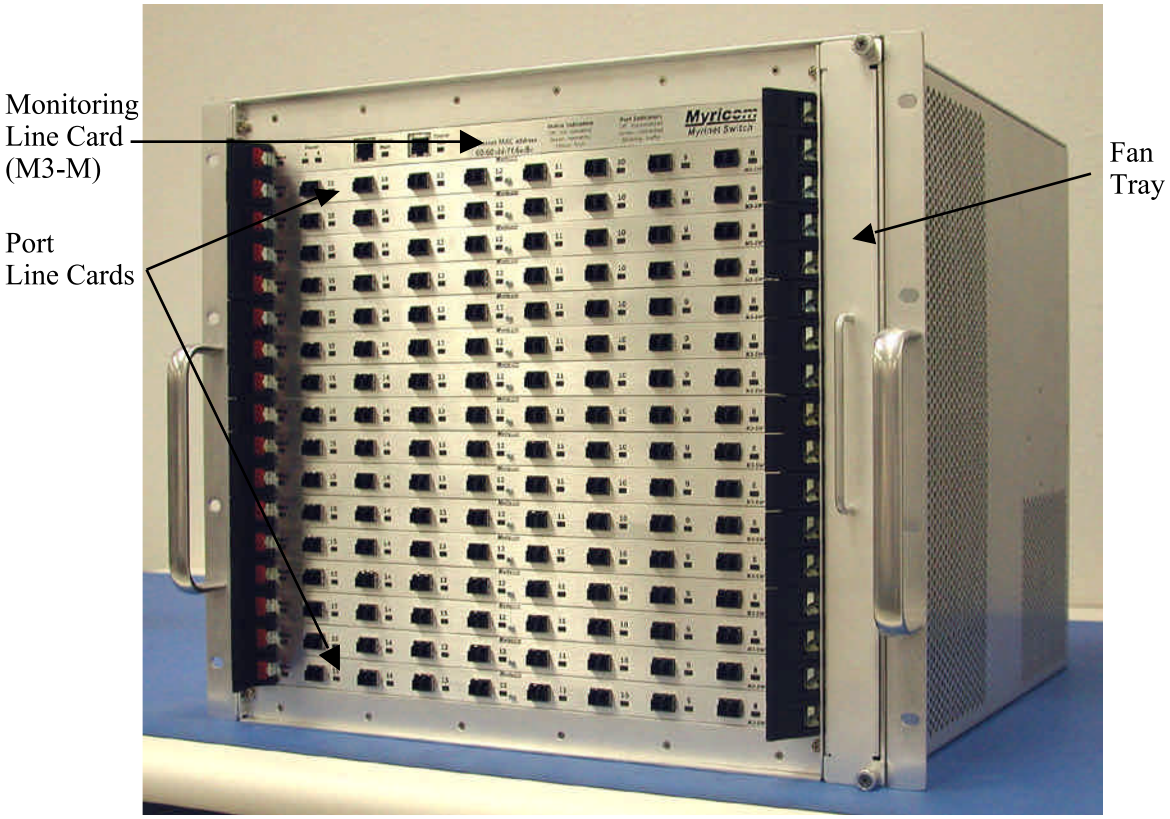
locally by efficient switching power regulators. These power regulators provide a means for selectively turning power off to line cards or backplanes under fault conditions, such as over-temperature.

- a hot-pluggable fan tray. The fan tray is replaced from the front (see page 7). The fans on the fan trays are Panasonic Panaflo fans with an unusually high MTBF of 500,000 hours (each). Also, these fans have tachometer outputs. The fan µC converts the tachometer outputs to fan RPMs, which can be read by the monitoring line card.

The modularity of the Myrinet-2000 switches and switch networks provides configuration flexibility, manufacturing economies, and, most importantly:

- easy expansion or reconfiguration of switches and switch networks already installed, and
- the ability to repair switches by replacing just a line card or the fan tray.

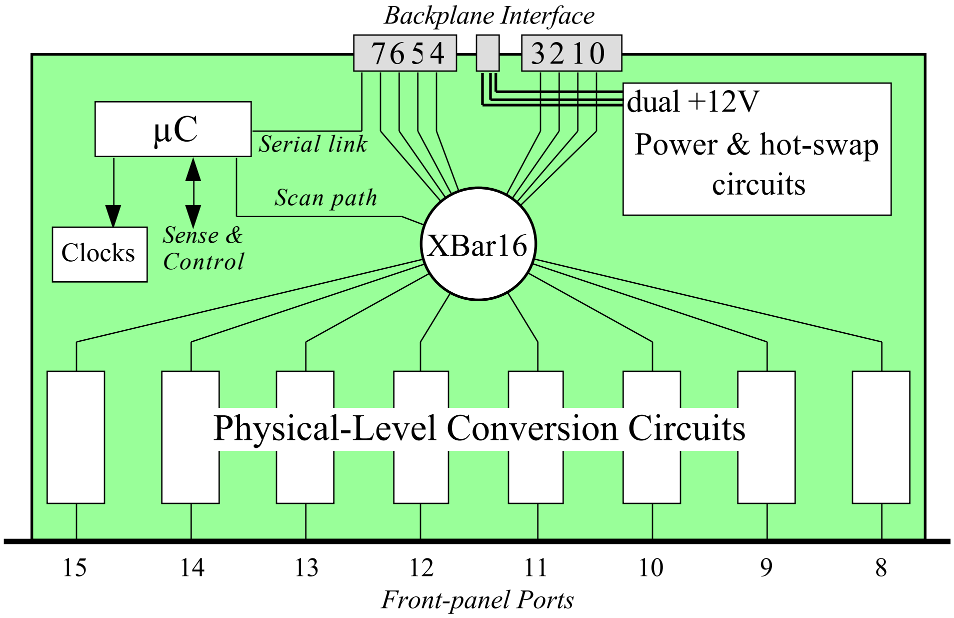
Although the MTBF of Myrinet components is exceptionally high, field failures do occur. The monitoring line card allows nearly any fault to be identified unambiguously. A line card or a fan tray is referred to as a field-replaceable unit (FRU). The enclosure is also an FRU. When hot-swappable power supplies become available, they will also be FRUs. It is obviously much easier to replace a line card, a fan tray, or a power supply than it would be to ship an entire switch network back to Myricom for repair.



