Network Working Group INTERNET-DRAFT Q. Xie Motorola R. R. Stewart Cisco Systems

expires in six months

November 15,2000

Enpoint Name Resolution Protocol (enrp) <draft-xie-rserpool-enrp-01.txt>

Status of This Memo

This document is an Internet-Draft and is in full conformance with all provisions of <u>Section 10 of RFC 2026</u>. 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.

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/lid-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Abstract

Endpoint Name Resolution Protocol (ENRP) is designed to work in conjunction with Aggregate Server Access Protocol (ASAP) [<u>ASAP</u>]. ENRP when combined with ASAP provdes a high availablitity data transfer mechanism over IP networks. ASAP uses a name-based addressing model which isolates a logical communication endpoint from its IP address(es), thus effectively eliminating the binding between the communication endpoint and its physical IP address(es) which normally constitutes a single point of failure.

The high availablitity server pooling is gained by combining the two parts, namely ASAP and the Endpoint Name Resolution Part (ENRP). ASAP provides the user interface for name to address translation, load sharing management, and fault management. ENRP defines the fault tolerant name translation service.

Table Of Contents

<TBD>

1. Introduction

ENRP is designed to provide a fully distributed fault-tolerant real-time translation service that maps a name to a set of transport addresses pointing to a specific group of networked communication endpoints registered under that name. ENRP employs a client-server model with which an ENRP server will respond to the name translation service requests from endpoint clients on both the local host and remote hosts.

This document defines ENRP client-server interface and the ENRP server functionalities, including the establishment and management of a fully distributed fault-tolerant endpoint name space.

1.1 Motivation

In this section, we will discuss the motivation for developing ASAP. Our discussion will be focused on the analysis of the inadequateness of two existing technologies, namely CORBA and DNS, in providing solutions to fault-tolerance design of IP distributed applications.

<u>1.1.1</u> CORBA and Its Limitations

Often referred to as a Distributed Processing Environment (DPE), CORBA was mainly desinged to provide location transparency for distributed applications. However, the following limitations may exist when applying CORBA to the design of real time fault-tolerant system:

 CORBA is traditionally weak in fault tolerance. The recent development of a high availablitity version of CORBA by OMG is perhaps a step in the right direction towards fixing this weakness. Nevertheless, the maturity, implementablity, and real-time performance of the design is yet to be proven.

[Editor's Note: the fault tolerance mechanism being developed by

OMG for CORBA bears quite some similarities to ASAP.]

- 2) CORBA's distribution model encourages an object-based view, i.e., each communication endpoint is normally an object. We consider this kind of granularity too fine to be efficient and effective for designing real-time fault-tolerant applications.
- 3) CORBA in general has a large signature that makes the use of it a challenge in real-time environments. Small devices with limited memory and CPU resource (e.g., H323 or SIP terminals) will find CORBA hard to fit in.
- 4) CORBA uses TCP as its transport (Note, some effort is currently underway to separate CORBA from its transport). This makes CORBA suffer from the same limitations of TCP in terms of real-time and fault-tolerance performance.
- 5) CORBA has long lacked easily usable support for the asynchronous communication model, and this may be an issue in many applications. An apparently improved API for asynchronous communication has been added to the CORBA standards only recently, and many, if not most, CORBA implementations still do not support it. There is as yet insufficient user experience with it to make conclusions regarding this new feature's usability.

1.1.2 DNS and Its Limitations

Undoubtedly DNS is the best-known and proven IP namespace mechanism. However, namespace function alone does NOT provide real time fault-tolerance solution. Other mechanisms and procedures, such as server process failure detection, back-up server control and selection, fast fail-over/switch-over, load balancing, etc., also play crucial roles. These functions are supported by ASAP but not by DNS DNS provides a loose binding where as ASAP and ENRP are designed to provide a tight binding.

As will be further elaborated later in this document, the fault tolerant design for server pools is made up in two parts, namely ASAP and ENRP, where ENRP defines a light-weight yet highly efficient namespace mechanism optimized for building real time fault-tolerant applications. Nevertheless, ASAP does not restrict itself to ENRP for namespace services. In fact, it is not only feasible but also desirable in the future to generalize ASAP design so that the ENRP can provide a generic interface that is capable of inter-working with different namespace, including DNS, at the ASAP implementor's choice.

