Network Working Group
Internet Draft
Expires in six months

Danny Cohen
Myricom
Craig Lund
Mercury Computers
Tony Skjellum
Mississippi State University
Thom McMahon
Mississippi State University
Robert George
Mississippi State University

October 1997

Part-1 of
The Router-to-Router (RRP) PacketWay Protocol for
High-Performance Interconnection of Computer Clusters
<draft-ietf-pktway-protocol-rrp1-spec-00.txt>

Status of this Memo

This document is an Internet-Draft. Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

To view the entire list of current Internet-Drafts, please check the "1id-abstracts.txt" listing contained in the Internet-Drafts Shadow Directories on ftp.is.co.za (Africa), ftp.nordu.net (Europe), munnari.oz.au (Pacific Rim), ds.internic.net (US East Coast), or ftp.isi.edu (US West Coast).

Table of Content:

1. Introduction

The PktWay protocol is introduced in the "The End-to-End (EEP) PacketWay Protocol for High-Performance Interconnection of Computer Clusters". This document defines the basic part (Part 1) of the Router-to-Router protocol (RRP) of PacketWay.

The shorter "PktWay" is used for "PacketWay".

More information about the PktWay activity is available from the PktWay web site {http://www.erc.msstate.edu/PktWay}.

The architecture of PktWay is very similar to the IP family (indeed, it heavily borrows from IP), with emphasis on performance not generality and scaleability as was selected for IP.

Like IP, PktWay is based on an End-to-End protocol (EEP) that assumes that if an address (or equivalent specification of the destination) is placed in the appropriate field in the packet header, then the packet will arrive to that destination. Neither IP nor EEP specify how this happens.

Routers are responsible to transfer packets from their source networks to their destination networks (possibly via other networks).

The communication among the routers (such the entire family of the GGPs [Gateway/Gateway Protocols] as they were originally called) is NOT a part of IP (as defined originally in RFC-791 and MIL-STD-1777). Similarly, nor is it a part of EEP.

Like the IP family, PktWay defines separately its Router-to-Router Protocol (RRP), in a device- and network-independent manner.

However, the model of routers in PktWay is slightly different from the original model in the IP family. IP routers (or gateways as they were called then) are monolithic devices, provided by their vendors. Each IP-router is a bona-fide host on two (or more) networks. The communication among these intra-router hosts is an internal "private" issue, handled by each vendor as it sees fit, not subject to published standards.

In the PktWay model a router is (like in the IP model) a set of cooperating bona-fide hosts on two (or more) networks. These hosts,

each being a full-fledged host on its SAN are called "half-routers" (HRs).

However, the intra-router communication among these hosts is a "public" issue, handled according to the RRP which defines only the Network-level [Level-2], and not the Physical-level [Level-1], of this communication.

PktWay does not define the nature of this interconnection. However, we believe the PCI Local Bus de facto standard and internal SANs will become a very popular link for short distances, and serial fiber for long ones.

Such an HR may be implemented by separate "boxes" with a long inter-SAN communication link between them, or inside a single "multi-homed" box that has an interface to each SAN, with these interfaces being interconnected via a bus or an internal-SAN.

RRP defines (via message structure and behavior) the interactions between HRs, and between HRs and nodes. RRP does not define the lower level (PHY) protocols that deliver its messages (over links, or between processes). In particular, RRP does not define the inter-SAN interconnection links between the HRs -- these are left for mutual agreements among the implementors of each HR.

RRP defines (like IP's GGP) the router/router and the intra-SAN node/router communication of PktWay. Nodes usually do not communicate explicitly with HRs on other SANs.

The HRs within a single router are called "twins". A router that is connected to N SANs has N HRs, each being a twin of all the other ones. ("Half" and "twin" do not imply that there are only two.)

All the HRs that are connected to the same SAN (being parts of different routers) are called "buddies".

An HR communicates with nodes on its own SAN, with its twins that are on other SANs, and with its buddies that are on its SAN. RRP defines all these communications.

Nodes may ask routers to forward messages to destinations specified either by L2-routes or by L3-addresses. Routers may provide L2-routes to nodes upon their own initiative, or upon request by the nodes.

A node may ask (by [HRTO] messages) any router on its SAN, which router on their SAN is the best to use for a given destination (the

nodes will typically ask their default routers for this information).

In response, the router redirects (using [RDRC] messages) the node to the best router for the specified destination.

At any time routers may "redirect" the node by providing more appropriate local routers for certain destinations, either upon request by the node, or upon the initiative of the router (e.g., to circumvent a fault).

Nodes may ask (by [TELL]) routers for information about other nodes, typically using PktWay-address, name, or capabilities to specify those nodes. In response, routers may provide (by [INFO]) a slew of data about the specified node(s), including physical-address, and optionally logical-addresses, name, and capabilities, if any.

PktWay nodes may use a SRVLOC to locate required resources.

It is assumed that each HR has a Routing Table (RT) for its own SAN (aka Local Routing Table, LRT), with (at least) the addresses of all the nodes and the source routes to each of them from the HR (and possibly also names and capabilities for each node). This information could be dynamic or static, even manually configured. The HRs may (or may not) perform dynamic mapping of their SANs.

It is also assumed that each node, on each SAN/LAN, knows the SR to at least one HR on its SAN/LAN, and that it has a default-HR defined.

In order to be able to provide the nodes with such information, each HR must collect this information about all the nodes in its own SAN. This may be performed dynamically, or statically, in either an automated or manual manner. RRP does not sepcify how this information is gathered.

Each HR gives its Local Routing Table to all his twins. HRs always share with twins information received from buddies, and with buddies information received from twins. This yields the global mapping of the PktWay.

All the various RRP messages are composed of a small set of common records. This document defines the messages, their structure, their common records, and their format. Several examples are used to illustrate the operation of the RRP.

RRP specifies a series of options that allow system designers to deploy PktWay nodes and routers of varying levels of capabilities ("intelligence").

There are four implementation levels of PktWay, indicated by a letter code. The higher the letter code ("A" = lowest), the more interoperability and adaptability result. System designers may choose the level of implementation to best suit their needs.

2. A note about the PacketWay Documents

The PacketWay protocol is defined by a series of documents:

* EEP (End-to-End Protocol)
* RRP-1 (basic Router-to-Router Protocol)
* RRP-2 (dynamic inter-SAN routing)
* PktWay enumerations

Each of these documents should include the same "PacketWay at a Glance (Cheat-Sheet)", this note, and the Notations page. They should include also (as appendices) a copy of the PacketWay glossary of terms and its acronyms and abbreviations list.

The EEP and the RRP documents will be published first as Internet-Drafts and later as Proposed-Standards, Draft-Standards, and Standards.

The Enumeration Document will be first published as an "Informational-RFC" and later will be maintained by IANA.

The enumeration document may be attached to the EEP/RRP documents, as a matter of convenience. The enumeration is NOT a part of the PktWay standard, just as RFC0739 (the original "Assigned Numbers" RFC) is not a part of RFC0791, that defines IP.

Similarly, the EEP-document has "Appendix-A: A Recommendation for PktWay Address Assignment" which is a recommendation only and NOT a part of the PktWay standard, just as IP-address-assignment is not

a part of RFC0791, that defines IP.

The appendices are brought for clearance and convenience. They are not a part of the PktWay specification.

Information about the PktWay activity may be found in the URL: http://www.erc.msstate.edu/PktWay/

3. Notations

The shorter "PktWay" is used for "PacketWay".

8B means "8-byte" (64 bits).

0x indicates hexadecimal values, e.g., 0x0100 is 2^8=256(decimal).

0b indicates binary values, e.g., 0b0100 is 4(decimal).

xxxx indicate a field that is discarded without any checking (e.g., padding).

[fff] indicates that fff is an optional field, when appropriate.

[exp] in equations, is the integral part, rounded down, of `exp`. e.g., [23/8]=2.

All length fields do not include themselves, and therefore may be 0.

Lengths are specified either (a) by byte count, implying that some padding bytes may follow to fill 8B-words, or (b) by 8B-word count and PL, the number of trailing padding bytes (with PL between 0 and 7).

4. The Four Implementation Levels of RRP

Level-A: Hosts have pre-wired (static) native routing. It's an L2 ("MAC"-based) operation. HRs do not provide any info to nodes, nor to other HRs. No RRP-messages are used in this level.

Level-B: L2 (MAC based) or L3 forwarding (planner transfers, IP-like operation). Nodes may ask HRs for L2 routing and for the HR to use for given destinations. In this level the following

RRP-messages are used: [GVL2], [L2SR], [HRTO], and [RDRC]. In addition the [WRU]? and [INFO] messages may be used, too.

Level-C: Node discovery (with static or dynamic routing). In this level nodes may ask HRs for information about other nodes, including their capabilities. In this level the [TELL] and the [INFO] RRP-messages are used in addition to those of Level-B.

Level-D: In this level there is a dynamic exchange of routing tables among the HRs. This create globals mapping of the PktWay, and allows for dynamic circumvention of faults. The [GVRT] and the [RTBL] RRP-messages are used for this exchange among the HRs.

Level-D applies only to routers, not to nodes.

Level-B is an extension of Level-A (i.e., Level-B exists only with Level-A). Level-C and Level-D are independent extensions of Level-B.

Level-B is the basic level of RRP. This document, RRP Part-1 (aka RRP1), defines the RRP messages used for Level-B and Level-C. Level-D is defined in RRP Part-2 (aka RRP2).

In L2 operation under Level-B , when a source node, SN, needs to send a message to a destination node, DN, it first uses a [GVL2] message to ask any of the HRs on the SN's SAN for a source route (SR) from HR to DN. That HR would either (1) use an [L2SR] message to provide such an SR, or (2) use an [RDRC] message to "re-direct", by suggesting to SN to use the specified HR (which is also on SN's SAN), or (3) use an error message to report no knowledge of DN (using the UNK error message).

SN may ask more than one HR for SRs to the same DN and use any algorithm to choose which of these SRs to use.

RRP does not specify whether (and how) to cache SRs.

In L3 operation, when a source node, SN, needs to send a message to a destination node, DN, it sends that message to any of the HRs on its SAN, using L2, expecting L3-forwarding to DN, using DN's PktWay address. That HR would either (1) forward the message toward DN, and possibly return to SN a "re-direct" message, suggesting to use, in the future, another HR on SN's SAN for DN, or (2) report no knowledge

of DN (using the UNK error message).