M3-E128 with an M3-M monitoring line card and 16 M3-SW16-8F Fiber line cards

4. Port Line Cards

The following block diagram illustrates the structure of any of the port line cards that include an XBar16. This block diagram also approximates the physical layout of these circuit boards.



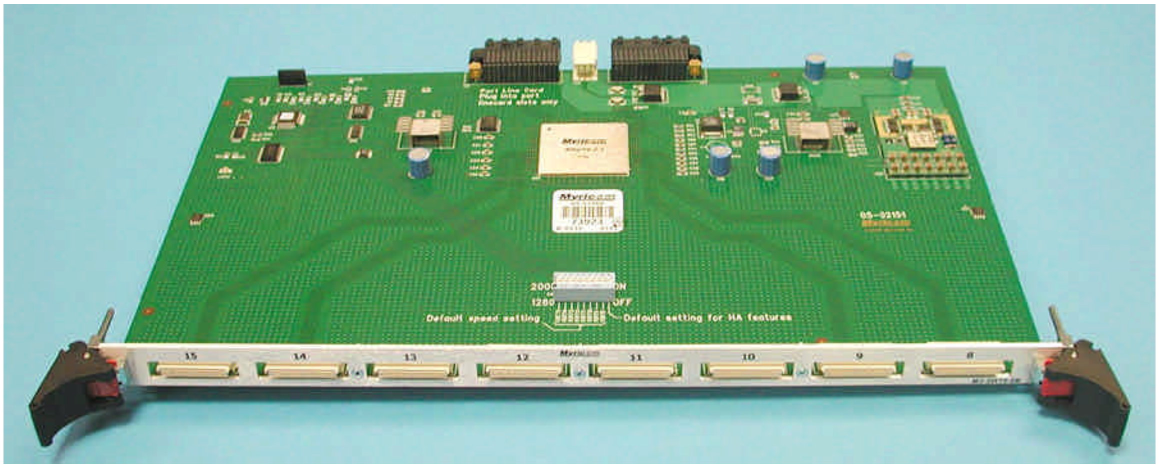
The backplane interface includes 8 SAN-2000 links numbered 0...7, as shown, corresponding to the port numbering of the XBar16. The backplane interface also includes dual +12V power, and a serial-link interface to the monitoring line card. The front-panel ports may be:

- 8 Fiber ports (M3-SW16-8F line card),
- 8 Serial ports (M3-SW16-8S line card),
- 8 SAN ports (M3-SW16-8M and M3-SW16-4DM line cards), or
- 4 legacy LAN ports (M3-SW12-4L line card),

depending upon the Physical-level conversion circuits. The port numbers, 8...15 (8...11 for the M3-SW12-4L line card), appear on the front panel silkscreen, and correspond to the port numbering of the XBar16.

Inasmuch as the native Physical level (PHY) of the XBar16 ports is SAN, there are no conversion circuits for the M3-SW16-8M or the M3-SW16-4DM. The only difference between these two line cards is that for the M3-SW16-8M, the 8 front-panel ports appear on the A link of 8 SAN connectors, with the B link unused; whereas for the M3-SW16-4DM, the 8 front-panel ports appear on the A&B links of 4 SAN connectors.

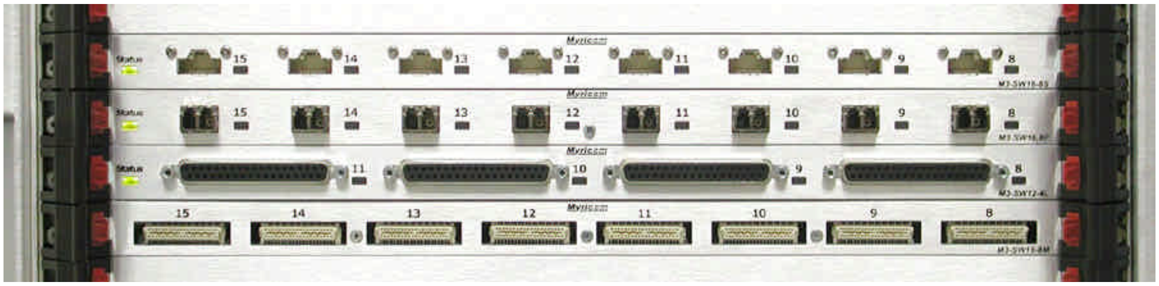
The speed of the front-panel SAN ports of the M3-SW16-8M and M3-SW16-4DM line cards can be set individually to SAN-2000 or to SAN-1280. Both line cards include a set of mechanical switches for setting the default data rate, and for setting the default state of the high-availability (HA) features to 'on' or 'off.' The line card must be ejected from an enclosure slot to change the switch settings. The default settings determined by the switches can be over-ridden via an M3-M monitoring line card.



M3-SW16-8M Line Card

The purpose of the M3-SW12-4L line card is to provide for backward compatibility with 2nd-generation Myrinet-1280 LAN ports.

