Myrinet-on-VME Protocol Specification Draft Standard

VITA 26-199x
Draft 1.1
31 August 1998

This standard is being prepared by the
VITA Standards Organization (VSO)
and is
Unapproved.

Do not specify or claim conformance to this document.

VSO is the Public Domain Administrator of this
specification and guards the contents
from change except by sanctioned
meetings of the task group under due process.

VITA Standards Organization
7825 E. Gelding Drive, Suite 104
Scottsdale, AZ 85260

Tel: 602-951-8866
Fax: 602-951-0720
URL: “http://www.vita.com”

1. General

1.1 List of Tables

1.2 List of Figures

1.3 List of Rules

1.4 List of Recommendations

1.5 List of Permissions

1.6 List of Observations

1.7 Purpose

The purpose of this specification is to provide the information needed to design VMEbus boards that interface with the Myrinet, high-performance, packet network.

Myricom Incorporated of Arcadia, California, developed the specification of Myrinet but makes no proprietary claims to this specification or to any technique. Myricom disclaims responsibility or liability for its use, or for any infringement of patents or other rights of third parties resulting from its use.

1.8 Scope

This standard describes the high-performance, inter-computer, Myrinet packet network that is fully compatible with existing VMEbus standards and their extensions. This standard addresses communication between VME boards using interconnect either on the front panel or on the backplane. The communication may use cables or an overlay (such as a backplane). The standard defines the interface between a VME board and Myrinet, allowing not only intra-subrack, board-to-board communication, but also a uniform extension for inter-subrack, inter-cabinet, and even local-area-network (LAN) communication.

This standard includes, either directly or by reference, the specification of the Data Link level, timing information, character set, signals, and the details of the connectors.

1.9 Task Group Members

When this standard was drafted, the Task Group for this standard had the following membership:

	Name		Company	E-address
1	Gorky Chin		Vista Controls Corporation	gorky@vistacc.com
2	Elwood Parsons	B	AMP	etparson@amp.com
3	J. J. Dumont		Framatome Connectors France	jdumont@iway.fr
4	David Wright		Hybricon Corporation	davidw@hybricon.com
5	Holly Sherfinski	B	Harting Inc. of North America	Holly.Sherfinski@harting.com
6	David Robak	B	Harting Inc. of North America	David.Robak@harting.com
7	Richard Jaenicke	B	Sky Computers, Inc.	jaenicke@sky.com
8	Istvan Vadasz		Force Computers	isva@force.de
9	Michael Thompson	B	Schroff, Inc.	m_thompson@ids.net
10	Martin Blake	B	VERO Electronics	martin_blake@vero-uk.com
11	Richard Morgan		Spectrum Signal Processing	rick_morgan@spectrumsignal.com
12	John Bratton		VERO Electronics	john_bratton@vero-usa.com
13	Danny Cohen	B	Myricom, Inc.	cohen@myri.com
14	Jim Waggett	B	CSPI, Inc.	jwaggett@cspi.com
15	Robert McKee		MITRE Corporation	rmckee@mitre.org
16	Michael Munroe		ERNI Components Inc.	mmunroe@compuserve.com
17	Wade Peterson	O	Consultant to Industry	peter299@maroon.tc.umn.edu
18	Douglas Endo	B	Raytheon System Company	dendo@msmail4.hac.com
19	Jing Kwok		Nexus Technology, Inc..	
20	John Rynearson	B	VITA	techdir@vita.com
21	Ray Alderman	B	VITA	exec@vita.com
22	Harry Andreas	B	Raytheon System Company	handreas@msmail4.hac.com
23	Joe Bedard	B	Hewlett-Packard Co.	bedard_j@apollo.hp.com
24	Tony Lavely	B	Mercury Computers, Inc.	atl@mc.com
25	Bob Patterson	B	AMP	rapatter@amp.com

The “B” in the third column indicates participation in the ballot group.
The “O” in the third column indicates Observers.

The following companies were on the sponsor balloting committee:

AMP
CSPI
Myricom

When the ANSI Standards Board approved this standard on xx/xx/199x, the balloting committee had the following membership:

Ray Alderman	VITA	<exec@vita.com>
Harry Andreas	Raytheon Systems Company	<haandreas@mail.hac.com>
Danny Cohen	Myricom, Inc.	<cohen@myri.com>
Adrian Cox	Transtech	<apc@transtech.co.uk>
Doug Doerfler	Sandia National Labs	<dwdoerf@sandia.gov>
Richard Jaenicke	Sky Computers, Inc.	<jaenicke@sky.com>
James Judd	Lockheed Martin Corp.	<jjudd@motown.lmco.com>
Elwood Parsons	AMP	<etparson@amp.com>
John Rynearson	VITA	<techdir@vita.com>
Michael Stern	CSPI	<mstern@cspi.com>
Rich Tallarico	Lockheed Sanders	<rtallari@sanders.com>
Michael Thompson	Schroff, Inc.	<m_thompson@ids.net>
James Waggett	CSPI	<jwaggett@cspi.com>

1.10 VSO and Other Standards

For information on other standards being developed by VSO, VME Product Directories, VME Handbooks, or general information on the VME market, please contact the VITA office at the address or phone number given on the front cover.

1.11 Draft Summary

This draft standard incorporates the comments received during the VSO ballot period and the ANSI ballot period throughout August 1998.

1.12 Draft History

Version	Date	Comments
D0.1	March, 1996	First draft, derived from various documents
D0.2	October 1997	Separation of the Data Link layer specifications from the Physical layer specifications
D0.3	November 1997	Examples added
D0.4	January 1998	Comments from ballot period (Dec-97) incorporated
D0.5	27 January 1998	Additional comments (throughout Jan-98) incorporated
D1.0	31 July 1998	Additional comments (throughout July-98) incorporated
D1.1	31 August 1998	Observation 3.5-1 (Progress guarantee) incorporated

1.13 Task Group Ballot

Task Group Ballot was conducted from 21 November 1997 through 31 December 1997.

The ANSI Ballot was conducted from 22 April 1998 through 27 July 1998.
A reballot was conducted from 29 July 1998 through 31 August 1998.

2. Introduction to the Myrinet-on-VME Standard

2.1 Definitions and References

See Appendix D for the glossary of the new terminology specific to the added features described in this standard. Some of terminology described in this standard is repeated there.

2.2 References

The following publications are used in conjunction with this standard.

IEC 61076-4-101	IEC Standard for Hard Metric 2 mm Connectors
ANSI/VITA 1-1994	VME64 Standard, Approved April 10, 1995
VITA 1.1-199x	VME64 Extensions Draft Standard, Draft 1.4, September 20, 1996.[1]
ANSI/VITA 1.3-1997	ANSI/VITA 1.3-1997, VME64x 9U x 400 mm Format
ISO 7498	ISO Reference Model (ISORM) for Open Systems Interconnection, October 1984
NRZI	Chapter 19 by David R. Brown and Jack I. Raffel in Handbook of Automation, Computation, and Control, Ed. Grabbe, Ramo, and Wooldridge, John Wiley & Sons, 1959.

2.3 Connector Notes

The front-panel application uses SAN (System-Area Network) connectors that are in the Microstrip family. This family of controlled-impedance connectors can be used as board-to-board connectors as well as board-to-cable connectors.

Backplane applications use either P0 or P4 (as defined by ANSI/VITA 1.3-1997) as specified by IEC 61076-4-101. It is a 95 pin connector (P0/J0) that fits between the VME64x P1/J1 and P2/J2 connector pairs or the same connector (P4/J4) that fits between P2/J2 and P3/J3 or between P2/J2 and P5/J5. The 95 pin connector is a 2 mm Hard Metric style connector.

The interconnection of the VME environment to external environments is based on LAN cables that use low-voltage differential signaling (LVDS) over twisted-pairs, and use DB-37 connectors.

2.4 Standard Terminology

[1] An unapproved draft standard being prepared by the VITA Standards Organization (VSO) that has been proposed as an American National Standard and is currently in the ANSI Ballot Process. When it achieves ANSI recognition this reference will be changed accordingly.

To avoid confusion and to make very clear what the requirements for compliance are, many of the paragraphs in this standard are labeled with keywords that indicate the type of information they contain. The keywords are listed below:

Rule
Recommendation
Suggestion
Permission
Observation

Any text not labeled with one of these keywords describes the Myrinet structure or operation. It is written in either a descriptive or narrative style. These keywords are used as follows:

Rule:
Rules form the basic framework of this draft standard. They are sometimes expressed in text form and sometimes in the form of figures, tables or drawings. All rules shall be followed to ensure compatibility between designs. All rules use the "shall" or "shall not" words to emphasize the importance of the rule. The "shall" and "shall not" words are reserved exclusively for stating rules in this draft standard and are not used for any other purpose.

Recommendation:
Whenever a recommendation appears, designers would be wise to take the advice given. Doing otherwise might result in some awkward problems or poor performance. While the Myrinet-on-VME architecture has been designed to support high-performance systems, it is possible to design a system that complies with all the rules but has abysmal performance. In many cases a designer needs a certain level of experience in order to design boards that deliver top performance. Recommendations found in this draft standard are based on this kind of experience and are provided to speed the learning curve. The "should" and "should not" words are reserved exclusively for stating rules in this draft standard and are not used for any other purpose.