Under Level-C nodes may be located by PktWay-addresses, names, or capabilities, but only addresses may be used for routing.

5. Use of RRP Messages by Levels

Level-A: no RRP messages used

```
Level-B: nodes send:      HRTO, GVL2, WRU?, INFO
         nodes receive:   RDRC, L2SR, INFO, WRU?
         routers receive: HRTO, GVL2, WRU?, INFO
         routers send:    RDRC, L2SR, INFO, WRU?

Level-C: nodes send:      HRTO, GVL2, WRU?, INFO, TELL
         nodes receive:   RDRC, L2SR, INFO, WRU?
         routers receive: HRTO, GVL2, WRU?, INFO, TELL
          routers send:    RDRC, L2SR, INFO, WRU?

Level-D: nodes send:      HRTO, GVL2, WRU?, INFO
         nodes receive:   RDRC, L2SR, INFO, WRU?
         routers receive: HRTO, GVL2, WRU?, INFO, GVRT, RTBL
         routers send:    RDRC, L2SR, INFO, WRU?, GVRT, RTBL
```

This RRP1 document defines the 7 messages required for Levels B and C (HRTO, RDRC, GVL2, L2SR, TELL, INFO, and WRU?). The RRP2 document defines the 2 messages required for Level D (GVRT and RTBL). In addition, a few error messages are also defined.

6. Node Attributes

Each node must have a Physical Address. Optionally it may also have Name, Capabilities, and Logical-Addresses:

Physical Address: 23 bits, flat, unique in this PktWay.

Name: flat, globally unique (e.g., IP address), arbitrary length

Capabilities: regular GP node, router, PktWay-server, NFS, paging server, M/C server, SRVLOC-server, DSP, printer,...

Some capabilities may need additional parameters (e.g., SAN-ID for routers, and resolution+colors

for printers).

There parameters are capability-specific.

The capabilities are defined in the PktWay Enumeration document.

Logical-Addresses: a set of (logical) addresses to which this node requests to listen. Logical addresses designate multicast and broadcast groups.

The control of the Logical-Addresses (a la IGMP) is not defined in this document. This will be designed by the applications that use it (e.g., PktWay-multicast).

The management of logical addresses (e.g., JOIN and LEAVE) is not defined here.

7. RRP Messages

RRP messages are PktWay messages with PT="RRP" and TE=RRP-type, in their EEP-header, followed by some (zero or more) RRP-records according to their RRP-type, followed (always) by the PktWay-TAIL which is the EI field.

The RRP-records constitute the Data Block (DB) of the PktWay-message. They must be in Big-Endians order, with e=0 in the EEP-header.

Following are the 7 RRP messages (for Level B and C), with their RRP-type, and the related error messages. The column S->D (Source to Destination) shows who sends such messages to whom, where N is

for Node, H is for HR, and A is for Any.

RRP-Type	S->D	Description
[GVL2]	N->H	Please give me L2-routes to node (address) Replies to [GVL2]: [L2SR], [RDRC], or [ERR/UNK].
[L2SR]	H->N	Here are L2-routes to node (address)
[HRTO]	N->H	Which HR should I use for node (address)? Replies to [HRTO]: [RDRC] or [ERR/UNK].
[RDRC]	H->N	Re-direct to node (address) via an HR on same SAN
[TELL]	N->H	Please tell me about node (address, name, capa's) The reply to [TELL] is [INFO], or [ERR/UNK].
[INFO]	A->A	Info about node (address, name, capabilities, LAs)
[WRU?]	A->A	Who/what-Are-You? (Tell me all about yourself) The reply to [WRU?] is [INFO] about the replier.

RRP also uses the following error messages:

[ERR/UNK]	Destination Unknown (address)
[ERR/HRDOWN]	HR Down
[ERR/LKDOWN]	Link Down
[ERR/GENERAL]	General error message

8. RRP Message Structure

The RRP-messages are made of RRP-records, distinguished by their Record-Type (RTyp). These RRP records are:

RTyp	Description
ADDR	Address
NAME	Name

CAPA Capability
LADR Logical Addresses
SRQR Source Route and its Quality (SR,Q)
MTUR MTU (for the preceding SRQR)

The RRP-records are made of 8B-words. The following shows the RRP-records that make each of the RRP-messages. Each message starts with a PH (PktWay-header), and ends with a PT (PktWay-TAIL). The TAIL is not shown here.

* [GVL2] Please give me L2-routes from you to node (address)

PH (with [PT/TE]=[RRP/GVL2])
ADDR (address of the node for which SR is requested)

* [L2SR] Here are L2-routes from me to node (address)

PH (with [PT/TE]=[RRP/L2SR])
ADDR (address of the node for which SR is provided)
SRQR (SR with Q) possibly with a few
L2RH records MTUR (MTU for the above SR)

This message may have several (SRQR,MTUR)s, one for each SR.

* [HRTO] Which HR should I use for node (address)

PH (with [PT/TE]=[RRP/HRTO])
ADDR (address of the node for which initial HR is requested)

* [RDRC] Re-direct to destination node (address) via a HR (address), on the same SAN.

PH (with [PT/TE]=[RRP/RDRC])
ADDR (address of the destination node)
ADDR (address of the HR to be used for that destination)

The above addresses are expected to be physical (but they be otherwise).

* [TELL] Please tell me about node (address | name | capabilities)

```
      PH   (with [PT/TE]=[RRP/TELL])
      ADDR (address of that node)
   or
      PH   (with [PT/TE]=[RRP/TELL])
      NAME (name of that node)
   or
      PH   (with [PT/TE]=[RRP/TELL])
      CAPA (capabilities for which nodes are requested)
```

  This message may have several CAPA's, one for each capability.

  [TELL] identifies a node by an address and/or a name and/or capabilities. If more than one attribute is specified (e.g., an address and name(s)) any nodes that meets any of them should be considered (like an implied OR).

* [INFO] Info about node(s) (address, name, capabilities)

```
      PH   (with [PT/TE]=[RRP/INFO])
      ADDR (address of that node)
      NAME (name of that node)
      CAPA (capabilities for which nodes are requested)
      LADR (Logical-Addresses for the requested node)
```

  This message may have several CAPA's, one for each capability. For nodes without NAME, LADR, or any CAPA, these records are omitted.

  [INFO] provides all the known information about all the nodes that match the [TELL]. The [ADDR} records are the separators between the nodes.

* [WRU?] (CD) Who/what-Are-You?

```
      PH   (with [PT/TE]=[RRP/WRU?] and [DD]=0x7FFFFE)
```

* [ERR/UNK] Destination Unknown (address)

```
      PH   (with [PT/TE]=ERROR/UNK)
      XXXX (XXXX of the Destination node for which the requested
            information is not available), where XXXX is the ADDR
            and/or NAME and/or CAPA of the node(s) about which this
            message is sent
```

* [ERR/HRDOWN] HR Down (or Router-Down)

```
PH   (with [PT/TE]=[ERROR/HRDOWN])
ADDR (address of the HR that is down)
ADDR (the other address of the router that is down)
```

* [ERR/LINKDOWN] Link Down

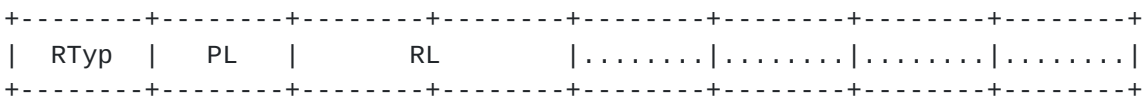
```
PH   (with [PT/TE]=[ERROR/LINKDOWN])
ADDR (address of one end of the link that is down)
ADDR (address of the other end of the link that is down)
```

* [ERR/GENERAL] General Error (i.e., none of the above)

```
PH   (with [PT/TE]=[ERROR/GENERAL])
XX   (The entire message that caused the error PH+OH+DB+TAIL)
```

9. RRP Record Format

Each RRP-record starts with an 8B-word header as shown below. Its first byte identifies the record type (RTyp). The second byte is the Pad-Count byte (PL) indicating the number of padding bytes. The third and the fourth bytes (RL) are the length (in 8B-words) of the record, excluding the record header, hence it may be zero. The rest of the header bytes depend on the record type (RTyp).



Some records that have an arbitrary length are "right justified" and have PL padding bytes before the data. Padding Before Data [PBD].

Some records that have an arbitrary length are "left justified" and have PL bytes after the data. Padding After Data [PAD].

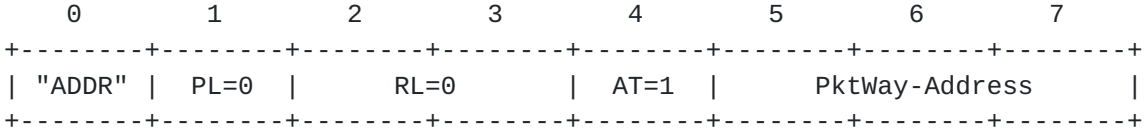
In either case the total number of data bytes is: (8*RL+4-PL).

Following are the RRP-records. These records are the building blocks used to construct RRP-messages.

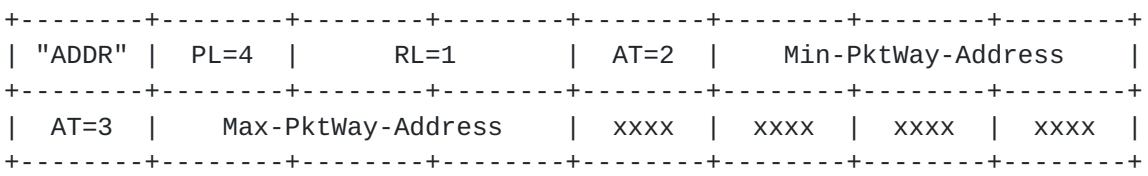
In the following xxxx indicate bytes that are discarded, such as for padding. It is recommended to set them to all-0.