The following photograph shows the front panels of four types of M3-SW16 port line cards – M3-SW16-8S, M3-SW16-8F, M3-SW12-4L, and M3-SW16-8M (top to bottom) – installed and powered in adjacent slots of a switch enclosure.



The M3-SW16-8S, M3-SW16-8F, and M3-SW12-4L include LEDs for each port. The port LEDs become green when a port is connected through a cable to an active port, and will blink when traffic is flowing.

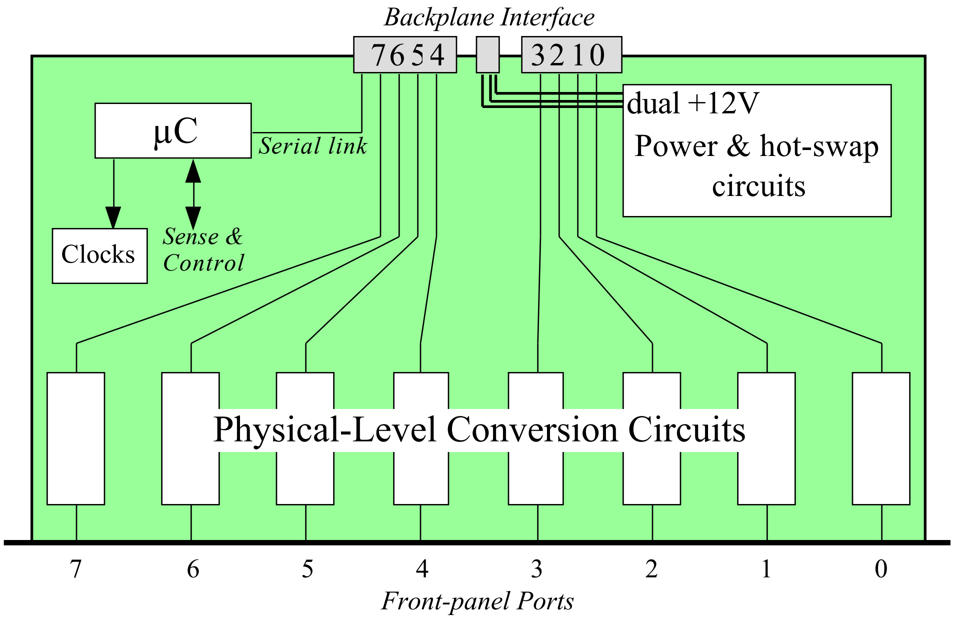
The M3-SW16-8S, M3-SW16-8F, and M3-SW12-4L also include a Status LED for the line card. When a line card is inserted, the Status LED shows green within ~0.5 seconds if the line card has passed its self-test and all voltages, temperatures, and internal status bits are at nominal levels. If the line card has a fault, the Status LED will be yellow. The Status LED is controlled by the line card's µC, and all of the detail information about the status of the line card is available from the monitoring line card. This information includes:

- The type of line card and its serial number,
- All information from the XBar16 scan path (port state, packet counts, CRC errors),

- Internal status bits for the Physical-level conversion circuits,
- Internal voltages, and
- Temperature upstream and downstream in the air flow.

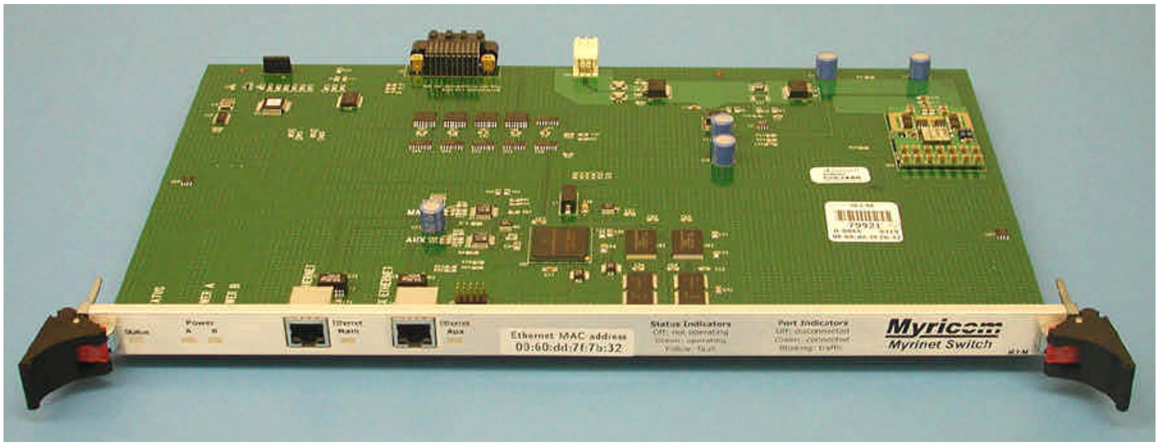
The same monitoring information is available for the M3-SW16-8M and M3-SW16-4DM port line cards, but these line cards do not include any indicator LEDs.

The M3-SPINE-8F and M3-SPINE-8S port line cards are similar to their M3-SW16 counterparts except for the absence of the XBar16:



The front-panel ports are labeled 0...7, corresponding to the ports of the backplane interface. The indicator LEDs are the same, and the µC information is the same except that there is no XBar16 scan path.

5. Monitoring line card



M3-M Monitoring Line Card

The M3-M monitoring line card is powered from the dual +12V sources in the same way as the port line cards, and likewise includes a small µC to monitor internal functions, voltages, and temperatures. The monitoring line card is keyed by the position of the backplane connector so that it can be inserted only in the top slot of an enclosure, and so that port line cards cannot be inserted into this slot.