Suggestion:
A suggestion contains helpful but not vital advice. The reader is encouraged to consider the advice before discarding it. Some design decisions are difficult without prior experience. Suggestions are included to help a designer who has not yet gained this experience. Some suggestions pertain to designing boards that can be easily reconfigured for compatibility with other boards, or to designing boards to ease system debugging.

Permission:
In some cases a rule does not specifically prohibit a certain design approach, but the reader might be left wondering whether that approach might violate the spirit of the rule or whether it might lead to some subtle problem. Permissions reassure the reader that a certain approach is acceptable and will cause no problems. The lower case word "may" is reserved exclusively for stating permissions in this draft standard and is not used for any other purpose.

Observation:
Observations do not offer any specific advice. They usually follow naturally from what has just been discussed. They spell out the implications of certain rules and bring attention to things that might otherwise be overlooked. They also give the rationale behind certain rules so that the reader understands why the rule must be followed.

2.5 The Structure of the Myrinet-on-VME Standard

This document specifies the Myrinet-on-VME Standard. It is written principally as an exposition, but interspersed with rules written in the restrictive or formal language of specifications. In addition, terms with formal meanings are shown in italics when first introduced and defined.

Section 2.6 describes Myrinet, in general, and its operation.

Chapter 3 specifies Myrinet at Level 2 of the ISO Reference Model (ISORM) for Open Systems Interconnection. This is the Data Link layer, the low level logical protocol. This chapter specifies also the abstract requirements that the Data Link layer imposes on the Physical layer underneath.

The Data Link level is the most appropriate starting point for the Myrinet specifications. Myrinet may be supported by various Physical-level (level 1 of the ISORM) implementations, for example, with electrical cables or a backplane, each with their own specified characteristics. At the Data Link level, however, Myrinet links are specified in more abstract terms that apply to all of the Physical-level implementations. Similarly, networks such as ethernet, ATM, and FDDI are defined at the Data Link level, and have multiple Physical-level implementations. The Data Link level also provides the foundation for higher protocol levels. The constancy of the Data Link specification, even when the Physical-level technology may change, protects investments in software at these higher levels. Chapter 3 defines the abstract Myrinet.

Chapter 4 summarizes the requirements that any Physical level implementation of Myrinet must meet, and defines the data rates of 640, 1,280 and 2,560 Mbits/sec.

Chapter 5 and Chapter 6 define the SAN (System-Area Network) and the LAN (Local-Area network) Physical levels of Myrinet-on-VME respectively. SAN implementations are intended for intra-cluster applications (including intra-PCB, intra-subrack, intra-cabinet, and even inter-cabinet) whereas the LAN implementations are intended for inter-cluster applications.

Both Chapter 5 and Chapter 6 specify for their types of networks:

(1) The channel character set (data bytes and control symbols)
(2) The encoding of the characters on signals
(3) The electrical characteristics of these signals (including rate, timing, and voltage)
(4) Pin-assignments, connectors, and cables.

The inter-board SAN links are expected to use either the front panel or the backplane. The LAN links are expected to run either from front-panel or from rear-panel connectors for high-performance communication with other clusters.

Chapter 5 specifies the standards for Myrinet-1280/SAN, Myrinet-2560/SAN, Myrinet-1280/BP, and Myrinet-2560/BP. Chapter 6 specifies the standards for Myrinet-640/LAN, and for Myrinet-1280/LAN

The above designations of the Myrinet standards include both the speed of the implementation and some reference to the implementation method (similar to the definition of the various ethernet variants such as E10, E100, E1000, E100BaseT, and E1000BaseT).

Appendix A shows several examples of Myrinet-on-VME. Appendix B describes the existing practices in Myrinet routing. Appendix C lists the assignment of Myrinet packet types, as of January 1998. Appendix D provides a glossary.

The following is a "map" of this document:

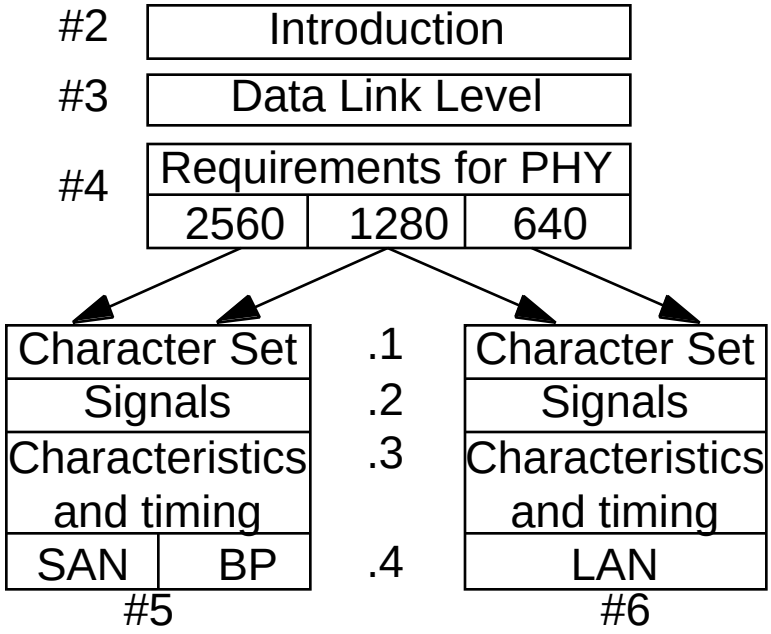

Figure 2.5—1: The map of this document

2.6 General Description of Myrinet *(not a part of the standard)*

Speed

Myrinet connects computing nodes (cards, workstations, and PCs) through full-duplex point-to-point Gigabit-per-second links (e.g., 1.28+1.28 Gbits/sec), and low-latency cut-through switches.

Robustness and Convenience

Any network topology is allowed, because the network is self-configuring. The Myrinet host interfaces may periodically map the network, both to provide fault-tolerance through the dynamic use of alternate routes to circumvent faults, and also as a convenience for installation.

Network Throughput

As a switched network, analogous in structure to Ethernet segments connected by switches, a Myrinet can carry many packets concurrently, each traversing the network at 1.28 Gbits/s. Unlike unswitched Ethernet or FDDI networks, which share a common communication medium, the aggregate traffic capacity of a Myrinet increases with the number of hosts. For example, a Myrinet connecting 16 hosts with a single 16-port switch can carry 16 packets at once, an aggregate traffic capacity (and switch-bisection data rate) of 20.48 Gbits/s.

Versatility

Myrinet packets may be of any length, and thus can encapsulate other types of packets, including IP and special APIs packets, without an adaptation layer.

Each packet is identified by type, so that Myrinet, like Ethernet, can carry packets of many types or protocols simultaneously.

Software Interfaces and End-to-End Performance

Myrinet users achieve short-message latencies between UNIX user processes smaller than 5 μseconds, better than most distributed-memory MPPs (Massively Parallel Processors, or multicomputers, such as the Intel Paragon or the IBM SP-2/3). Users achieve also sustained, one-way data rates (at the user level) exceeding 1 Gbits/s. There exist direct implementations over Myrinet of many of the standard, distributed-memory-MPP software interfaces, including MPI and PVM. There are also Myrinet implementations for a wide variety of operating systems including NT, the many “dialects” of BSD , Linux, Solaris, OSF1, HP-UX, and the realtime operating systems VxWorks.

Technology and Reliability

Myrinet components are implemented with the same advanced technology – full-custom-VLSI CMOS chips – as today's microprocessors that are used in workstations, PCs, and single-board computers. The use of CMOS technology ensures that Myrinet performance will continue to advance in step with advances in the hosts, without changes to the network architecture and software. These CMOS-based Myrinet components are also extremely reliable, having MTBF in the range of several million hours.

3. Specification of the Data Link Level

This chapter defines the Data Link level specification of Myrinet-on-VME. This level, known as Level 2 of the ISO Reference Model (ISORM) for Open System Interconnection, defines the logical (as opposed to physical) protocol of Myrinet.

3.1 Myrinet Channels, Links, and Ports and Flow Control

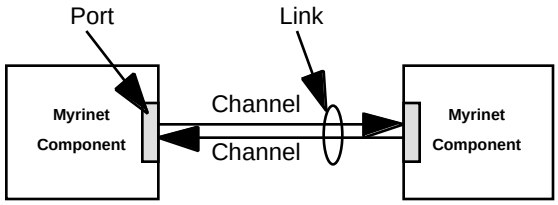

Figure 3.1—1: Myrinet Ports, Channels, and Links

Rule 3.1–1: Myrinet Links

A Myrinet link shall be a full-duplex pair of opposite-going Myrinet channels. A Myrinet channel shall be a unidirectional, point-to-point, communication channel that conveys characters of information. The flow from the sender can be blocked (stopped) temporarily by the receiver at any time, during or between packets, using Flow Control.

Hence, by definition, Myrinet channels appear in links. The existence of the opposite-going channel simplifies automatic network mapping, and could be used at the Physical level to implement flow control and to monitor link continuity.

A Myrinet *Port* is the connection of a link to a component or to a system.

Rule 3.1–2: Character Set

The character set of a Myrinet channel shall include all 256 data bytes and several control symbols.

3.2 Myrinet Flow Control

Rule 3.2–1: Flow Control

The Myrinet Data Link level shall use the data-byte-level flow control, implemented at the Physical level.