===> [ADDR] Node-Address Record [PAD]

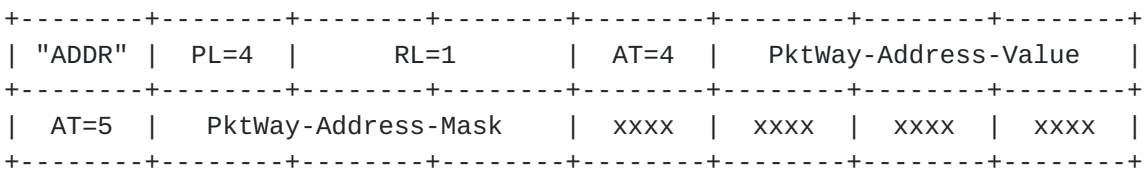
This record specifies either a single address (with AT=1) or a range of addresses (with AT=2 followed by AT=3, or by AT=4 followed by AT=5). AT is the "Address-Type".



or:



or:



The address-mask follows the address-value.

The above addresses may be physical or logical.

The address X is specified by an ADDR record if:

if AT=1: X == PktWay-Address

if AT=2,3: Min-PktWay-Address <= X <= Max-PktWay-Address

if AT=4,5: (PktWay-Address-Mask & X) == PktWay-Address-Value

An ADDR-record defines only one PktWay-address (or one range), unlike an LADR record (see below) that may specify multiple addresses and multiple address-ranges.

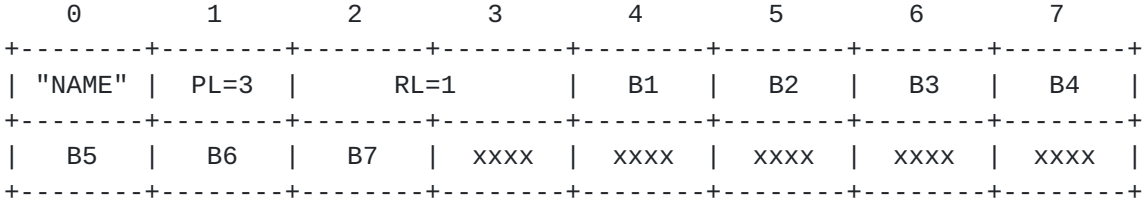
If the ADDR record is followed by other records that describe the same node (such as NAME, CAPA, LADR, SRQR, and MTUR) then the RL of the ADDR records also covers all these records. All these records apply to all the addresses specified in this ADDR-record. Needless to say that NAME is not expected to appear within a record that specifies more than one

address.

Hence, if an ADDR-record with AT=1 has RL>1, or if an ADDR-record with AT>1 has RL>2, then this ADDR-record includes additional records (such as CAPA, LADR, SRQR, and/or MTUR) about the specified address(es).

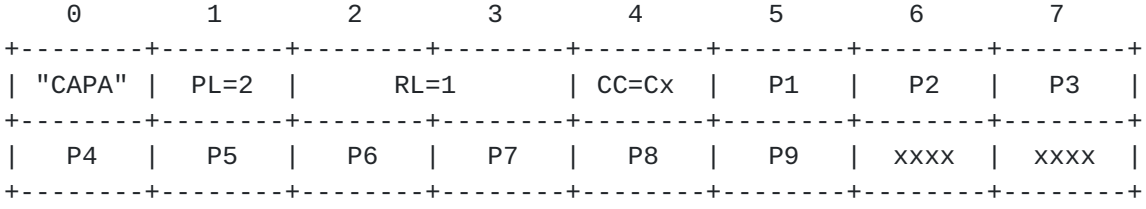
The enumeration is guaranteed not to have overlap between the AT and the RTyp codes.

===> [NAME] Node-Name Record [PAD] (e.g., a name with 7 bytes B1..B7)



The number of bytes in the name is 8*RL+4-PL.

===> [CAPA] Node-Capability Record [PAD] (e.g., 9 parameter bytes):

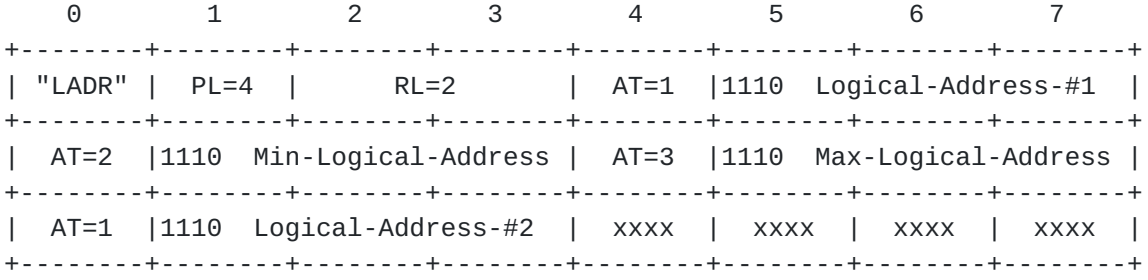


Byte#4 is the Capability Code, CC, followed by as many parameter bytes as needed (9 in the above example).

The capability codes are listed in the PktWay Enumeration document.

The number of bytes used by the parameters is 8*RL+3-PL.

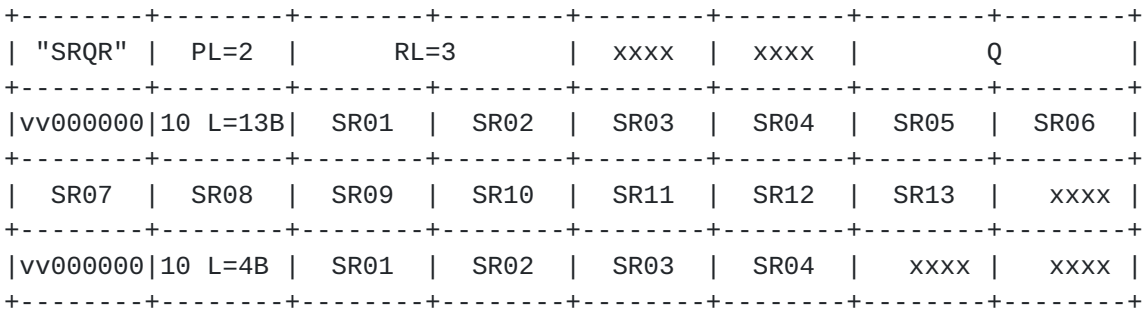
===> [LADR] Logical-Addresses Record [PAD]
(e.g., 2 logical addresses and a range of logical addresses)



Whereas an ADDR-record defines only one PktWay-address (or one range), an LADR record may specify multiple addresses (each with AT=1) and multiple ranges (each with a pair of AT=2,3 or AT=4,5).

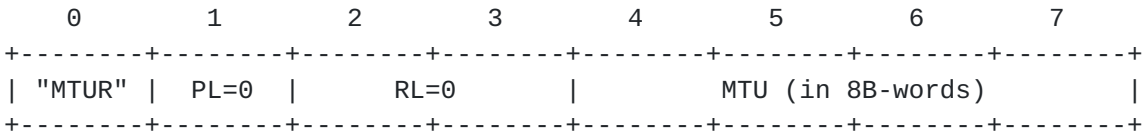
===> [SRQR] Source-Route Record [PBD], with Q for that route.
(e.g., an SR combined of 2 L2RHs, one with 13 bytes and one with 4 bytes)

This record carries one, or more, L2RHs (2 in the following example, one with SR of 13B, followed by an SR of 5B).



Q (the Route Quality) is an unsigned 16-bit integer. The units are not defined here. It is assumed that it is monotonic with all-0 being the best and all-1 the worst. If there is an MTUR (MTU-record) for that SR it should follow this SRQR record. However, the RL of the SRQR does not include the RL of the MTUR.

===> [MTUR] MTU record [PBD]:



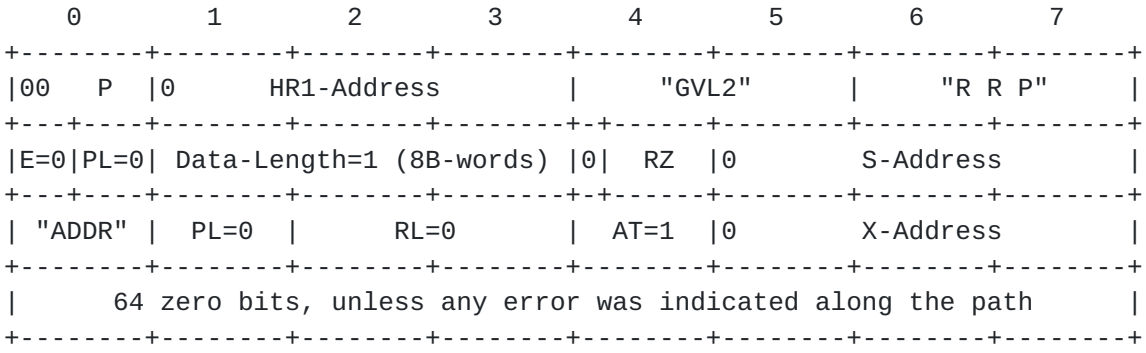
The MTU record provides the MTU for the SR defined before (by an SRQR).

The value of 0 means indefinite MTU (i.e., any length is OK).