The main circuitry on this line card is a PowerPC microcontroller with dRAM, flash memory, and dual Ethernet ports. The M3-M is supplied with standard firmware that runs under VxWorks, a real-time operating system. The firmware can be updated via the Ethernet ports.

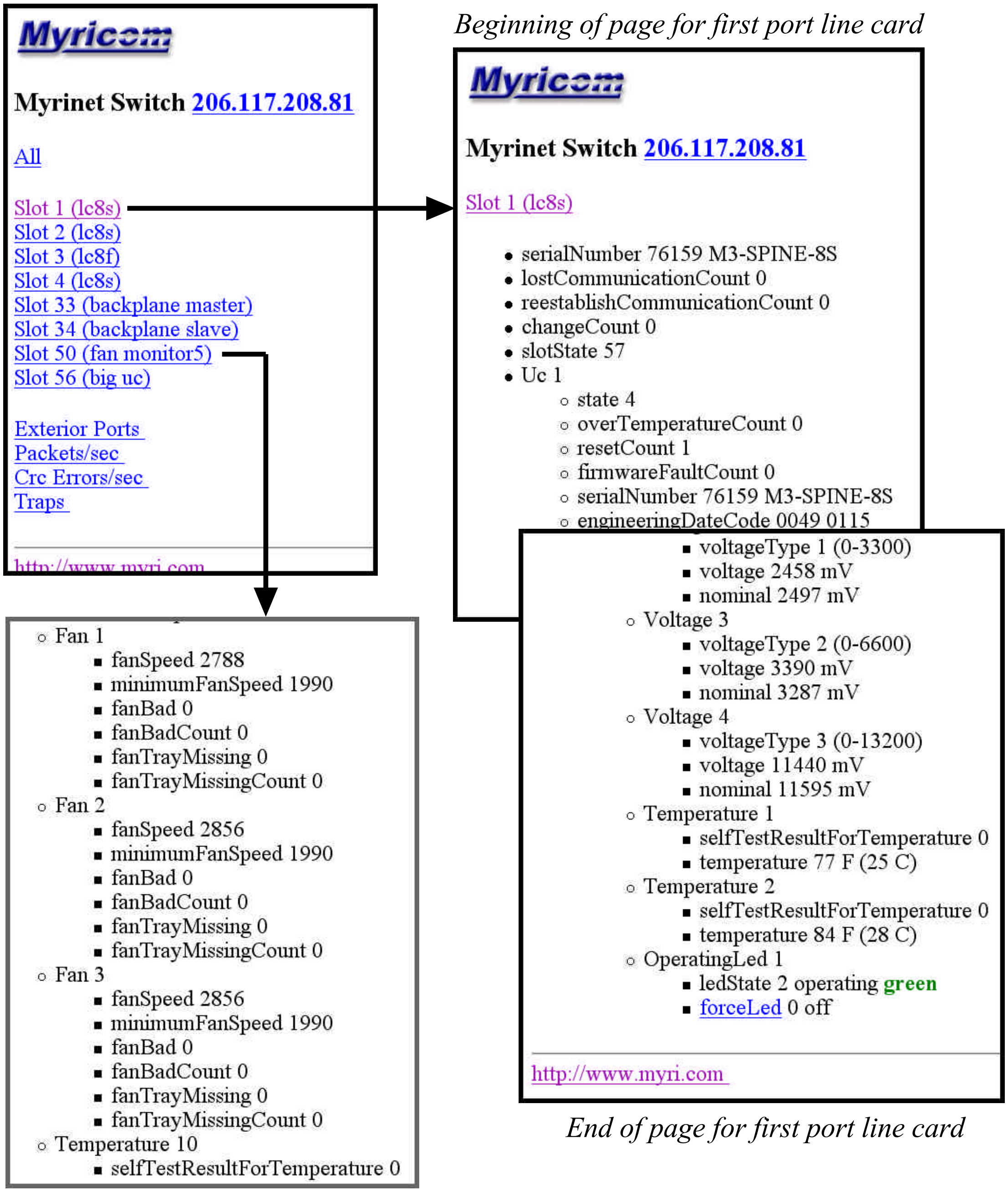
This main microcontroller can communicate internally within a switch enclosure via 26 separate and independent serial links:

- 16 serial links to µCs on port line cards,
- 8 serial links to µCs associated with backplane XBars,
- 1 serial link to the fan-monitoring µC, and
- 1 serial link to its own small µC.

The front panel includes the usual Status LED, green for operating, and yellow for fault. There are two LEDs for the +12V 'A' and 'B' internal power buses. Each of the RJ-45 10-Base Ethernet ports has a LED indicator: off if not connected to an active port, green for connected, and blinking green for traffic. The Ethernet MAC address is provided on a label on the front panel. The Ethernet ports are dual-redundant, and have the same MAC address. The dual-redundant, automatic-failover feature is implemented in the firmware.

The Ethernet ports require a DHCP server on the network to give this small computer an IP address and subnet mask.

The standard firmware supports two software interfaces over the Ethernet connections, SNMP and a web server. Current tools, documentation, the SNMP MIB, and the firmware is linked from the http://www.myri.com/scs/ (Software & Customer Support) page. The following collection of screen snapshots from a web browser will give you some impression of the capabilities of the monitoring and of the web-browser interface.