Flow control is required for the logical operation of a Myrinet. In addition, it also allows the freedom in the Physical-level implementations for different components and channels within a Myrinet to operate at different data rates.

3.3 Myrinet Packets

Myrinet channels convey packets that are sequences of data bytes. The channel provides the framing of packets, by identifying the first byte (head) of the packet, and the last byte (tail).

Rule 3.3–1: Packet Framing

> The Myrinet Data Link level shall use the packet framing, implemented at the Physical level.

Rule 3.3–2: Packet length (and MTU)

> A Myrinet Data Link level shall convey packets over the channels with any number of data bytes, limited only by an Maximum Transmission Unit (MTU) that shall be at least 4 MBytes.

A byte is the smallest unit of information conveyed on a Myrinet channel.

Myrinet is "byte-wide" at the Data Link level. Independent of whether the Physical-level implementation of a Myrinet channel can carry information as, for example, bit-serial data, byte-parallel data, or multiple bytes in parallel, the Myrinet channel must appear at the Data Link level to be byte-wide. Similarly, software interfaces could restrict data carried from one host to another to multiples of 4 or 8 bytes, but such restrictions appear only at protocol levels higher than the Data Link level.

This specification imposes no limit on the maximum length of a Myrinet packet, but practical system limitations such as timeout[2] periods and buffer sizes limit the packet length.

To be compliant with this Myrinet Data Link specification, the physical layer must support at least packets of 4 MBytes. The software at higher levels can impose more restrictive requirements (e.g., smaller MTU).

Observation 3.3–1: Setting the MTU

> The MTU is not a parameter of Myrinet, but a convention introduced by the application software, above Myrinet. System administrators may set the MTU of their systems to any desired value to meet their special needs. Rule 3.3–2 guarantees that the Myrinet components are capable to handle any MTU of at least 4 MBytes.

[2] See section 3.9 (Timeout and Deadlocks)

3.4 Myrinet Packet Format

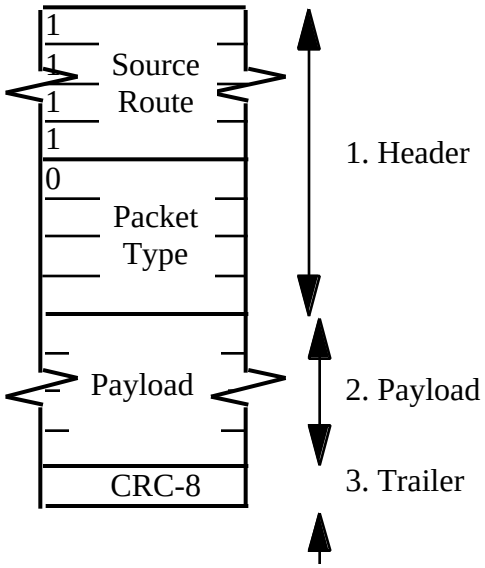

Figure 3.4—1: The Structure of a Myrinet Packet

Rule 3.4–1: Packet Structure

A Myrinet packet shall include in sequence:
1 The header, which shall be four or more bytes.
2 The payload, which shall be zero or more bytes.
3 The trailer, which shall be one byte.

3.4.1 Myrinet Packet Header

Rule 3.4–2: Header Structure

The Myrinet packet header shall include in sequence:
1.1 The source route, which shall be zero or more bytes.
The most-significant bit of each of these bytes shall be 1.
1.2 The packet type, which shall be 4 bytes.
The most-significant bit of the first byte of the packet type shall be 0.

The *source route* is the beginning of a Myrinet packet that is interpreted sequentially by the switches encountered on the route, with each switch "stripping off" (removing) that part of the source route used by that switch. Thus, the source route is sequentially consumed enroute.

Rule 3.4–3: Valid Routing Bytes

For any Myrinet switch, a valid routing byte shall specify an existing port, and also shall have its most significant bit (MSbit) set to 1.

The *packet type* is a field that is used to identify the protocol and suitable software for handling an incoming packet. A Myrinet can carry packets of many types at once, for example, "native" packets used by streamlined application-programming interfaces, and mapping packets used to explore and map the network. The packet type is composed of a two-byte primary type, the most-significant bit of the first of these bytes being 0, followed by a two-byte secondary type. Myricom, Inc., maintains the registry of the primary packet types. Please see the URL: {http://www.myri.com/myri-types.html}. Appendix C lists the assignment of Myrinet packet types, as of July 1998.

Although the header is composed of a variable-length source route followed by a four-byte type , it is always possible to determine from the most significant bits of these bytes where the packet type begins. If all packets followed known routes, this information would be redundant, but mapping packets explore unknown routes. The redundancy in the encoding of the source route in the header of a packet allows a mapping packet routed to a possible host interface to be dropped if it encounters a switch. Similarly, a mapping packet routed to a possible switch can be identified if it encounters a host interface.

3.4.2 Packet Payload

The format and interpretation of the payload part of a Myrinet packet is governed by protocols above the Data Link level. However, it helps in understanding why more information is not required in the Myrinet header to realize that the payload is most often itself a packet that starts with its own header. For example, Myrinet can readily encapsulate Ethernet packets, and thus can carry packets of any protocol supported by Ethernet. The header of the packet carried in the payload can carry information such as the destination address, the packet length, and checksums, which together provide, as required, a degree of error control against data errors or misrouting.

3.4.3 Packet Trailer

Rule 3.4–4: CRC-8

> The last byte of a Myrinet packet shall be originated as the CRC-8 of all the preceding data bytes in that packet, where the CRC-8 is computed using the polynomial X^8+X^2+X+1.

The one-byte trailer of a Myrinet packet contains a cyclic-redundancy-check (CRC-8) character. The CRC-8 detects packets whose data have been corrupted by port-circuit, cable, or connector faults. Packets corrupted by timeout[3] events can additionally be detected at higher protocol levels by the packet being of the wrong length.

The CRC-8 used by Myrinet is the polynomial X^8+X^2+X+1, the same CRC-8 that is used in the header of ATM packets, but without the exclusive-or with hexadecimal 55 (binary 01010101) at the end of the CRC computation. Thus, a packet composed of all zero bytes has a Myrinet-CRC-8 of zero.

[3] See section 3.9 (Timeout and Deadlocks)

Operationally, the CRC-8 is computed on the entire preceding part of the packet, including the header. Because the source-route part of the header is normally modified at each switch, the CRC-8 must be recomputed and checked for each link. At a switch or at an interface, the port circuits compute the CRC-8 of the incoming packet, and substitute the exclusive-or of the computed and received CRC-8 in the trailer of the packet. A received packet whose CRC-8 is correct will have a zero trailer at this point, whereas a packet whose CRC-8 is incorrect will have a non-zero trailer. On the outgoing link, the port circuits compute the CRC-8 as the packet is sent, and exclusive-or the computed CRC-8 into the trailer. If the packet had a correct CRC-8 when received, it will have a correct CRC-8 when sent; whereas if the packet had an incorrect CRC-8 when received, it will have an incorrect CRC-8 when sent.

3.5 Myrinet Components

A Myrinet interface to or within a host computer nominally has one port. A host may include more than one Myrinet interface, but from the network's viewpoint these ports operate independently. The ports of Myrinet interfaces are the only points where new packets are injected into a Myrinet network, and the only points at which they are properly consumed.

A Myrinet switch is a multiple-port component that switches (routes) packets entering on the incoming channel of a port to the outgoing channel of a port selected by a source route in the initial bytes of the packet. A very brief description of Myrinet switching and routing is included in Appendix B.

There is a general rule related to technicalities of "progress" that applies to Myrinet components:

Rule 3.5–1: Progress Guarantee

A Myrinet port shall not block progress indefinitely.

A packet directed to the incoming port of a Myrinet interface must be consumed eventually. Although the interface can block an incoming packet for some interval, for example, while it allocates a new receive buffer, such blocking must be minimized. Similarly, a Myrinet switch must eventually send out on the specified port any packet that it has received, or drop it if that port is blocked for too long (i.e., longer than the timeout[4] period).

Observation 3.5–1: Progress guarantee

The primary mechanism guaranteeing progress and packet consumption is the insistence on consumption at the endpoints of the network, together with the routing being deadlock-free.

If deadlocks occur upon initialization or due to hardware or software errors, the timeout is the fallback mechanism to clear them out.

[4] See section 3.9 (Timeout and Deadlocks)

For any Myrinet switch, a valid routing byte must specify an existing port, and also must have its most significant bit (MSbit) set to 1.

Observation 3.5–2: Dropped Packets

A packet is dropped if it is received and ignored (e.g., not being transmitted by a switch). A dropped packet does not block other packets.

Rule 3.5–2: Invalid Routing

A Myrinet switch shall drop all packets that do not start with a valid routing byte.

Invalid routing bytes can result from software or hardware malfunction, or from unsuccessful explorations by a dynamic mapping process.

A switch drops packets by accepting them without sending them.

3.6 Myrinet Source Routes and switching

Myrinet switching is not specified at the Data Link level. Indeed, the only rule that could be stated is the rule against a rule:

Rule 3.6–1: Packet Routing by Switches

The Data Link-level specification shall not restrict the method of packet routing performed in a Myrinet switch.