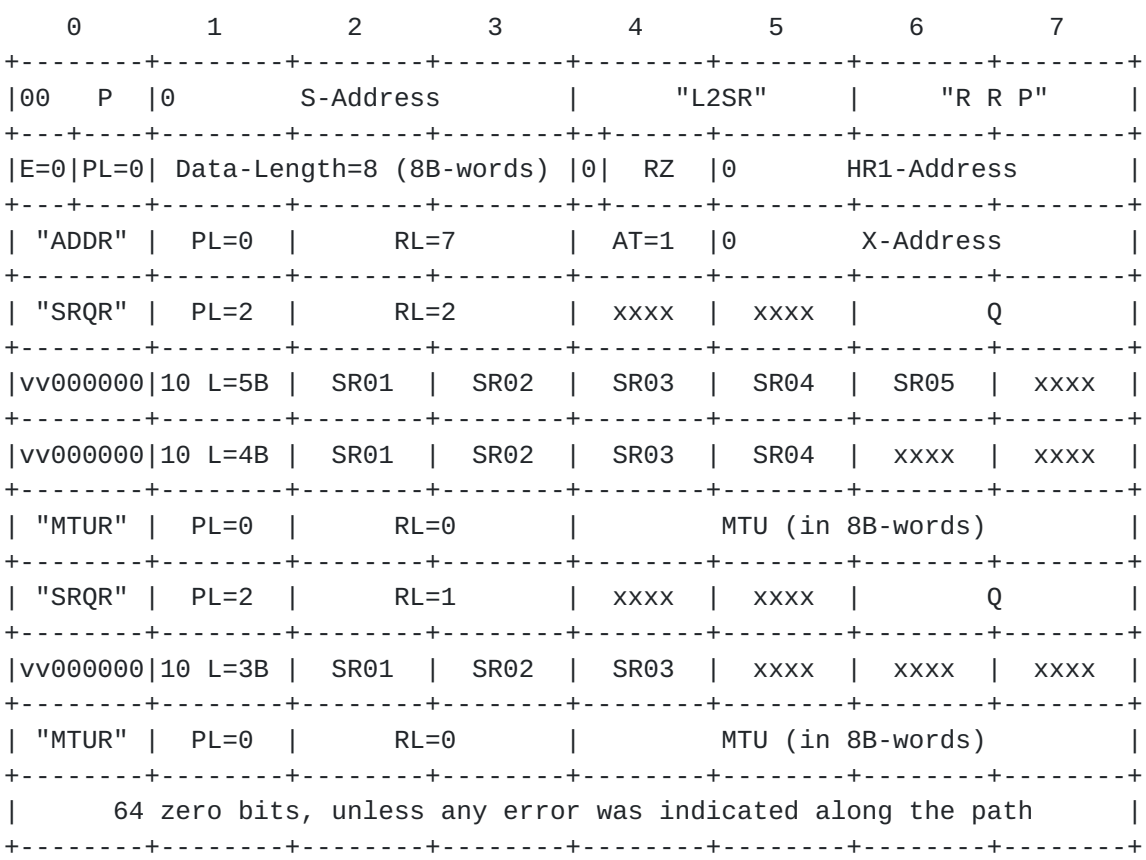
10. Examples for RRP Message

Node-S asks HR1 to provide an L2RH to node-X:

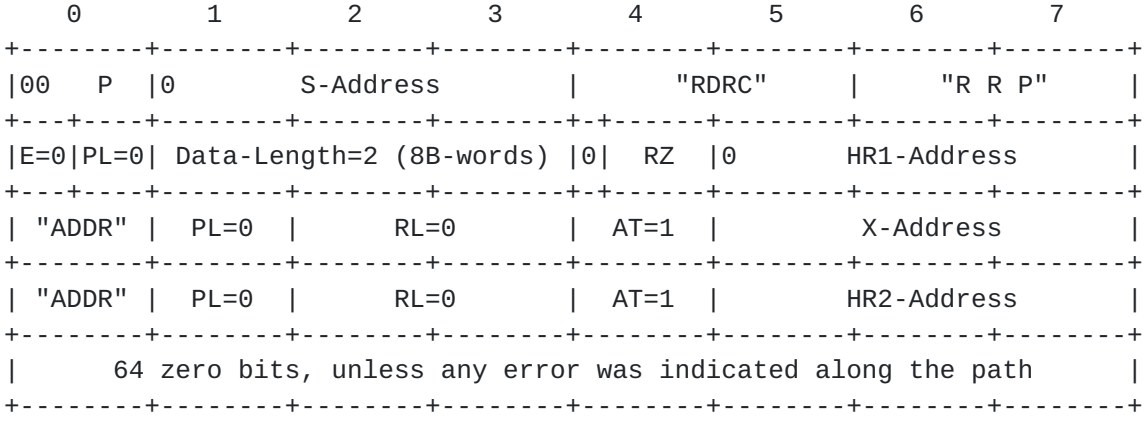
==> [GVL2] Please give me L2-routes from you to node-X



==> [L2SR] HR1 replies with two L2-routes to node-X with Qs and MTUs (e.g., an SR of 2 L2RHs (of 5+4 bytes), and an SR of 1 L2RH of 3 bytes)



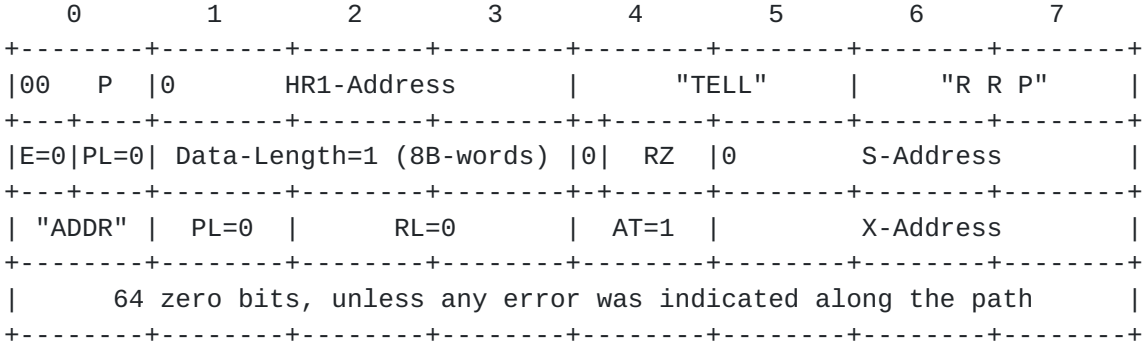
==> [RDRC] HR1 redirects Node-S to use HR2 for node-X



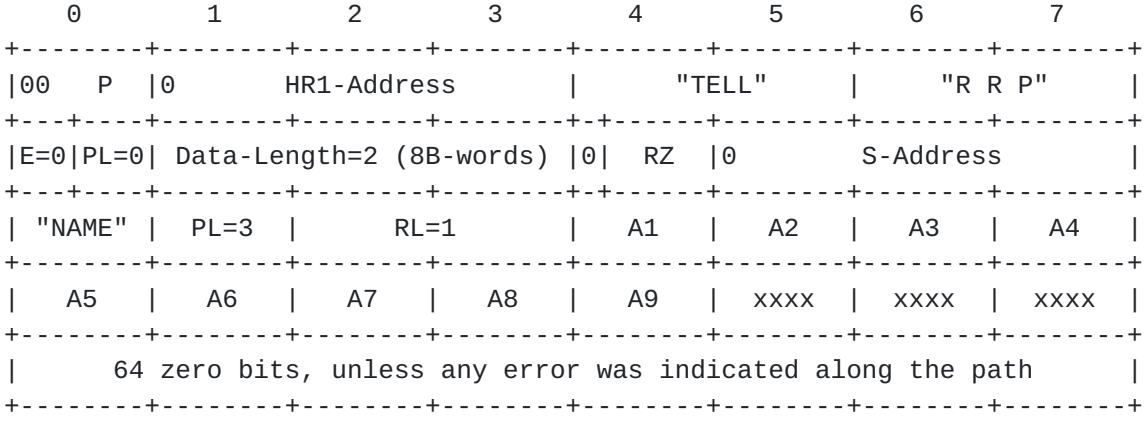
==> [TELL] Please tell about Node-X (address | name | capabilities)

This message may have any of the following 3 forms:

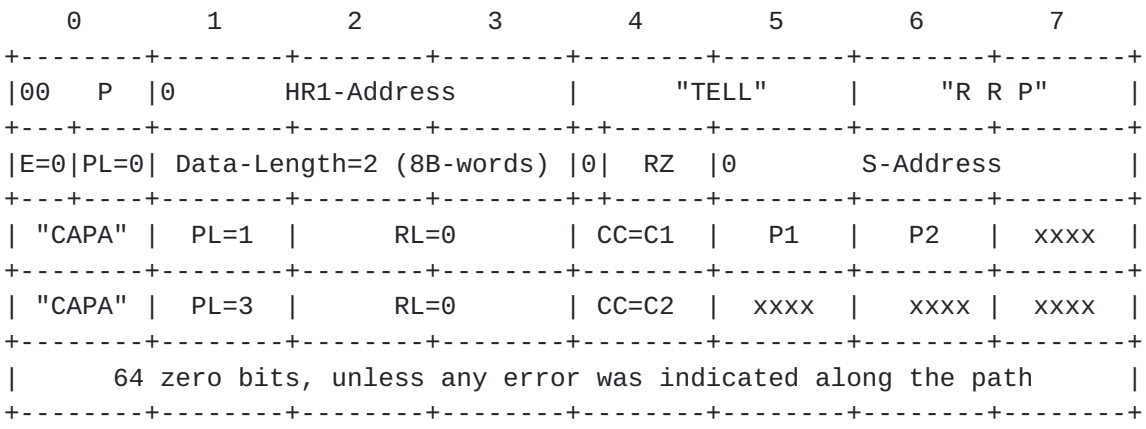
If by PktWay-address:



If by name (e.g., a name with 9 characters: A1...A9):

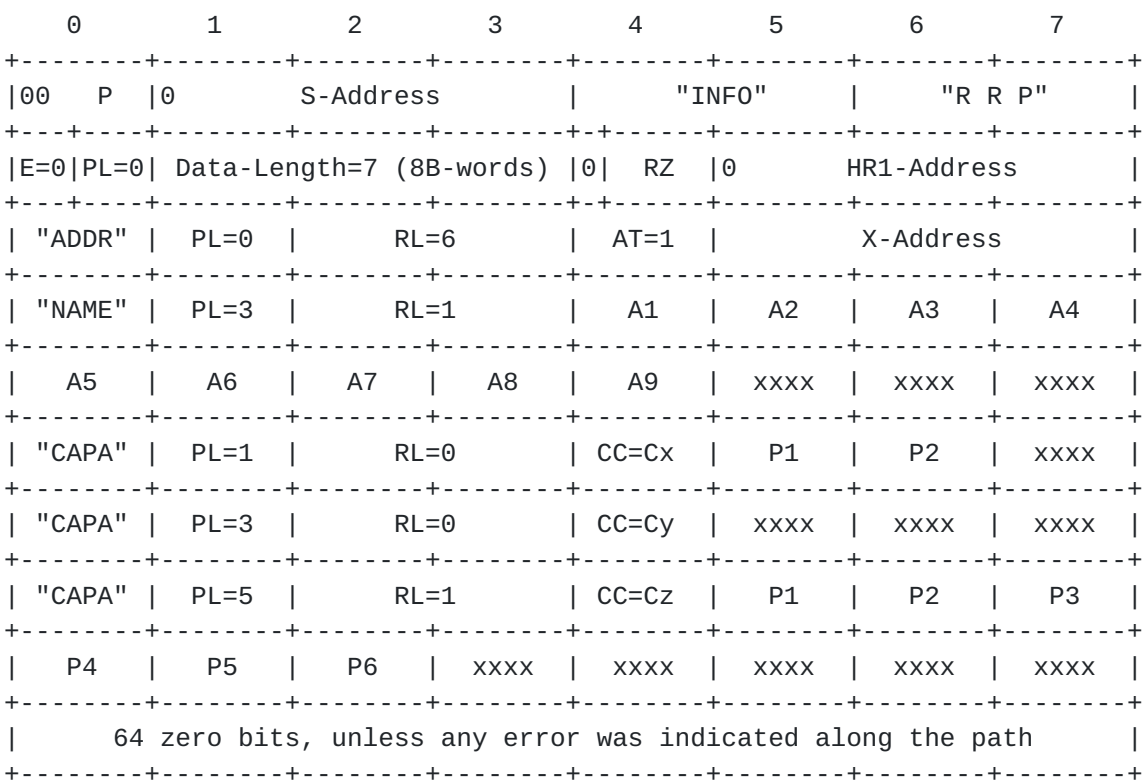


If by capabilities (e.g., 2 capabilities, C1 with 2 parameter bytes, and C2 with no parameter bytes):



A [TELL] may specify several nodes, by addresses, names, and capabilities. Any node that matches any of the specifications in [the TELL] should be included in the replying [INFO], in separate ADDR records.