Home page of an M3-E32

Myrinet Switch 206.117.208.81

All

Slot 1 (lc8s)
Slot 2 (lc8s)
Slot 3 (lc8f)
Slot 4 (lc8s)
Slot 33 (backplane master)
Slot 34 (backplane slave)
Slot 50 (fan monitor5)
Slot 56 (big uc)

Exterior Ports
Packets/sec
Crc Errors/sec
Traps

http://www.myri.com

Beginning of page for first port line card

Myrinet Switch 206.117.208.81

Slot 1 (lc8s)

- serialNumber 76159 M3-SPINE-8S
- lostCommunicationCount 0
- reestablishCommunicationCount 0
- changeCount 0
- slotState 57
- Uc 1
    - state 4
    - overTemperatureCount 0
    - resetCount 1
    - firmwareFaultCount 0
    - serialNumber 76159 M3-SPINE-8S
    - engineeringDateCode 0049 0115

End of page for first port line card

- voltageType 1 (0-3300)
- voltage 2458 mV
- nominal 2497 mV

- Voltage 3
    - voltageType 2 (0-6600)
    - voltage 3390 mV
    - nominal 3287 mV
- Voltage 4
    - voltageType 3 (0-13200)
    - voltage 11440 mV
    - nominal 11595 mV
- Temperature 1
    - selfTestResultForTemperature 0
    - temperature 77 F (25 C)
- Temperature 2
    - selfTestResultForTemperature 0
    - temperature 84 F (28 C)
- OperatingLed 1
    - ledState 2 operating **green**
    - forceLed 0 off

http://www.myri.com

A part of the fan-monitor page

- Fan 1
    - fanSpeed 2788
    - minimumFanSpeed 1990
    - fanBad 0
    - fanBadCount 0
    - fanTrayMissing 0
    - fanTrayMissingCount 0
- Fan 2
    - fanSpeed 2856
    - minimumFanSpeed 1990
    - fanBad 0
    - fanBadCount 0
    - fanTrayMissing 0
    - fanTrayMissingCount 0
- Fan 3
    - fanSpeed 2856
    - minimumFanSpeed 1990
    - fanBad 0
    - fanBadCount 0
    - fanTrayMissing 0
    - fanTrayMissingCount 0
- Temperature 10
    - selfTestResultForTemperature 0

6. Configurations for up to 128 Hosts

Optimal, full-bisection networks for 128 or fewer hosts do not require any detailed understanding of network topologies.

A network for 8 or fewer hosts requires only the front-panel ports of an M3-SW16 line card. Although small switches can be ordered with single product codes, here are the recipes in order to give you a sense of the modularity. All of the configurations listed include the optional M3-M monitoring line card. If it is omitted, it should be replaced with an M3-BLANK panel.

Product Code	*Description*	*Modular assembly*
M3F-SW8M	**Fiber switch:**	M3-E16 + M3-M + M3-BLANK + M3-SW16-8F
M3S-SW8M	**Serial switch:**	M3-E16 + M3-M + M3-BLANK + M3-SW16-8S
M3M-SW8M	**SAN switch:**	M3-E16 + M3-M + M3-BLANK + M3-SW16-8M

As described on page 12, 16-port switches are, when possible, assembled from an M3-SW16 line card and an M3-SPINE line card in the M3-E16 enclosure. Here are the recipes for the modular assembly of the 16-port switches (with monitoring) that appear in the product list:

Product Code	*Description*	*Modular assembly*
M3F-SW16M	**Fiber switch:**	M3-E16 + M3-M + M3-SPINE-8F + M3-SW16-8F
M3S-SW16M	**Serial switch:**	M3-E16 + M3-M + M3-SPINE-8S + M3-SW16-8S
M3M-SW16M	**SAN switch:**	M3-E16 + M3-M + 2*M3-SW16-8M

There is no M3-SPINE-8M product due to signal-integrity limitations on carrying SAN signals across backplanes and then across line cards to cables. Hence, the 16-port, all-SAN switch has a redundant layer of switching (see page 12).

What if you wanted a switch to connect 8 hosts with Fiber links and an additional 8 hosts with SAN links? You can readily specify a 16-port switch for this application as either:

Product Code	*Description*	*Modular assembly*
(none)	**8 Fiber** +	M3-E16 + M3-M + M3-SPINE-8F + M3-SW16-8M, or
	8 SAN	M3-E16 + M3-M + M3-SW16-8F + M3-SW16-8M

The second of these configurations has a extra level of switching and has a slightly higher cost, but uses line cards that are more versatile. (The extra level of switching is nearly invisible from the standpoint of software, performance, and reliability.) If you planned to expand this cluster to up to 64 hosts, it would make sense to leave room for expansion by specifying:

Product Code	*Description*	*Modular assembly*
(none)	**8 Fiber +**	M3-E64 + M3-M + 6*M3-BLANK +
	8 SAN	M3-SW16-8F + M3-SW16-8M

For each 8 hosts added for a particular PHY, just add another line card.

All configurations larger than 16 hosts and up to 128 hosts escape the (page 12) 'special case' optimization of the M3-E16 with an M3-SPINE line card, and use all M3-SW16 line cards. Indeed, the configuration of Myrinet-2000 switch networks for $16 < N \leq 128$, where N is the number of hosts, is particularly simple.

> First, count the number of hosts you wish to connect using each Myrinet PHY (Fiber, Serial, SAN). You'll need one M3-SW16-8{F|S|M} line card for each 8 hosts of each PHY. Then, pick an enclosure that will accommodate all of these port line cards, plus as many spare slots as you may expect to add line cards when you expand the cluster[3].