Rule 3.6–2: Source Routing

Myrinet switches shall steer packets according to the valid routing bytes at the beginning of the packets, and modify (remove or replace) these bytes.

Observation 3.6–1: Source Routing

By removing routing information from the head of packets, the head is an "iterative" source route.

However, because the design rationale for Myrinet cannot be appreciated without understanding the existing practice in Myrinet routing, Appendix B of this specification offers a brief exposition of that practice.

3.7 Myrinet Unused / Disconnected Ports

.

Rule 3.7–1: Disconnected Ports

An outgoing packet on a port that is unused, disconnected, or connected through a link to a powered-off component shall not be blocked, but shall be dropped.

A Myrinet must continue to operate under conditions in which some of its ports are unused, some of its links are disconnected, or some of its links are connected to unpowered components. The outgoing packet on a port that is in any of these conditions must not be blocked, or else it would block other links and ports. Instead, outgoing packets on such ports are dropped.

This Data Link level requirement must be implemented at the Physical level. Port circuits can employ signaling such that blocking is disabled on a link that is open or connected to a powered-off component. Alternatively, port circuits can employ methods such as carrier detection or a "heartbeat" protocol to monitor link continuity. In high-availability systems, loss of link continuity or dropped packets can be reported through a monitoring network.

Recommendation 3.7–1: Self-Monitoring

The Physical layer should continuously monitor the continuity of the links, and issue alerts when discontinuities are discovered.

Sections 5.1 and 6.1 suggest the use of an optional control symbol, Beat, for that purpose.

3.8 Myrinet Topology

The network topology can be viewed as an undirected graph. Any way of linking together interfaces and switches is allowed. The graph can contain cycles; indeed, topologies with cycles are required to provide multiple-path redundancy. The physical network can include unpowered host interfaces and unused switch ports.

3.9 Timeout and Deadlocks

Under certain error conditions a Myrinet is capable of deadlock, a condition in which one or more packets cannot make progress because they are mutually blocked. It is generally the responsibility of the software that establishes sets of routes between hosts to assure that the set of routes is deadlock-free, and it is the responsibility of the hardware to clear any deadlock, if it occurs.

In the event of a deadlock caused by a data error in a packet header or by a software error, it is necessary to clear the deadlock by dropping packets that are contributing to the deadlock.

Rule 3.9–1: Timeout

If a packet is not terminated for more than a given timeout period since it began, then the port circuits shall terminate and/or drop any such packet.

This rule applies to both sending and receiving ports. Lack of packet termination can result from reasons such as blocking (flow control) or from hardware malfunction.

It is not intended that the timeout be used or depended upon in the normal operation of a Myrinet. Rather, because a Myrinet can be distributed across a large area, this timeout mechanism is regarded as a more practical way to reinitialize a Myrinet than distributing a system-reset signal.

Recommendation 3.9–1: Timeout Period

The timeout period should be on the order of a second.

Observation 3.9–1: Timeout Period

This timeout period is expected to be long enough to allow transmission times that manyfold exceeds the MTU to allow for queuing that can occur as a result of "hot spots".

The use of Timeout may significantly vary between installations. Fast-response systems (e.g., military fire control) are expected to require very short timeout periods, whereas non realtime systems may require long periods. Other systems may use yet other mechanisms to monitor progress ignoring the timeout all together.

Suggestion 3.9-1: Setting the Timeout Period

Designers are encouraged to make it possible to set the timeout period in the field (e.g., by jumpers).

4. The Requirements for the Physical Level

This chapter summarizes the requirements imposed by the Data Link level on the Physical level, and also defines various rates for Myrinet links, regardless of their physical implementations.

A summary of Myrinet requirements from any Physical-level implementation:

- The Physical level channels must convey characters, including all 256 data bytes, and several control symbols.
- The Physical level must provide full-duplex point-to-point links with flow control and packet framing.
- A single byte must be the smallest unit of information conveyed on Myrinet channel. If the physical layer is wider than a byte, padding insertion and removal (if needed) must be provided.
- The MTU of the Physical level must be at least 4 MBytes.
- Packets that require transmission time beyond the timeout must be properly terminated
- (i.e., all bytes are ignored until the next Gap, and a Gap is sent).
- The Physical level must direct packets according to their source route, and modify it.
- The Physical level of a Myrinet channel must provide a CRC-8 packet protection in the packet trailer.

Because all Physical-level implementations must support the above requirements, they differ only in incidental details, not in essential properties. Therefore, it is easy to translate between different Physical implementations using conversion circuits.

4.1 Myrinet Link Rates

Observation 4.1–1: Myrinet-RRR/PHY Designations

The designation "Myrinet-RRR" specifies any Myrinet implementation that transfers characters at the RRR rate, measured in Mbits/second, regardless of the actual modulation frequency of the links or of the encoding.

The "PHY" defines the specific Physical implementation, such as:

- "SAN" for short cables used on the front and the rear panels, and for PCB traces;
- "BP" for "SAN" over the P0/J0 or P4/J4, used for interconnection with the backplane;
- "LAN" for the longer cables that are used on the front and the rear panels.

4.1.1 640 Mbits/sec Rate

Rule 4.1–1: Nominal Character Period for Myrinet-640

The Nominal character period of Myrinet-640 shall be 12.5 nsec.

4.1.2 1280 Mbits/sec Rate

Rule 4.1–2: Nominal Character Period for Myrinet-1280

The Nominal character period of Myrinet-1280 shall be 6.25 nsec.

4.1.3 2560 Mbits/sec Rate

Rule 4.1–3: Nominal Character Period for Myrinet-2560

The Nominal character period of Myrinet-2560 shall be 3.125 nsec.

These rates are defined uniformly for all the Physical-level implementations (e.g., the same definition of the character period holds both for Myrinet-1280/SAN and for Myrinet-1280/LAN).

Chapter 5 defines Myrinet-1280/SAN, Myrinet-1280/BP, Myrinet-2560/SAN, and Myrinet-2560/BP.

Chapter 6 defines Myrinet-640/LAN and Myrinet-1280/LAN.

5. Specification of the SAN Physical Level

This chapter specifies the SAN (System-Area-Network) Physical-level implementation of Myrinet-on-VME. This level, known as Level 1 of the ISORM, defines the physical (electrical and mechanical, as opposed to logical) protocol of Myrinet.

SAN communication is intended for intra-cluster distances, such as intra-card, intra-subrack, and intra- cabinet. For inter-cluster distances LAN are intended to be used.

This chapter specifies, for SAN Physical implementations, the following:
(5.1) The channel character set (data bytes and control symbols)
(5.2) The encoding of the characters on signals
(5.3) The electrical characteristics of these signals (such as timing and voltage)
(5.4) Pin-assignments, connectors, and cables.

Section (5.4) includes the specifications for front-panel and rear-panel applications using SAN connectors and SAN cables, and also for backplane applications using P0 or P4 connectors.

The designation of the backplane implementation is "BP", which is SAN using the P0/J0 or the P4/J4 connectors. Hence, every specification of SAN (except the details of the SAN connectors) applies also to BP.

5.1 The SAN Channel Character Set and Flow Control

Rule 5.1–1: The SAN character set

The SAN character set shall include, in addition to all 256 data bytes, at least the following control symbols: Gap, Res1, and Res2.

The Gap symbol is used for packet framing. Res1 and Res2 are reserved for future use, such as for interoperability with future versions of Myrinet as they evolve with time.

Rule 5.1–2: Gap Symbol

The Gap symbol shall indicate that the previous data byte was the trailer of the previous packet, and that the next data byte (i.e., not a control symbol) will be the first byte of the header of the next packet.

Rule 5.1–3: Asynchronous Control Symbols

The Gap symbol shall be sent only between packets. All other control symbols can be sent either between or inside packets.

Permission 5.1–1: Repeated Gap Symbols

Successive Gap symbols may be sent repeatedly between packets.

Recommendation 5.1–1: Beat Symbol (Continuity Monitoring)

The Beat symbol should be used to monitor the continuity of links. Beat should be sent at a the frequency defined below. If no Beat arrives within a given period an alert should be generated regarding that link.

Rule 5.1–4: Beat Frequency

If the Beat symbol is used, it shall be sent every 10 microseconds (±10%). The receiver's timeout period is left for implementer's discretion.

The flow control is handled on SAN channels by using the B-bit (the "Block signal"), as described in section 5.2.

Rule 5.1–5: Flow Control

Upon receiving an asserted IB signal, the output channel of a port shall stop its transmission of data bytes until the IB is unasserted.

Observation 5.1–1: Flow Control Applicability

The flow control applies only to data bytes, not to control symbols.

Observation 5.1–2: Bytes in Flight

From the time a component issues a Block signal (by asserting its OB) to block the flow of incoming bytes, until the flow actually stops, more bytes may still arrive. The number of these bytes is N=2LR/c'+K, where L is the length of cable, R the transmission rate, c' the speed of electronic propagation over this cable (typically c'=0.6c=180Mm/s in copper), and K is the number of bytes transmitted since the IB arrives until the flow actually stops.