==> [INFO] Info about Node-X (address, name, capabilities) e.g., a name with 9 characters (A1...A9) and 3 capabilities (Cx, Cy, and Cz):

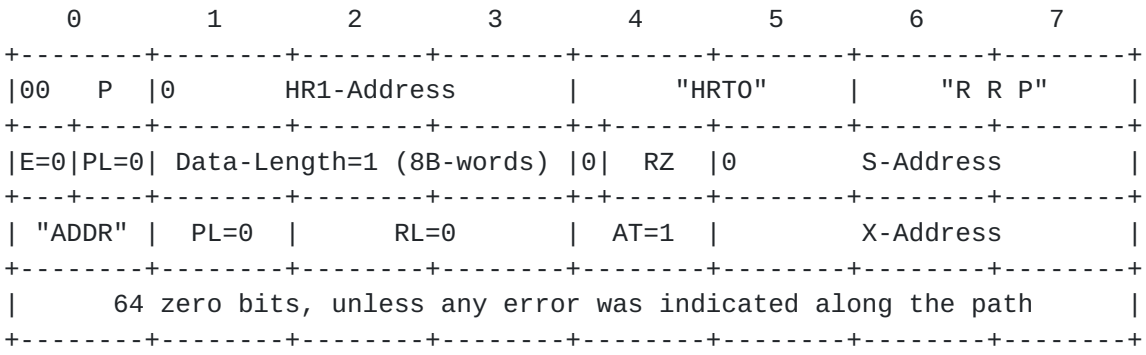


The INFO records should specify all the nodes that meet any of the attributed specified in the TELL record. When such aggregation is used, the DL (data length) in the PH is the sum of the (RL+1)s of all the ADDR fields.

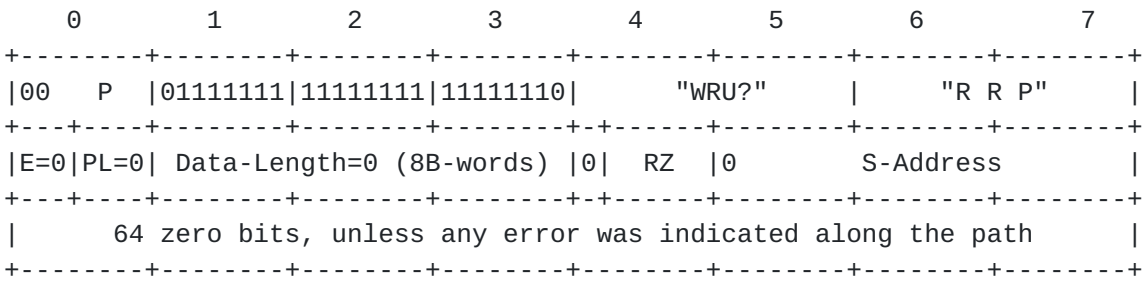
(*) The ADDR, NAME, and CAPA records are repeated for each applicable node. Same also for LADR, SRQR, and MTUR, if any.

If several capabilities are specified in [TELL], any node that has any of these capabilities should be reported in [INFO].

==> [HRTO] Node-S asks HR1 which HR to use for Node-X.

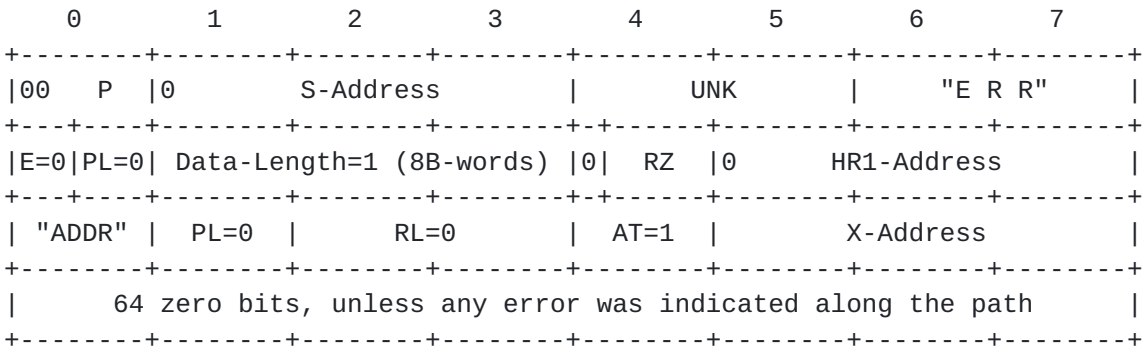


==> [WRU?] Who/what-Are-You?



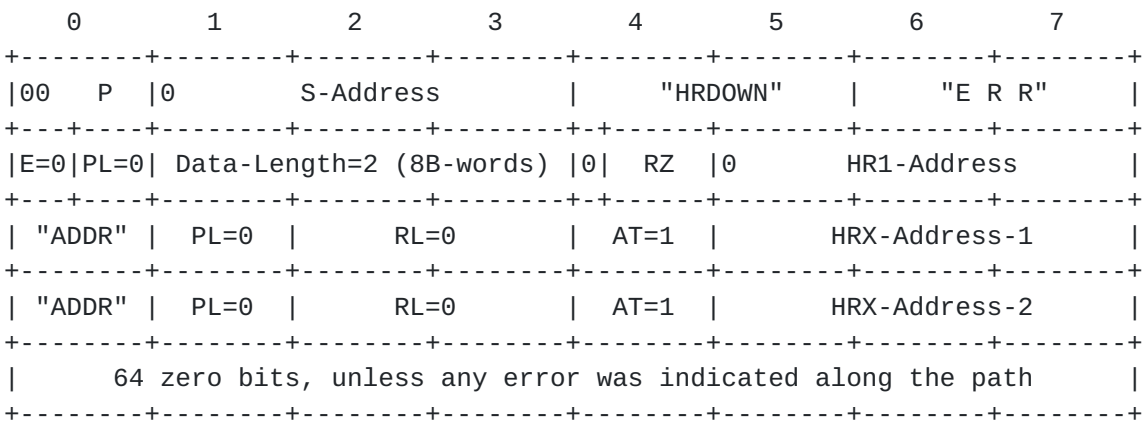
This is addressed to 0x7FFFFE, the "Hey-You" address.

==> [ERR/UNK] Destination Unknown (address). HR1 tells Node-S that he does not know about Node-X.



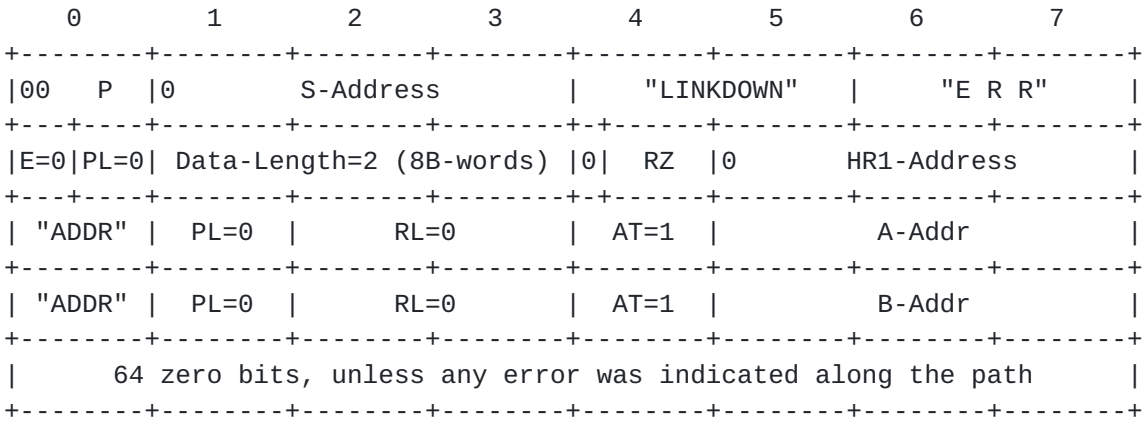
This message reports that host (X) is unknown to S.

==> [ERR/HRDOWN] HR Down (2 addresses).
HR1 tells Node-S that HR-X is down



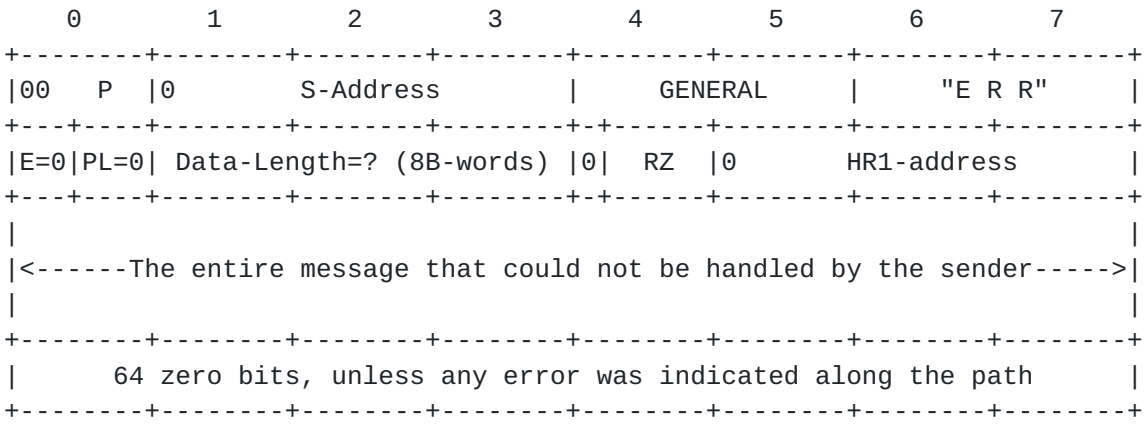
HR1 knows 2 addresses of the downed router.