Example 1: You are under contract to build a cluster with 20 hosts within a cabinet, for which you have decided to use SAN links, and 6 external hosts connected by Fiber links. You'll need 3 M3-SW16-8M line cards, and 1 M3-SW16-8F line card, in an enclosure that will accommodate at least 4 port line cards (M3-E32). The cluster is for a fixed, embedded application, and will not grow, and rack space is at a premium. So, you order the following components for the switch network:

Qty	*Modular component*
1	M3-E32 (enclosure)
1	M3-M monitoring card (highly recommended) or M3-BLANK
3	M3-SW16-8M (port line card with 8 SAN ports)
1	M3-SW16-8F (port line card with 8 Fiber ports)

Example 2: You plan to build a 64-host cluster that you expect to expand soon to 128 hosts, and perhaps to an even larger size eventually. Accordingly, you order the following components for the switch network:

Qty	*Modular component*
1	M3-E128 (enclosure)
1	M3-M monitoring card (highly recommended) or M3-BLANK
8	M3-SW16-8F (port line card with 8 Fiber ports)
8	M3-BLANK (blank front panels)

Hosts and line cards can then be added incrementally up to 128 hosts.

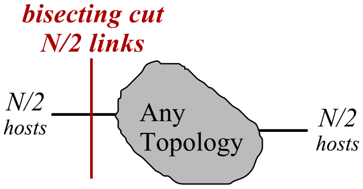
Scenario 2a: The management (or sponsor) is so happy with the results from the 64-host and then the 128-host cluster that they want you to expand it to 256 hosts. Every 128-host switch component, as well as interfaces and cables, can be re-used effectively in a 256-host configuration (see the following sections).

[3] We're not trying to promote the sale of empty enclosure slots or of M3-BLANK panels. Myricom sales records show that *most* Myrinet clusters will be expanded during their lifetime, and often within months of their initial installation.

7. Topology Concepts

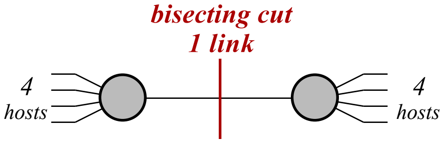
A message-passing network's *minimum bisection,* measured in links, is defined mathematically as the minimum number of links crossing any cut that bisects the hosts (half of the hosts on one side of the cut, the other half on the other side). The minimum bisection provides a metric for the minimum traffic-handling capacity of a network, no matter what the communication patterns between the hosts may be.

The upper bound on the minimum bisection of a network of *N* hosts is *N/2* links, because, no matter what the internal topology of the network, there will be bisecting cuts possible across half of the host links:



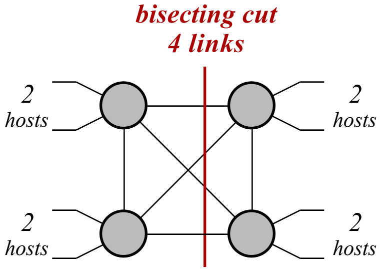
A network topology that achieves this *N/2* upper bound is said to have *full* bisection.

As an example of the relevance of the minimum bisection, suppose that we wished to connect 8 hosts, and the largest (crossbar) switches available had 5 ports. One way to connect the 8 hosts would be:



The bisecting cut that exhibits the minimum bisection is evident. If the 4 hosts on the left were sending messages to the 4 hosts on the right, and vice versa, the total traffic-handling capacity of the network would be throttled to that of a single link. The host links would be limited to an average of 1/4 of the link data rate, corresponding to the bisection being 1/4 full.

Here is a full-bisection network to connect the 8 hosts with 5-port switches:



The bisecting cut shown is just one of many that must be considered to determine that the minimum bisection of this topology is 4 links. For a large, unstructured network, there are so many ways to bisect the set of hosts, and so many possible cuts for each bisection, that the minimum bisection of the network may be difficult to compute. However, large networks will generally have a regular structure that can be used to simplify the calculation. Also, at least for full-bisection networks, there are some mathematical shortcuts that can be used to verify that the bisection is full.

A message-passing network is said to be *rearrangeable* if it can route any permutation without blocking. For example, an 8-host rearrangeable network must be able to route:

from host	0	1	2	3	4	5	6	7
	↓	↓	↓	↓	↓	↓	↓	↓
to host	3	5	1	0	7	6	2	4

and every other of the 8! (40,320) possible permutations. Unlike a single crossbar switch, a special case of a rearrangeable network, the set of routes used from host to host is permitted in a rearrangeable network to be different for different permutations.

It should be evident that a network that is rearrangeable necessarily has full bisection. The converse is not true. A network may have full-bisection but not be rearrangeable.

For the distributed computations performed on clusters, it is generally not practical to coordinate the routing to take advantage of a network being rearrangeable. However, it is often easiest to demonstrate that a network has full bisection by showing that it is rearrangeable.

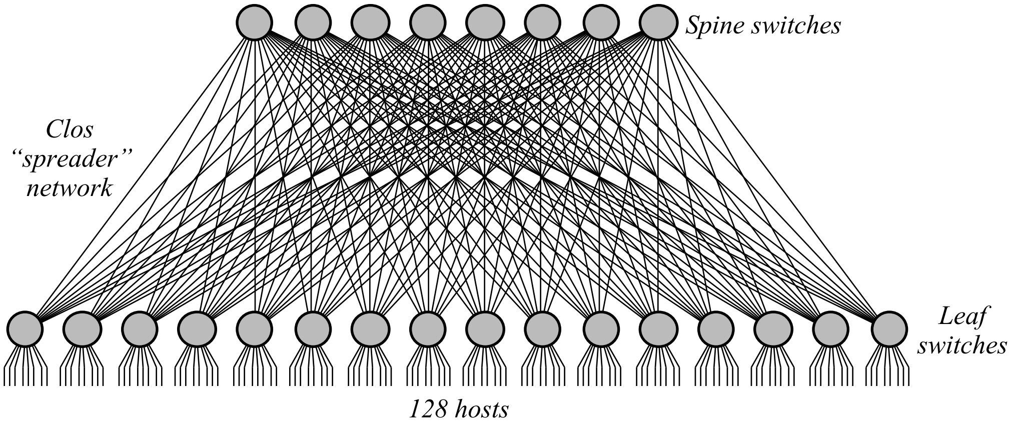
Inasmuch as the path-formation latency of Myrinet switches is much smaller than software latency, the maximal or average *diameter*, measured in switches traversed, is a relatively unimportant metric of the network topology.