For example, a copper cable of L=3m, at the rate of R=160MB/s, with K=5B can still deliver after a Block signal is asserted, at least:

$$N = 2*3m*160M(B/sec) / (180Mm/sec) + 5B = 11B.$$

Rule 5.1–6: Cable Slack Buffers to Prevent Data Loss

Components that receive data from cables shall have enough slack-buffer memory to absorb the data that is already in flight when a Block signal is issued, thus preventing data loss. This amount compensates for the cable length and for the time it takes senders to block the flow after receiving the Block signal.

Recommendation 5.1–2: Cable Slack Buffers to Prevent Data Starvation

In order to prevent potential "data starvation" by receivers after unblocking the data flow, the size of the slack buffers should be at least twice of what is needed just to prevent data starvation.

5.2 The Encoding of the SAN Characters on Signals

The SAN physical layer for a Myrinet link is comprised of 20 signals organized in two independent channels of 10 signals for communication in each direction.

Rule 5.2–1: SAN signals

The organization for each SAN channel is:
8 data bits (0 through 7)
1 Data/Control bit (D: 1 = Data, 0 = Control)
1 Flow Control "Block" bit (B: 1 = stop, 0 = go)

Rule 5.2–2: Encoding of the SAN Character Set

The encoding of the SAN character set shall be as shown in Table 5–1.

	Bit:	D	7	6	5	4	3	2	1	0
Mandatory	Data Byte	1	d	d	d	d	d	d	d	d
Mandatory	Gap Symbol	0	0	0	0	0	1	1	0	0
Mandatory	Res1 Symbol	0	0	0	0	0	0	0	1	1
Mandatory	Res2 Symbol	0	0	0	0	0	1	1	1	1
Optional	Beat Symbol	0	1	1	1	1	1	1	1	1

Table 5–1: Encoding of SAN Characters

The signals are denoted by either "I*" (inputs from the Myrinet into the component) or "O*" (outputs to the Myrinet from the component) for * being any of {0,1,2,3,4,5,6,7,D,B}. The SAN cables connect each O* to the corresponding I*. Hence, the O* of a port is the I* of the port with which it is connected.

The direction of B, the flow-control signal, is opposite to that of all other signals. It is an output of the component for which the rest of the channel is an input.

Hence, the output channel consists of {O0,O1,O2,O3,O4,O5,O6,O7,OD,IB} and the input channel consists of {I0,I1,I2,I3,I4,I5,I6,I7,ID,OB}.

OB and IB are used for flow control. IB is an input to the sending part of a port, and commands the port to stop sending. Similarly, OB is an output from the receiving part of a port, and signals the sending part of the port on the other side of the link to stop sending.

Recommendation 5.2–1: Unknown Characters

Characters that are not defined by Rule 5.2–2 should be ignored or treated as errors.

Rule 5.2–3: SAN NRZI Encoding

The 8 data signals and the D signal shall be NRZI (Non-Return-to-Zero, Invert, or "transition") encoded, such that a transition on a line represents a 1 and no transition represents a 0.

The B signal is level encoded (not NRZI).

A Myrinet receiver operates asynchronously, and accepts characters whenever they arrive; however, they cannot be separated in time by less than the character period.

All characters include at least one transition. Myrinet receivers group together the transitions of multiple signals within a certain amount of time ("window") after the first transition is detected on any signal, and treat all of them together as a single character.

Observation 5.2–1: Signal frequency

This encoding keeps the maximum fundamental frequency on any line from exceeding half of the data rate. For example, while transferring data at 160Mbits/s per signal line, no signal exceeds 80MHz in the fundamental frequency.

Rule 5.2–4: Skew

The total skew (difference in arrival time from earliest to latest transition) from sender to receiver shall be less than 40% the character period.

Recommendation 5.2–2: Skew

To maximize timing margins and improve bit error rate, the total skew of the PCB traces, connectors, and cable should be made as small as possible.

Permission 5.2–1: Skew reduction

Adaptive techniques may be used to reduce the effective skew at the receiving port.

Recommendation 5.2–3: Cable Skew

The skew of any cable should not exceed one quarter of the character period.

5.3 The Characteristics of SAN Signals

5.3.1 Packet Timing

Rule 5.3–1: IPG (Inter-Packet Gap)

Myrinet receivers shall be able to handle packets with an Inter-Packet Gap of (minimum) one Gap symbol.

5.3.2 SAN Signal Characteristics

Rule 5.3–2: SAN Signal characteristics

The SAN signals shall have controlled edge rate and shall have the characteristic specified in Table 5–2.

	Level	Character Period	Switching time Min	Switching time Max	Source Termination	Type
Myrinet-1280	1.25 V	6.250 ns	0.5 ns	2.0 ns	50 Ω	Single Ended
Myrinet-2560	1.25 V	3.125 ns	0.2 ns	1.0 ns	50 Ω	Single Ended

Table 5–2: SAN Signaling

Recommendation 5.3–1: Pulldown for Input Signals

All SAN input signals (I*) should have high impedance pulldown of at least 5 KΩ (20 KΩ nominal).

These pulldowns maintain floating inputs at a voltage sufficiently far from the switching threshold to avoid noise inputs from unused ports or disconnected cables.

Rule 5.3–3: Common GND for SAN Signaling

Myrinet components that use SAN links shall have common GND.

Recommendation 5.3–2: Common GND for SAN Signaling

The best way to achieve common ground is by using the same power+GND source. If this is not the case, ground straps should be used, but this still does not guarantee common GND, e.g., between "floating chasses".

5.4 SAN Connectors, Cables, and Pin-Assignments

This section specifies low-level details of the connectors, and recommends certain connectors and cables. SAN connection through the front and/or the rear panels must be as specified in section 5.4.1, whereas SAN connection through the backplane must use either the P0 or the P4 connectors, as specified in section 5.4.2.

Rule 5.4–1: Connector Gender

Myrinet components shall have receptacle (female) connectors. Myrinet cables shall have plug (male) connectors.

5.4.1 Front-Panel and/or Rear-Panel Connection (SAN)

This section describes the SAN connection scheme. Items covered are the SAN connectors and their locations, and the basic SAN pinout.

Rule 5.4–2: SAN Connectors

For applications when the SAN connector are provided on the front panel of the VME board, the board mounted SAN connector should be a 40-pin Microstrip receptacle [5].

Recommendation 5.4–1: Positioning of SAN Connectors.

The SAN connector should be positioned as close as possible to the board edge and the front panel should be designed to allow for the latching cable connector.

Observation 5.4–1: Dual Link SAN Connectors

Since the SAN connector has 40 pins, and since a SAN link uses only 20 pins, each SAN connector carries 2 full-duplex links, designated as A and B. The A link uses the 20 middle pins (with the output pins 11 to 20 of each device connected to the input pins 30 to 21, respectively, of the other device; and the B link uses the 20 outside pins (with the output pins 1 to 10 connected to the input pins 40 to 31).

Rule 5.4–3: The A Link is the Primary Link

The A link shall be considered as the primary link of the connection. Components that use one link only shall use the A link.

[5] The board mounted connector should be the 40-Pin right-angle Microstrip receptacle, AMP part number 536295-1, or equivalent.

Rule 5.4–4: Unused SAN Links

When only one link is used by Myrinet, the other link (on dual link connectors) shall not be used for other purposes.

Rule 5.4–5: SAN Pin Assignment

The pin assignments in the board mounted SAN connector shall be as shown in Table 5–3. The signals S* (for * being any of 0,1,2,3,4,5,6,7,D,B) are the output signals of the B link, similar to O* of the A link. The R* are the inputs of the B link, similar to the I* of the A link.

Sending Channels			Receiving Channels		
Pin #	Signal Name	Link	Link	Signal Name	Pin #
1	S0	B	B	R0	40
2	S1	B	B	R1	39
3	S2	B	B	R2	38
4	S3	B	B	R3	37
5	S4	B	B	R4	36
6	S5	B	B	R5	35
7	S6	B	B	R6	34
8	S7	B	B	R7	33
9	SD	B	B	RD	32
10	SB	B	B	RB	31
11	O0	A	A	I0	30
12	O1	A	A	I1	29
13	O2	A	A	I2	28
14	O3	A	A	I3	27
15	O4	A	A	I4	26
16	O5	A	A	I5	25
17	O6	A	A	I6	24
18	O7	A	A	I7	23
19	OD	A	A	ID	22
20	OB	A	A	IB	21

Table 5–3: Dual Link SAN Connector Pinout

5.4.1.1 SAN Cables

Rule 5.4–6: SAN Cables

The SAN cables shall have 40-pin Microstrip plugs[6] that mate with the board mounted receptacles, and shall connect the signals defined in Rule 5.4–5 such that they connect each Ox signal of one component with the Ix of the other component, and each Sx to the other Rx.

Observation 5.4–2: SAN Cable Connectivity

Each connector pin#N is connected with connector pin#(41-N) at the other cable end (e.g., pin 1 of each connector to pin 40 of the other connector, pin 2 to pin 39, pin 3 to pin 38 and so on).

Each row of Table 5–3 indicates the pins (of the opposite ends) that are interconnected by the SAN cable.

Rule 5.4–7: SAN Cables for Myrinet-1280

Myrinet-1280/SAN shall be used only over SAN cables not longer than 10ft.

Rule 5.4–8: SAN Cables for Myrinet-2560

Myrinet-2560/SAN shall be used only over SAN cables not longer than 10ft.

Rule 5.4–9: SAN Cables – Impedance

The SAN cables shall have 50 Ω controlled impedance.