==> [ERR/LINKDOWN] Link Down (2 addresses)



This message reports that the link between A-Addr and B-Addr is down.

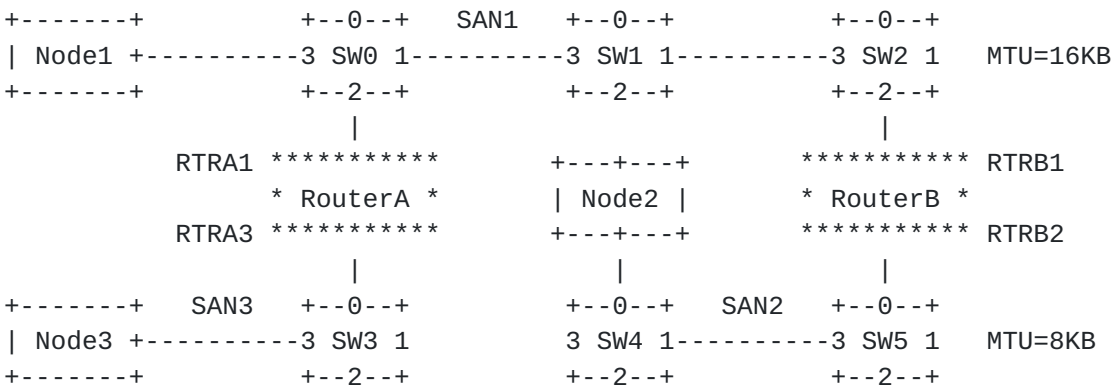
==> [ERR/GENERAL] General error: HR1 tells node-S that it (HR1) could not handle the enclosed message)



This message reports that the enclosed message could not be handled by its receiver (the sender of this error message).

11. Appendix-A: Example of the use of RRP

The following PktWay is used for the example. It included 3 SANs, interconnected via 2 routers, Router-A (RTRA) between SAN1 and SAN3, and RTRB between SAN1 and SAN2.



In this example Node1 on SAN1 (with MTU=16KB) is looking for Node2 which is on SAN2 (with MTU=8KB). It first asks its default router (RTRA1) for an L2RH to Node2. RTRA1 redirects Node1 to RTRB1 regarding Node2.

Node1 asks RTRA1 (by [HRTO], in message M1) which router to use for Node2. RTRA1 suggests (using [RDRC], M2) to use RouterB. Node1 uses L3-forwarding ([WRU?], M3), via Router-B, to verify that RTRB can indeed get to Node2, by asking Node2 for information about itself. Node2 provides this information ([TELL], M4) which Node1 likes. Node1 asks RouterB ([GVL2], M5) for L2RH(s) to Node2. RouterB

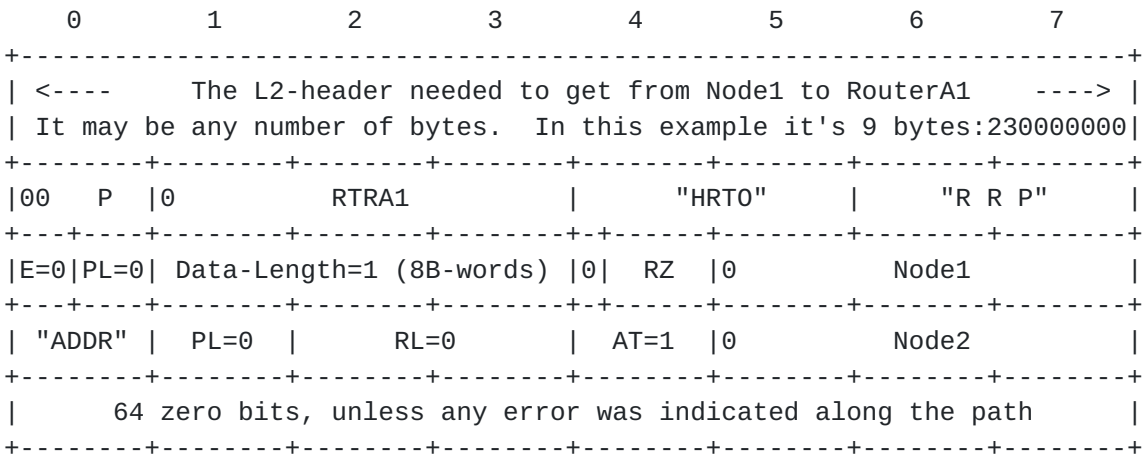
provides ([L2SR], M6) the requested L2RH with its MTU of 1,024 8B-words (8KB).

Finally, Node1 sends data (by M7) to Node2 using L2-forwarding. Similarly, Node2 may ask its default router which HR to use for Node1 and for L2RH(s) to Node1.

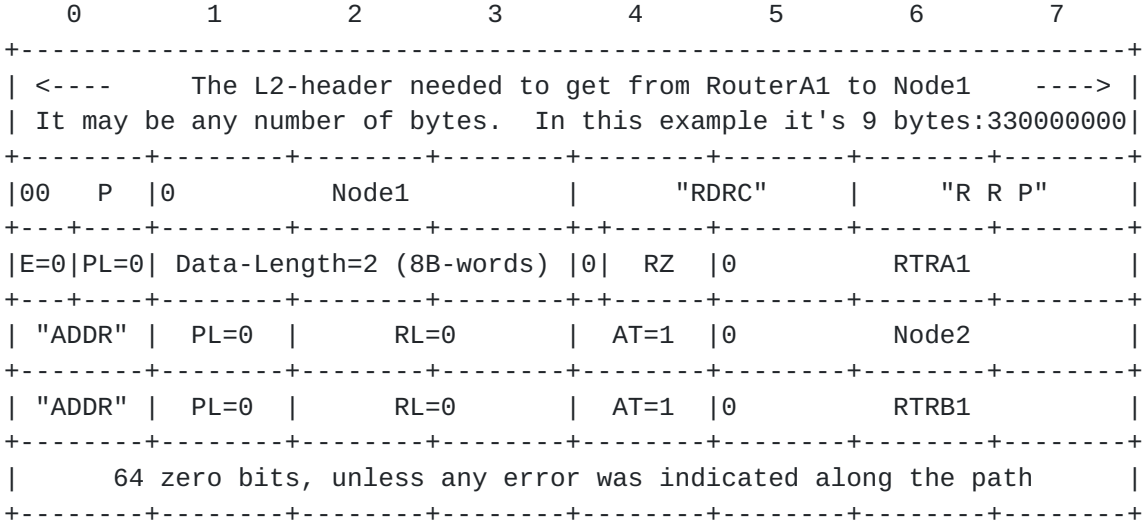
If Node1 had only Level-A implementation then it should have the combined L2RH from itself to RouterB and from there to Node2 pre-wired, saving all this message exchange.

The sequence of messages (M1 thru M7) is shown below.

(M1) Node1 sends [HRTO] to its default router RTRA1 asking which HR to use for node2.

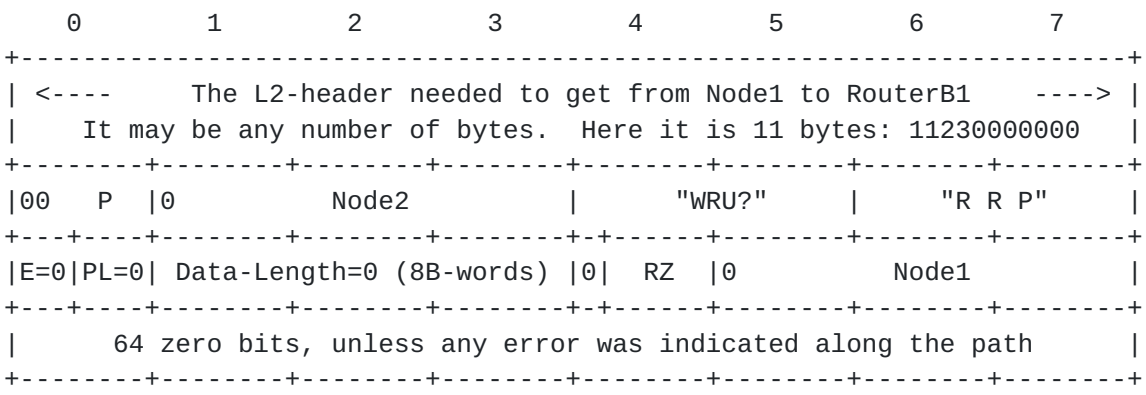


(M2) RTRA1 uses [RDRC] to re-direct to Node2 via RouterB.

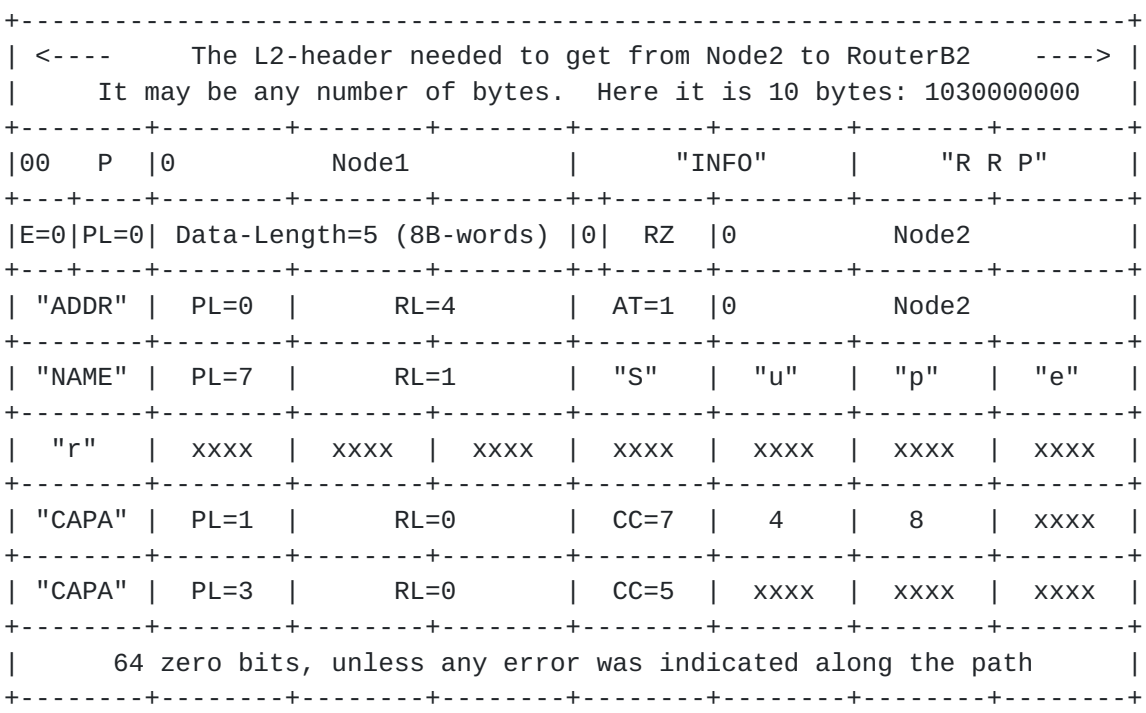


Node1 knows how to get to RouterB over its SAN.