8. Clos Networks

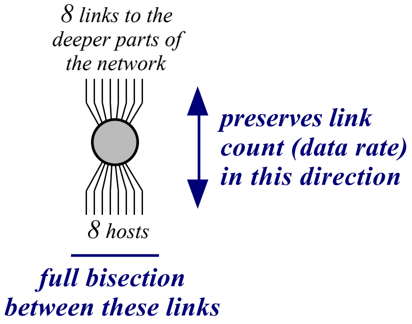
Clos networks are named for Charles Clos, who introduced them in a paper titled “A Study of Non-Blocking Switching Networks,” published in the *Bell System Technical Journal* in March 1953. Clos networks are rearrangeable: researchers at that time were seeking ways to route arbitrary permutations of telephone connections. Clos networks thus exhibit full bisection. The Clos network topology has other excellent properties – scaling to large sizes, modularity, and multiple-path redundancy – that make it an ideal topology for cluster networks.

Our basic building block, the crossbar switch, is itself a rearrangeable and full-bisection network. However, technology sets a practical upper limit on the degree of crossbar switches. For Myrinet this limit is currently 16-port crossbar switches, which are implemented in a single chip, the XBar16. The Clos topology provides systematic ways of connecting these 16-port crossbar switches to serve large numbers of hosts, and to do so in a way that achieves full bisection.

As a first step toward understanding these constructions, here is the Clos network of 16-port switches that connects 128 hosts, a network that we’ll refer to as the Clos128:



Observe that 8 ports of the leaf switches connect to hosts; the other 8 ports connect to the spine switches. It is possible that all of the packet traffic to and from the 8 hosts connected to a single leaf switch is to or from hosts connected to other leaf switches; thus, the leaf switches must have at least as many links to the spine switches as to the hosts.


8 links to the deeper parts of the network
preserves link count (data rate) in this direction
8 hosts
full bisection between these links

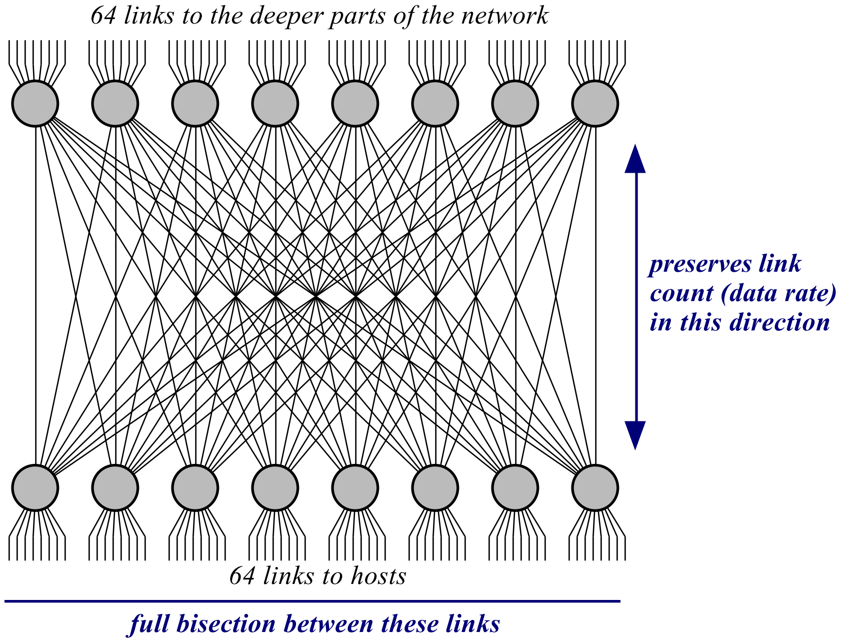
More generally, it is necessary for any component of a full-bisection network to provide at least as many links toward the deeper parts of the network fabric as toward the hosts. Another way to look at this necessity is to consider a bisecting cut that crosses the 8 host links. There is another, adjacent, bisecting cut across the 8 upper links. Thus, any structure used as a building block for larger networks must have this property of preserving link count and data rate.

The proof that the 128-host Clos network of 16-port switches is a rearrangeable network – a stronger statement than that it exhibits full bisection – proceeds by showing how a set of routes can be found for any permutation. Once the problem is cast into combinatorial terms, the proof is made by appealing to Hall's Theorem on systems of distinct representatives.