5.4.1.2 VME Front and/or Rear-Panel Access

A SAN cable connects to the SAN connector on the VME board through the Front Panel or boards that are on the bulkhead. The board-mounted connector is specified in Rule 5.4–6. Since connectivity is through a cable, the connector on the boards can be located anywhere on the Front Plate that the board designer chooses.

Permission 5.4–1: Right-side-up Connectors

The connector may be placed "right side up" for VME board mounting.

[6] The SAN cables should be 40-signal Microstrip cables, AMP part number 636349 (or equivalent).

Permission 5.4–2: Right-side-down Connectors

The connector may be placed "upside down" for mounting on a mezzanine board such as a PMC.

Permission 5.4–3: SAN Connector Location

The SAN connector may be placed any place along the edge of the VME board.

5.4.2 Backplane Connection (P0/J0 and/or P4/J4)

This section specifies the Myrinet connection through the backplane, using either P0/J0 or P4/J4.

Observation 5.4–3: Myrinet-RRR/BP

The Backplane Connection (via P0/J0 and/or P4/J4) has the Physical-level designation of "BP". Except for the mechanical details of the connectors Myrinet-RRR/BP is identical to Myrinet-RRR/SAN.

Rule 5.4–10: Conformity with VME64x

The P0/J0 connector shall conform to VITA 1.1-1997, VME64x, chapter 4, and the P4/J4 connector shall conform to ANSI/VITA 1.3-1997, VME64x 9U x 400 mm Format, chapter 4.

Observation 5.4–4: Caution on P0 and P4 Connectors

VME-6U and VME64x-9U boards using a P0 connector (not specified in this document) could conflict with VME backplanes that have a mechanical structure member between J1 and J2 connectors.

Similarly, VME64x-9U boards using a P4 connector (not specified in this document) could conflict with VME backplanes that have a mechanical structure member between J2 and J3 (or between J2 and J5) connectors.

Rule 5.4–11: Pin Assignment for the Backplane Connectors

The pin assignments for the P0 connector shall be as shown in Table 5–4.

Pos #	Row A	Row B	Row C	Row D	Row E
1	Res	GND	Res	GND	Res
2	GND	S0	GND	S1	GND
3	S2	GND	S3	GND	S4
4	GND	S5	GND	S6	GND
5	S7	GND	SD	GND	SB
6	GND	O0	GND	O1	GND
7	O2	GND	O3	GND	O4
8	GND	O5	GND	O6	GND
9	O7	GND	OD	GND	OB
10	GND	Res	GND	Res	GND
11	IB	GND	ID	GND	I7
12	GND	I6	GND	I5	GND
13	I4	GND	I3	GND	I2
14	GND	I1	GND	I0	GND
15	RB	GND	RD	GND	R7
16	GND	R6	GND	R5	GND
17	R4	GND	R3	GND	R2
18	GND	R1	GND	R0	GND
19	Res	GND	Res	GND	Res

Table 5–4: Backplane Connector Pinout

The signals S* (for * being any of 0,1,2,3,4,5,6,7,D,B) are the output signals of the B link, similar to O* of the A link. The R* are the inputs of the B link, similar to the I* of the A link. The “Res” indicate a reserved signal for future use.

Observation 5.4–5: P0 has 7 rows (copied from VME64x Observation 4.6)

The J0/RJ0 connectors have seven physical rows of contacts, with the z and f contact rows connected to the backplane's ground plane. On the VME64x board, there is no z row of contact holes. Depending on the connector design, the P0 and the RP0 connector ground contacts in the z and f row of the connector (which is on the connector shroud) will alternately connect to the board's f row of ground contacts.

6. Specification of the LAN Physical Level

This chapter specifies the LAN (Local-Area-Network) Physical-level implementation of Myrinet-on-VME. This level, known as Level 1 of the ISORM, defines the physical (electrical and mechanical, as opposed to logical) protocol of Myrinet.

LAN communication is intended for distances of up to a few tens of meters. For distances up to 3 meter (typical for intra-cabinet and between colocated cabinets) SAN can be used.

This chapter specifies, for LAN Physical implementations, the following:
(6.1) The channel character set (data bytes and control symbols)
(6.2) The encoding of the characters on signals
(6.3) The electrical characteristics of these signals (such as timing and voltage)
(6.4) Pin-assignments, connectors, and cables.

Section (6.4) includes the specifications for front-panel and rear-panel applications using LAN connectors and LAN cables. Unlike SAN, LAN cannot be used for backplane applications that use P0 or P4 connectors.

6.1 The LAN Channel Character Set and Flow Control

Rule 6.1–1: The LAN character set

The LAN character set shall include (in addition to all 256 data bytes) at least the following control symbols: Gap, Stop, and Go.

The Gap symbol is used for packet framing. The Stop and Go symbols are used for flow control.

Rule 6.1–2: Gap Symbol

The Gap symbol shall indicate that the previous data byte was the trailer of the previous packet, and that the next data byte will be the first byte of the header of the next packet.

Rule 6.1–3: Asynchronous Control Symbols

The Gap symbol shall be sent only between packets. All other control symbols can be sent either between or inside packets.

Permission 6.1–1: Repeated Gap Symbols

Successive Gap symbols may be sent repeatedly between packets.

Recommendation 6.1–1: Beat Symbol (Continuity Monitoring)

The Beat symbol should be used to monitor the continuity of links. Beat should be sent at a the frequency defined below. If no Beat arrives within a given period an alert should be generated regarding that link

Rule 6.1–4: Beat Frequency

If the Beat symbol is used, it shall be sent every 10 microseconds (±10%). The receiver's timeout period is left for implementer's discretion.

Rule 6.1–5: Flow Control (Stop and Go Symbols)

Upon receiving the Stop symbol, the output channel of a port shall stop its transmission of data bytes until a Go symbol is received.

Observation 6.1–1: Flow Control on Opposite Channel

The flow-control symbols that control the flow from a source to a destination are inserted in the opposite-going channel, of that link, from the destination to the source.

Observation 6.1–2: Flow Control Applicability

The flow control applies only to data bytes, not to control symbols.

Permission 6.1–2: Repeated Flow Control Symbols

Successive Flow Control symbols (Stop or Go, whichever is applicable at the time) may be sent repeatedly when data bytes are stopped, and between packets.

Observation 6.1–3: Bytes in Flight

From the time a component sends the Stop symbol to block the flow of incoming bytes until the flow actually stops, more bytes can still arrive. The number of these bytes is N=2LR/c'+K, where L is the length of cable, R the transmission rate, c' the speed of electronic propagation over this cable (typically c'=0.6c=180Mm/s in copper), and K is the number of bytes transmitted since the IB arrives until the flow actually stops.

For example, a copper cable of L=10m, at the rate of R=160MB/s, with K=5B can still deliver after a Stop symbol is sent, at least:

N = 2*10m*160M(B/sec) / (180Mm/sec) + 5B = 23B.

Rule 6.1–6: Cable Slack Buffers to Prevent Data Loss

Components that receive data from cables should have enough slack-buffer memory to absorb the data that is already in flight when a Block signal is asserted. This amount should compensate for the cable length and for the time it takes senders to block the flow after receiving the Stop symbol. This amount of slack buffers is needed to prevent data loss.

Recommendation 6.1–2: Cable Slack Buffers to Prevent Data Starvation

In order to prevent potential "data starvation" by receivers after unblocking the data flow, the size of the slack buffers should be at least twice of what is needed just to prevent data starvation.

6.2 The Encoding of the LAN Characters on Signals

The LAN physical layer for a Myrinet link is comprised of 18 signals organized in two independent channels of 9 signals for communication in each direction.

Rule 6.2–1: LAN signals

The organization for each LAN channel is:
8 data bits (0 through 7)
1 Data/Control bit (D: 1 = Data, 0 = Control)

Rule 6.2–2: Encoding of the LAN Character set

The encoding of the LAN character set shall be as shown in Table 6–1.

	Bit:	D	7	6	5	4	3	2	1	0
Mandatory	Data Byte	1	d	d	d	d	d	d	d	d
Mandatory	Gap Symbol	0	0	0	0	0	1	1	0	0
Mandatory	Go Symbol	0	0	0	0	0	0	0	1	1
Mandatory	Stop Symbol	0	0	0	0	0	1	1	1	1
Optional	Beat Symbol	0	1	1	1	1	1	1	1	1

Table 6–1: Encoding of LAN Characters

The signals are denoted by either "I*" (inputs from the Myrinet into the component) or "O*" (outputs to the Myrinet from the component) where * being any of {0,1,2,3,4,5,6,7,D}. The LAN cables connect each O* to the corresponding I* at the other end. Hence, the O* of a port, is connected with the I* of the port at the other end.

Hence, the output channel consists of {O0,O1,O2,O3,O4,O5,O6,O7,OD} and the input channel consists of {I0,I1,I2,I3,I4,I5,I6,I7,ID}.

Recommendation 6.2–1: Unknown Characters

Characters that are not defined by Rule 6.2–2 should be ignored or treated as errors.

Rule 6.2–3: LAN Differential Signaling

The 8 data signals and the D signal shall be differentially transmitted over a twisted-pair of wires.