(M3) Node1 uses [WRU?] (still using L3-forwarding via RouterB) to verify the capabilities of Node-2, and that RTRB can indeed get to it. This is done by asking Node2 for information about itself.

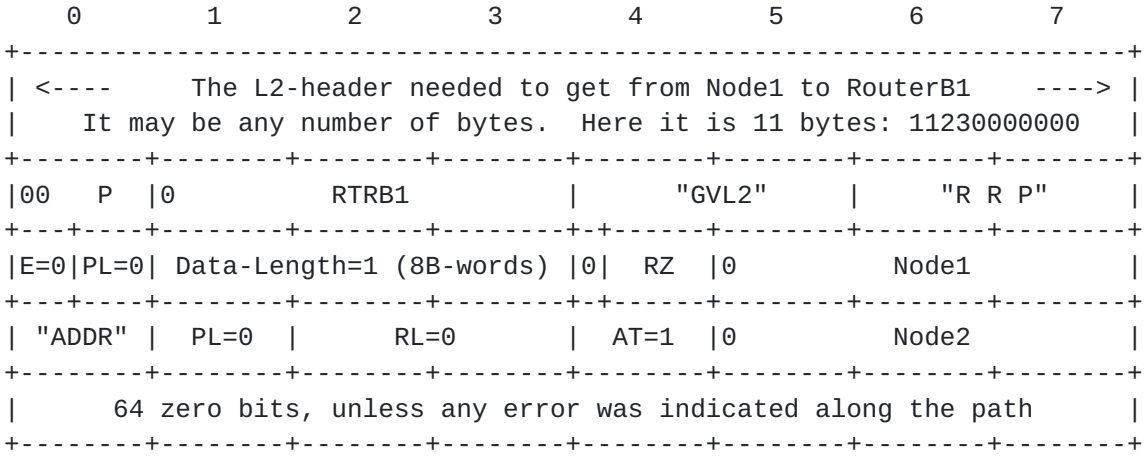


(M4) Node2 uses [INFO] (via RouterB2, also using L3-forwarding) to provide information about itself to Node1. This info includes its PktWay-address and its name ("Super"). If Node2 had implemented also Level-C of the RRP it would also provide a record about its capabilities (as shown in this example with 2 capabilities (with codes of 5 and 7).

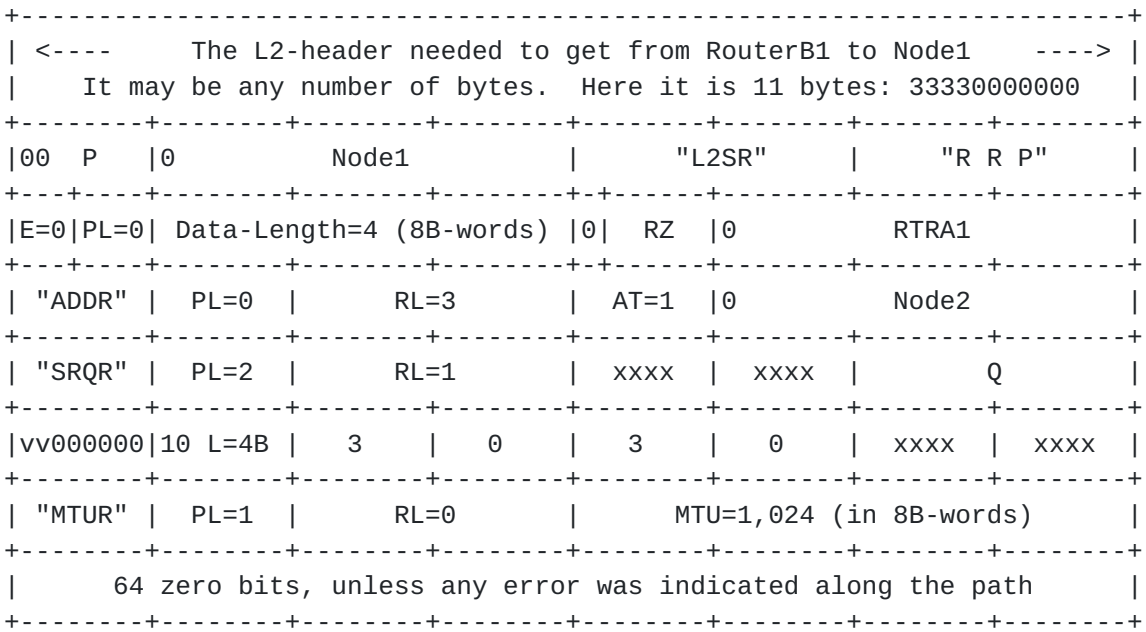


By receiving this message Node1 knows that RTRB could indeed be used for communication with Node2.

(M5) Node1 uses [GVL2] to ask RouterB for L2RH(s) from RouterB to Node2.



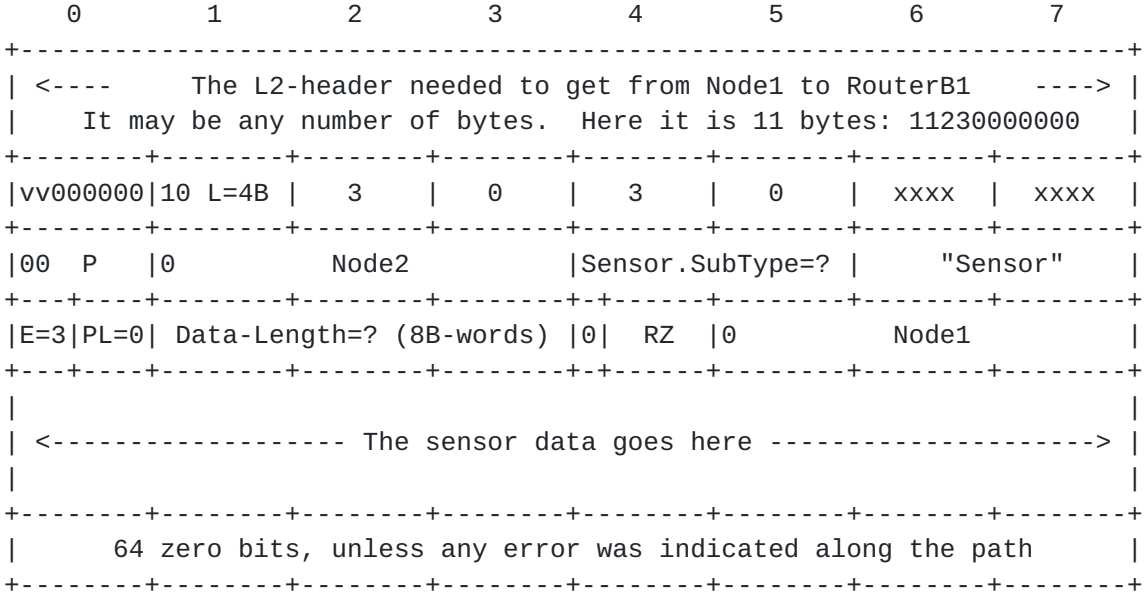
(M6) RouterB uses [L2SR] to provide Node1 with an L2RH from RTRB2 to Node2, with its Q and MTU. This L2RH is {3,0,3,0,0,0,0,0,0,0} from RouterB to Node2, and the MTU is 1,024 (meaning 8KB).



The MTU in the MTUR above is the lessor of the MTUs of both networks.

The RL (record-length) of the last MTUR-record is NOT included in the RL of the preceding SRQR-record, but is included in the RL of the preceding ADDR-record (since the RL of the SRQR is included in the RL of the ADDR). The RL=3 of the ADDR includes 2 words of SRQR and 1 word of MTUR.

(M7) Finally, Node1 sends data to Node2 using L2-forwarding.



E=3 (0b0011) indicates that all the data is 64-bit, Big Endian order.

Again, if Node1 had only Level-A implementation then it would have pre-wired the combined L2RH from itself to RouterB and from there to Node2, saving all this message exchange.

All the messages shown in this appendix start with local L2 routing bytes needed to get across either SAN1 or SAN2 (indicated with "The L2-header needed to get from ... to ...") which are not L2RHs. The difference is that these bytes are in front of the packet, exposed to the local switches, whereas the L2RHs are only exposed to PktWay-entities.

These local L2 routing bytes are the actual bytes required by the SANs and likely to be consumed as the messages traverses the SAN, unlike the L2RHs that are intact until converted to actual routing bytes.

The L2RHs start with 0bvv00000010 followed by the number of routing bytes in that L2RH, and possibly also by several bytes of padding.