9. Clos Networks for more than 128 Hosts

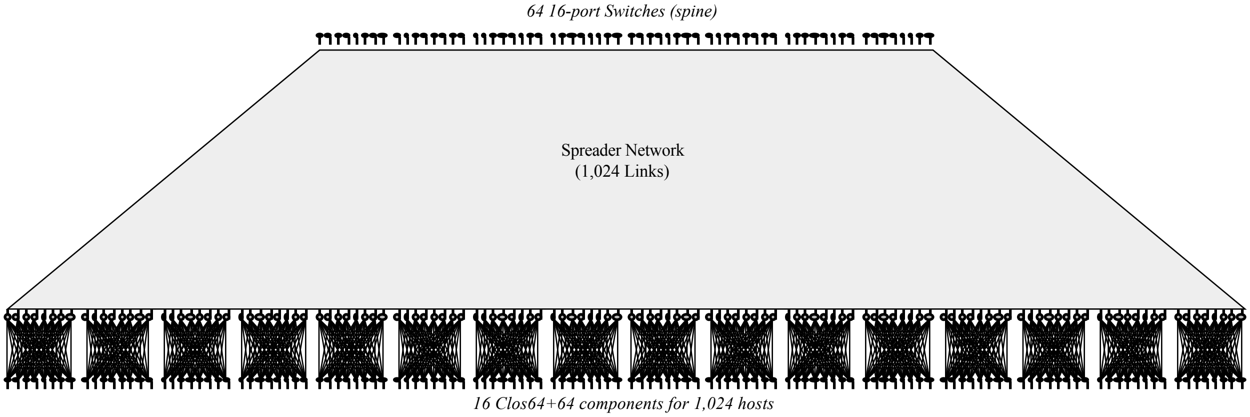
For the Clos128 network, the spine is 8 16-port switches, and 16-port switches are our limit. How do we scale the network to a larger size?

We can apply the principle illustrated above for the leaf switches of the 128-host Clos network. Let us use the standard construction for a 64-host network, which requires a spine of 8 8-port switches. However, we'll leave the other 8 ports of the spine XBar16s switches to connect to another, deeper layer in the network fabric.


64 links to the deeper parts of the network
preserves link count (data rate) in this direction
64 links to hosts
full bisection between these links

With respect to the host links, the topology is a 64-host Clos network; thus, this network will exhibit full bisection (and is rearrangeable) in the communication between the 64 hosts. In addition, the network preserves 64-link data rate between the hosts and the deeper parts of the network fabric, so that if these 64 hosts are communicating entirely with hosts outside their own group, there are enough links to carry all the traffic. This network is referred to as a "Clos64+64," and it can be implemented with an M3-E128 enclosure with 8 M3-SW16 line cards and 8 M3-SPINE line cards.

In analogy to the structure of the 128-host Clos network, we can build a 1,024-host Clos network by connecting 16 of the Clos64+64 components to a spine composed of 64 16-port switches:


64 16-port Switches (spine)
Spreader Network
(1,024 Links)
16 Clos64+64 components for 1,024 hosts

(A better diagram is linked from http://www.myri.com/myrinet/m3switch/guide/.)

Note that the diameter of this network is 5 switches. A packet may traverse two switches in its upward path through a Clos64+64, one spine switch, and two switches in its downward path through a Clos64+64.

For the spreader network, the 64 upper ports of each Clos64+64 should connect to one port on each of the 64 16-port spine switches. Any cabling plan that satisfies this description is correct; however, systematic cabling plans can be found linked from http://www.myri.com/myrinet/m3switch/guide/.

The most compact and economical way to implement the spine is with 8 M3-E128 enclosures, each with 16 M3-SPINE line cards, providing 8 16-port switches in each M3-E128.

If you recall the way in which a networks smaller than the 128-host Clos network are constructed (page 11) by using a 16-port switch as if it were 2 8-port switches, or 4 4-port switches, this same technique can be used with the Clos64+64 to build networks for 512 or for 256 hosts.

For example, a network for 256 hosts would require 4 Clos64+64 units and 64 4-port switches. Instead of 64 4-port switches, we could instead use 16 16-port spine switches. The 64 upper ports of each Clos64+64 should connect to four ports on each of the 16 16-port spine switches.

This basic scheme is suitable for any multiple of 64 hosts from 192 to 1024 hosts. For example, a 320-host network requires 5 of the Clos64+64 units. In the abstract, the spine should consist of 64 5-port switches. The minimal-cost implementation of the spine requires 3 M3-E128 enclosures and 40 M3-SPINE line cards. If you refer to the port layout and numbering linked from http://www.myri.com/myrinet/m3switch/guide/ , you will see that each M3-SPINE line card in an M3-E128 connects to the same port of all 8 backplane switches. Thus, if an M3-E128 is populated with 15 M3-SPINE line cards, it provides 8 15-port switches, which can be used as 24 5-port switches. Two such units, and another with 10 M3-SPINE line cards for 8 10-port switches used as 16 5-port switches, provide a total of 64 5-port switches. This optimized implementation would need to be recabled to expand the network to 384, 448, or 512 hosts, whereas the more symmetrical implementation of the spine with 4 M3-E128 enclosures, each with 10 M3-SPINE line cards, can be expanded without recabling.

The Clos networks for $16 < N \leq 128$ are particularly efficient because the switch-to-switch links are all contained within the "network in a box." The discontinuity in the switch-network cost from $16 < N \leq 128$ to $N > 128$ is due in part to the diameter of the Clos network increasing from 3 to 5, and in part to the need for a spreader network implemented with cables. A Clos network for $N = 8192$ can be implemented with 128 of the Clos64+64 units connected to a spine of 64 Clos128 networks. This network has a diameter of 7, but still requires only one external spreader network built with cables between the Clos64+64 units and the Clos128 units. Thus, the cost/host of Clos networks in the $1024 < N \leq 8192$ range is similar to the cost/host of Clos networks in the $128 < N \leq 1024$ range, and the optimization techniques are the same.

For additional information about Myrinet Clos networks for $128 < N \leq 8192$, please see the supplemental information at http://www.myri.com/myrinet/m3switch/guide/ .