Rule 6.2–4: NRZI Signaling

All the LAN signals shall be NRZI (Non-Return-to-Zero, Invert, or "transition") encoded, such that a transition on a line represents a 1 and no transition represents a 0.

A Myrinet receiver operates asynchronously, and accepts characters whenever they arrive; however, they cannot be separated in time by less than the character period.

All characters include at least one transition. Myrinet receivers group together the transitions of multiple signals within a certain amount of time ("window") after the first transition is detected on any signal, and treat all of them together as a single character.

Observation 6.2–1: Signal frequency

This encoding keeps the maximum frequency on any line from exceeding half of the data rate. For example, while transferring data at 160Mbit/s per line, no signal exceeds 80MHz, in the fundamental frequency.

Rule 6.2–5: Skew

The total skew (difference in arrival time from earliest to latest transition) from sender to receiver shall be less than 40% the character period.

Recommendation 6.2–2: Skew

To maximize timing margins and improve bit error rate, the total skew of the PCB traces, connectors, and cable should be made as small as possible.

Permission 6.2–1: Skew reduction

Adaptive techniques may be used to reduce the effective skew at the receiving port.

Recommendation 6.2–3: Cable Skew

The skew of any cable should not exceed one quarter of the character period.

6.3 The Characteristics of LAN Signals

6.3.1 Packet Timing

Rule 6.3–1: IPG (Inter-Packet Gap)

Myrinet receivers shall be able to handle packets with an Inter-Packet Gap of one Gap symbol only.

6.3.2 LAN Signal Characteristics

Rule 6.3–2: LAN Signal characteristics

The LAN signals shall have controlled edge rate and shall have the characteristic specified in Table 6–2.

	Level	Character Period	Switching time Min	Switching time Max	Source Termination	Type
Myrinet-640	1.2+/-0.4 V	12.50 ns	0.8 ns	4.0 ns	100 Ω	differential
Myrinet-1280	1.2+/-0.4 V	6.25 ns	0.5 ns	2.0 ns	100 Ω	differential

Table 6–2: LAN Signaling

Recommendation 6.3–1: Pulldown for Input Signals

All LAN input signals (I*) should have high impedance pulldown of at least 10KΩ.

Rule 6.3–3: Common GND for LAN Signaling

Myrinet components that use LAN links shall have GND levels that agree within the common-mode range.

6.4 LAN Connectors, Cables, and Pin-Assignments

This section specifies low-level details of the connectors, and recommends certain connectors and cables. LAN is used only to enter/egress subracks on the front or on the rear panels.

Rule 6.4–1: Connector Gender

Myrinet components shall have receptacle (female) connectors. Myrinet cables shall have plug (male) connectors.

6.4.1 Front-Panel and/or Rear-Panel Connectors (LAN)

Rule 6.4–2: LAN Connectors

The standard cable-connector for Myrinet ports shall be a 37-pin D-Subminiature Connector (DB-37). The component-mounted connectors shall be receptacle (female), and the cable-end connectors shall be plug (male).

In addition to these 37 pins, the cable's shield with its shell shall be connected to GND at both ends. When unshielded cables (e.g., "twist-n-flat" ribbon cables) are used, at least one wire shall carry GND between the end shells.

Rule 6.4–3: LAN Pin Assignment

The pin assignments in the connectors shall be as shown in Table 6–3.

A		B		C		D	
pin#	Name	pin#	Name	pin#	Name	pin#	Name
1	Rd+	19	Sd+	20	Rd-	37	Sd-
2	S0+	18	R0+	21	S0-	36	R0-
3	S1+	17	R1+	22	S1-	35	R1-
4	R7+	16	S7+	23	R7-	34	S7-
5	R6+	15	s6+	24	R6-	33	S6-
6	S2+	14	R2+	25	S2-	32	R2-
7	S3+	13	R3+	26	S3-	31	R3-
8	R5+	12	S5+	27	R5-	30	S5-
9	R4+	11	S4+	28	R4-	29	S4-
The cable's Pin#10 must be grounded if it is longer than 10m							

Table 6–3: LAN Connector Pinout

Each signal is encoded differentially between pins that are in the same row, either of columns A–C or of columns B–D.

Rule 6.4–4: LAN Cable Connectivity

The LAN cable shall connect each Sx to its corresponding Rx, as defined in Rule 6.4–3. This connects each pin from column A with the same-row pin from column B (i.e., pin 1 with pin 19, 2 with 18,..., and 9 to 11), and similarly pins from column C with same-row pins from column D (i.e., 20 with 37, 21 with 36, ..., and 28 with 29).

Recommendation 6.4–1: Twisted Pairs

The LAN cables should consist of at least 18 twisted pairs, a single wire for pin#10, and an electrical connection of the shells (by a cable shield or by a wire) .

Permission 6.4–1: LAN Connector Location

The LAN connector may be placed any place along the front panel of the VME board.

Rule 6.4–5: LAN Cables

The LAN cables shall have connectors that mate with the connector specified in Rule 6.4–2 such that they connect each Sx signal with its corresponding Rx at the other end. The LAN cables shall have 100Ω controlled impedance.

Links between Myrinet components could employ a variety of cable types in the 100Ω to 110 Ω range of characteristic impedance. Depending upon the EMI requirements, either unshielded or shielded cable may be used.

Rule 6.4–6: Pin #10 of the DB-37 Connector

LAN cables that are longer than 10 meters shall have pin#10 (on their connectors), connected to GND at both ends.

Rule 6.4–7: LAN Cables for Myrinet-640

Myrinet-640/LAN shall be used only over LAN cables that are not longer than 25 meters, and have pin #10 connected to the shell/shield GND.

Rule 6.4–8: LAN Cables for Myrinet-1280

Myrinet-1280/LAN shall be used only over LAN cables that are not longer than 10 meters, and have pin#10 floating (i.e., not connected to anything).

Rule 6.4–9: Auto-Speed for LAN ports

A Myrinet port shall send at the Myrinet-640 rate if pin#10 of the LAN cable is connected to GND, and shall send at the Myrinet-1280 rate if it is not connected to GND (indicating that the cable is not longer than 10 meters).

7. Appendix A: Examples of Myrinet-on-VME

7.1 Examples

The following figure demonstrates 4 SBCs interconnected over the front panel with a switching board (shown in the middle) providing internal and external connectivity.

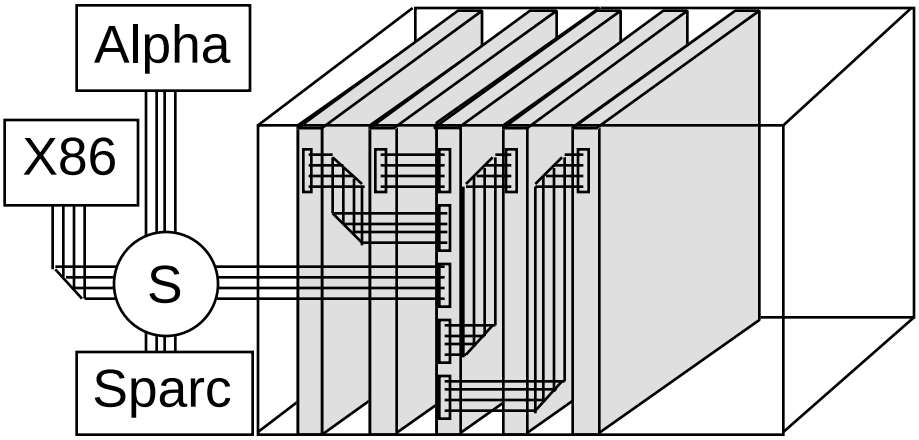

Figure 7.1—1: Front Panel Interconnection of 4 SBCs

Following is an example of a backplane overlay that connects 4 P0 connectors (with a total of 8 links) to 4 SAN connectors, using two 8x8 switches to provide both internal communication and flexible communication with the outside, in several possibilities ranging from using four SAN cables to using one SAN cable for accessing the four boards (that are behind these P0 connectors).

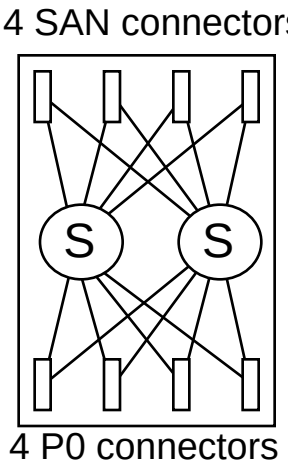

Figure 7.1—2: A Backplane Overlay (with Myrinet/BP and Myrinet/SAN)

The following figure shows how 4 such overlays may be used for interconnecting 16 boards with the full bisection bandwidth (of 10.28Gigabit/sec for Myrinet-1280/BP).