12. Appendix-B: Glossary

Address: A unique designation of a node (actually an interface to that node) or a SAN.
Buddy-HR: HRs are "buddies" if they are on the same SAN.
Cut-Thru: See wormhole.
Destination: The node to which a packet is intended
Dynamic-Routing: Routing according to dynamic information (i.e., acquired at run time, rather than pre-set).
Endianness: The property of being Big-Endian or Little-Endian (transmission order, etc.)
Ethertype: A 16-bit value designating the type of Level-3 packets carried by a Level-2 communication system.
HR: Half-Router, the part of a router that handles one network only.
L2-Forwarding: Forwarding based on Level-2 (i.e., data-link layer of the ISORM) information, e.g., the native technique of each SAN or LAN. Also called "source routing."
L3-Forwarding: Forwarding based on end-to-end (Level-3 i.e., network layer of the ISORM) addresses. Also called "destination routing."
Map: The topology of a network.
Mapper: A node on a SAN/LAN that has the map and an RT for that network. It is expected that the mapper dynamically updates the map and the RT.
Multi-homed Node: A node with more than one network interface, where each interface has another address.
Node: Whatever can send and receive packets (e.g., a computer, an MPP, a software process, etc.)
Node structure: A C-struct (or equivalent) containing values for some attributes of a node.
Planned Transfer: Transfer of information, occurs after an initial phase in which the sender decides which Level-2 route to use for that transfer.
RCVF: The "Received From" set includes all the physical addresses through which an RT was disseminated, starting with the address of the mapper that created that RT.
Re-direct-message: A message that tells nodes which HR should be used in order to get to a certain remote address.
Router: The inter-SAN communication device
Security Context: A relationship between 2 (or more) nodes that defines how the nodes utilize security services to communicate securely.
Source: The node that created a packet.
Source-Route: A Level-2 route that is chosen for a packet by its source.
Symbol: Data preceding the EEP header of a PktWay message,

interleaving with the L2RHs.

Twin-HR: Two HRs are twins if they both are parts of the same inter-SAN router.
Wormhole-routing: (aka cut-thru routing) forwarding packets out of switches as soon as possible, without storing that entire packet in the switch (unlike Stop-and-forward)
Zero-copy TCP: A TCP system that copies data directly between the user area and the network device, bypassing OS copies

13. Appendix-C: Acronyms and Abbreviations

0bNNNN	The binary number NNNN (e.g., 0b0100 is 4-decimal)
0xNNNN	The hexadecimal number NNNN (e.g., 0x0100 is 256-decimal)
8B	8 byte (64 bits) entity
ADDR	The Address-record of RRP
APIn	Application/Program Interface
AT	Address Type
ATM	Asynchronous Transmission Mode
B	Byte (e.g., 4B)
b	bit (e.g., 32b)
BC	Byte Count (of parameters)
BER	Bit Error Rate
CAPA	The CAPAbility-record of RRP
CC	Capability Code
CSR	Common Source-Route
DA	Destination Address
DB	Data Block
DL	Data Length (in 8B words)
DSP	Digital Signal Processor
DT	Destination-Type
E	The Endianness field (in the EEP header)
e	The MSbit of E
EEP	End/End Protocol
EI	Error Indication
GP	General Purpose
GVL2	An RRP message, requesting L2 route to a given destination
GVRT	An RRP message asking an HR to give its routing tables
h	Optional header fields flag
HR	Half Router
HRTO	An RRP message asking which HR to use for a given destination
ID	Identification
IGMP	Internet Group Management Protocol
INFO	An RRP message providing information about nodes
IP	The Internet protocol
ISORM	The ISO Reference Model
L	Length field (exclusive of itself)
L2	Level-2 of the ISORM (Link)
L2RH	Level-2 Routing Header
L2SR	Source Route

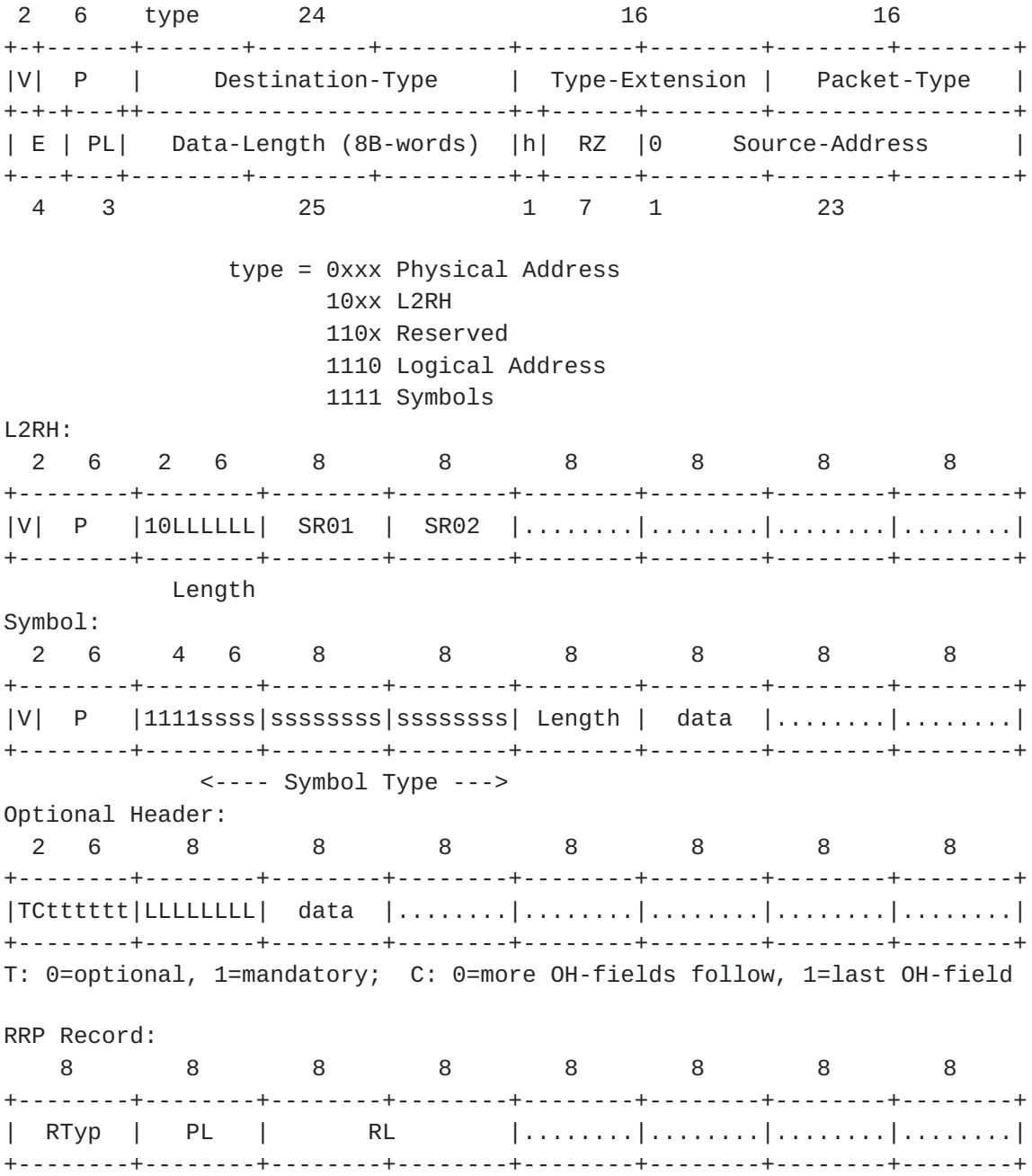
L3	Level-3 of the ISORM (Network)
LA	Logical Address
LADR	The Logical-addresses-record of RRP

LAN	Local Area Network
LRT	Local Routing Table
LSbit	Least Significant bit
LSbyte	Least Significant byte
MAC	Message Authentication Code / Media Access Control
MPI	Message Passing Interface
MPP	Massively Parallel Processing system
MSbit	Most Significant bit
MSbyte	Most Significant byte
MSU	Mississippi State University
MTU	Maximum Transmission Unit
MTUR	The MTU-record of RRP
M/C	Multicast
NAME	The name-record of RRP
NFS	Network File Server
OH	Optional Header field
OH-TYPE	The Type of an Optional Header field
OT	Optional Trailer field
P	The Priority field
PAD	Padding After Data
PBD	Padding Before Data
PCI	The Peripheral Component Interconnect "standard"
PH	PacketWay Header
PL	Padding Length (always in bytes)
PPP	The Point-to-Point Protocol
PROM	Programmable ROM (Read-Only-Memory)
PT	Packet Type (2B)
PVM	Parallel Virtual Machine
PW	The Myrinet Packet Type assigned to PktWay (PW=0x0300)
Q	Quality (of a path)
RCVF	Received-From list, or the Received-From record of RRP
RDRC	A re-direct message of RRP
RH	Routing Header
RID	Record ID
RL	Record Length (in 8B-words)
RRP	Router/Router Protocol
RT-hd	RT (Routing Table) header
RT	Routing Table
RTBL	An RRP message proving a Routing Table
RTHD	The Routing-Table-Header record of RRP
RTyp	RRP's Record Type
RZ	The Reserved field (in the EEP header)
SA	Source Address
SAN	System Area Network
SAN-ID	The 24-bit PktWay-address of a SAN
SAR	Segmentation and Reassembly
SN	Serial Number
SNID	SAN-ID
SNMP	Simple Network Management Protocol

SR	Source Route (always at Level-2)
SRQR	The Source-Route-and-Q-record of RRP
ST	Symbol Type

TAIL	PacketWay EEP Trailer
TE	Type Extension (2B)
TELL	RRP message requesting INFO about a partially specified node
UNK	Unknown
V	Version
WRU?	An RRP message asking its recipient to identify itself
XRT	External Routing Table
xxxx	A padding byte

14. Appendix-4: PktWay at a Glance (aka "The Cheat-Sheet")



```
RRP-messages: GVL2, L2SR, RDRC, TELL, INFO, HRTO, WRU?, GVRT, RTBL;
        RTyp: ADDR, NAME, CAPA, LADR, SRQR, MTUR, RCVF, RTHD;
```

15. Security Considerations

This RFC raises no security issues. PktWay is designed to work in clusters to which the access may be as controlled as needed.

PktWay has a security applique for securing the comminucation between classified/sensitive clusters, even when non-secure clusters and non-secure communication facilities have to be used. This applique uses cryptographic methods and equipment. More about that applique may be found in "Proposed Specification for Security Extensions to the PacketWay Protocol" {http://WWW.ERC.MsState.Edu/labs/hpcl/packetway/secure.txt}.

At the presence of security threats such applique should be used

16. Editor's Address

Danny Cohen
Myricom, Inc.
325 N. Santa Anita Ave
Arcadia, CA 91006

Phone: 626-821-5555
Fax: 626-821-5316
Email: Cohen@myri.com