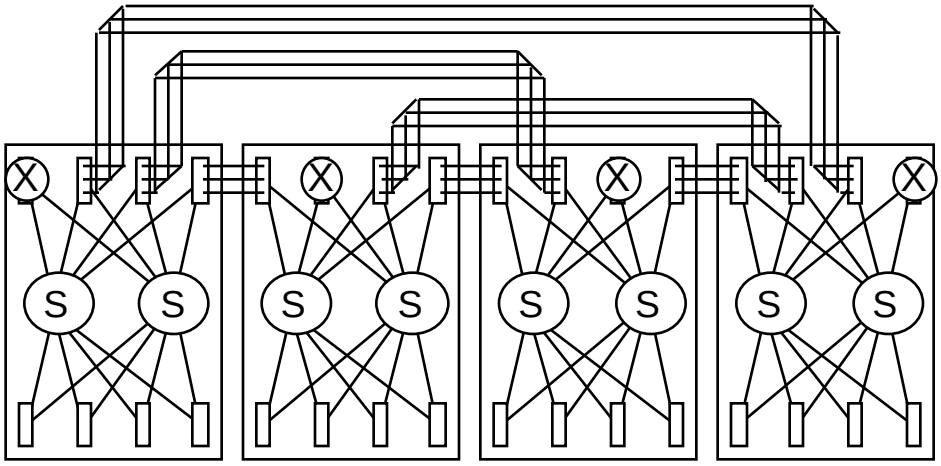

Figure 7.1—3: Full-Bisection-Bandwidth 16-Board Configuration

The following figure shows a subrack with 16 processing boards, each with an on-board Myrinet connecting 4 processing nodes (such as microprocessors and FPGA nodes) forming together a 64-processor system in a subrack.

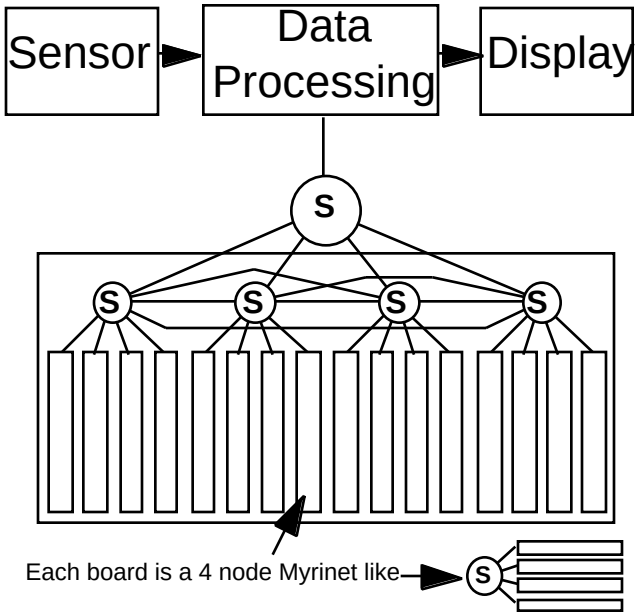

Figure 7.1—4: 64 Processors in a Subrack

8. Appendix B: Existing Practice in Myrinet Routing *(not a part of the standard)*

This Appendix is not a part of the Myrinet specification. It offers a brief exposition of the existing practice in Myrinet routing, in order to help designers familiarize themselves with the Myrinet technology and with the rationale behind it.

Cut-Through Routing

Myrinet switches may employ any type of routing, including store-and-forward routing, but Myrinet was designed to exploit the low latency of cut-through routing. In cut-through routing, the incoming packet is advanced into the selected outgoing channel immediately, provided that the selected outgoing channel is not already occupied by another packet. The packet is then spooled through this established path until the path is broken by the tail of the packet. If the selected outgoing channel is occupied by another packet, the incoming packet is blocked.

Source-Route Encoding

The source-route part of the packet header is generated initially by the hosts or host interfaces, and is interpreted by the switches. The encoding of this source route is at the option of the designers and manufacturers of Myrinet switches.

The Myrinet switches employ cut-through routing based on source routes encoded in single bytes, each with a most-significant bit of 1. Each switch uses a single source route byte and then removes it from the packet. It is expected that future switches of degree higher than 64 ports will use and remove multiple bytes of source route.

Existing Myrinet switches interpret the leading source-route byte as the difference between the outgoing port number and the incoming port number. The advantage of relative port addressing over absolute port addressing is that it is then possible, without knowing the complete map of a network, to "reverse" a route, e.g., for a route from host A to host B of {+3, -2, +5}, the reverse route from host B to host A is {-5, +2, -3}.

Myrinet switches drop packets whose first byte has a most significant bit of 0.

The computation of the outgoing port number does not "wrap," e.g., in a 16-port switch a packet arriving on port 2, with a relative port of -3 at the head of the packet, yields an out-port of -1, which is not valid (and is not wrapped to mean port 15).

Packets addressed to non-existing ports (e.g., to port 19 of a 16-port switch) are dropped. Hence, the valid relative-ports for an 8-port switch, for example, are as listed in the table below.

Extensions to the operation of Myrinet switches could be accessed or controlled through packets addressed to invalid switch ports.

To:	Port 0	Port 1	Port 2	Port 3	Port 4	Port 5	Port 6	Port 7
From Port 0	0	1	2	3	4	5	6	7
From Port 1	-1	0	1	2	3	4	5	6
From Port 2	-2	-1	0	1	2	3	4	5
From Port 3	-3	-2	-1	0	1	2	3	4
From Port 4	-4	-3	-2	-1	0	1	2	3
From Port 5	-5	-4	-3	-2	-1	0	1	2
From Port 6	-6	-5	-4	-3	-2	-1	0	1
From Port 7	-7	-6	-5	-4	-3	-2	-1	0

Table 8–1: The Relative Port Routing for a Myrinet 8-Port Switch

9. Appendix C: Assignment of Myrinet Packet Types

Assignment of Myrinet Types (as of July-31-1998)

(Copied from URL="http://www.myri.com/myri-types.html")

Type (range)	Assigned for (assigned to)
0x0000	Reserved (Myricom)
0x0001	V2 Data Packet (Myricom)
0x0002	V2 Mapping Packet (Myricom)
0x0003	V2 Data Packet (Myricom)
0x0004	Data Packet (Myricom)
0x0005	MyriMap (Myricom)
0x0006	MyriProbe (Myricom)
0x0007	MyriOption (Myricom)
0x0008	GM (Myricom)
0x0009-0x000E	Reserved (Myricom)
0x000F	Mapping (Myricom)
0x0010	Link test packets (Myricom)
0x0011-0x001F	Reserved (Myricom)
0x0020	Ethernet over Myrinet (GM) (Myricom)
0x0021-0x00FF	Reserved (Myricom)
0x0100	Active Messages (UC Berkeley)
0x0110	SSN-Multicasting (UCLA CS Dept.)
0x0200-0x0207	Fast Messages (FM)
0x0300	PacketWay Protocol (PacketWay)
0x0301-0x0302	BDM MCP (MPICH, Mississippi State U)
0x0310	Reserved (Myricom)
0x0330	CSPI packets (CSPI)
0x0340	MPI (MPI Software Technology)
0x0350	BIP (LHPC)
0x0360	Reserved (Myricom)
0x0370	CLF (Digital Equipment Corporation)
0x0400	PM (Real World Computing Partnership)
0x0500	MPI (Hughes Aircraft Co., MPI)
0x0600-0x0601	Isotach (UVa CS Dept. Isotach)
0x0700	VIA: Virtual Interface Architecture (Intel)
0x0800	Portals (Sandia)
0x0900	U-Net (Cornell)
0x0A00	Sanders Packets (Sanders)
0x0B00	AFRL Packets (Air Force Research Lab (AFRL))
0x1000-0xFFFF	Reserved (Myricom)

Consult the above URL for the current assignment of Myrinet packet types.

Requests for Myrinet packet types should be e-mailed to <help@myri.com>.

10. Appendix D: Glossary

BP	Backplane
Character	9 signals conveying a data byte or a control symbols
CRC	Cyclic-Redundancy-Check
Channel	A one-way, point-to-point, byte-wide communication medium
GND	Electrical "ground"
Gbits	Giga bits
Hot Spot	Congestion of packet traffic
I0-I7	Primary port (A-link) input data bits
IB	Primary port (A-link) input Blocking bit
ID	Primary port (A-link) input Data/Control bit
IPG	Inter-Packet Gap
LAN	Local Area Network
Link	A full-duplex pair of opposite going channels between the same ends
LV	Low Voltage
LVDS	Low Voltage Differential Signaling
MB	Mega Bytes
Mb	Mega bits
Mbps	Megabit per second
MTU	Maximum Transmission Unit (i.e., maximum packet length)
NRZI	Non-Return-to-Zero, Invert (see the NRZI reference)
O0-O7	Primary port (A-link) output data bits
OB	Primary port (A-link) output Blocking bit
OD	Primary port (A-link) output Data/Control bit
PCB	Printed Circuit Board
Port	A connection of a link to a component
R0-R7	Secondary port (B-link) or LAN input data bits
RB	Secondary port (B-link) or LAN input Blocking bit
RD	Secondary port (B-link) or LAN input Data/Control bit
Rear Panel	Outer rear surface of subrack containing opening(s) for rear-panel modules. The rear panel is often called bulkhead.
Rear-Panel Connectors	Connectors (typically SAN and LAN) mounted on rear-panel module visible to outside of the subrack
Rear-Panel Module	Module mounted in rear-panel opening with interface to the outer world
Res	Reserved signal for future use. This is not a user-defined pin
SAN	System Area Network
S0-S7	Secondary port (B-link) or LAN output data bits (B-link)
SB	Secondary port (B-link) or LAN output Blocking bit
SBC	Single Board Computer
SD	Secondary port (B-link) or LAN output Data/Control bit
Switch	A multiple port device that steers packets among its ports
Symbol	A control character

[End]