In the following, we list some limitations of the current DNS namespace capability when compared to that of ENRP:

1) DNS name registration and translation services have been primarily

optimized for host names. But we consider namespace services optimized for process names or endpoint names more appropriate and efficient for supporting real time server-level fault-tolerance applications.

- 2) DNS is primarily passive. It provides a query/response database that allows one to find information, but does NOT provide monitoring of hosts or processes to assure consistency within its database.
- 3) DNS has dynamic extensions but is not designed around the dynamic fast changing process address space that is typical to real time distributed applications.

It has also been suggested that ASAP be extended to work with DNS to bridge multiple ASAP planes and provide an "inter-ASAP-Domain" bridging function.

1.2 Definitions

This document uses the following terms:

Operation scope --- the part of the network visible by ENRP.

ASAP Endpoint --- a logical entity in the operation scope which implements the ASAP stack and is capable of sending and receiving messages.

ASAP Node --- a host machine in the network which contains one or more ASAP endpoints.

ASAP Plane or ASAP operational domain --- A relam of tight binding controlled by a set of one or more ENRP deamons. A process or application entitiy registers its name within this "plane" and is peroidically checked for sanity. Note that a plane does not imply geographical locality.

Endpoint name --- the registered tag of a ASAP endpoint, consisting of a NULL terminated ASCII string of fixed length.

Named group --- a group of ASAP endpoints sharing the same endpoint name in the name space.

Endpoint handle --- a logical pointer, consisting of a name and the primary destination transport address to a particular endpoint in a named group.

ENRP client --- a ASAP endpoint using ENRP to obtain name translation and other related services. In this document, the term "ASAP endpoint" is exchangeable with "ENRP client", unless otherwise stated. ENRP maintenance client --- a ASAP endpoint that has the additional capability of exchanging ENRP maintenance messages with an ENRP server in order to perform certain maintenance functions.

ENRP server --- a server program running on a node that manages the name space collectively with its peer ENRP servers and replies to the service requests from any ENRP client.

Home ENRP server --- the ENRP server to which an endpoint currently belongs. Endpoints normally choose the ENRP server on their local host as the home ENRP server, if one exists. An endpoint shall only have one home ENRP server at any given time, and both the endpoint and the server shall keep track of this master/slave relationship between them.

ENRP server takeover --- the event that a remote ENRP server takes the ownership of all the ENRP endpoints on a node and becomes their home server.

Caretaker ENRP server --- The ENRP server on a remote node which takes ownership of all endpoints on the local node because of the absence of an active local server.

ENRP client channel --- the communication channel through which a ASAP endpoint requests for ENRP service. The ENRP client channel is usually defined by the transport address of the home server and a well known port number.

ENRP server channel --- defined by a well known multicast IP address and a well known port number. All ENRP servers in an operation scope can communicate with one another through this channel. Endpoints are also allowed to communicate on this channel occasionally.

ENRP name domain --- defined by the combination of the ENRP client channel and the ENRP server channel in the operation scope.

Network Byte Order: Most significant byte first, a.k.a Big Endian.

1.3 Protocol overview

At startup, each endpoint in a ASAP operational domain registers its name to in name space. Here a name is defined as a NULL terminated ASCII string of fixed length.

When sending a message, the sender addresses the receiver endpoint by its name and passes the message to its ASAP layer. The ASAP layer, with the help from the ENRP name space daemon server(s), translates the name to a valid transport address (or a list of transport addresses if the receiver is multi-homed) and sends the message to the receiving endpoint. The following diagram illustrates the components of ASAP and their relationships.

```
Figure 1.
```

ENRP Server	Data Sender (ASAP-user) ++	Data Receiver (ASAP-user) ++
	ASAP /</td <td>> ASAP </td>	> ASAP
++	+	+
<> ENRP <	> ENRP	ENRP
To peer ++	++	++
ENRP server SCTP	SCTP	SCTP
++	++	++
IP	IP	IP
++	++	++

Multiple endpoints can register themselves under the same name. In that case they will be treated as a receiver pool, and ASAP, when sending a message addressed to that name, will use a predefined load-sharing policy to determine which endpoint(s) in the pool to send (or address) the message.

ASAP design has a high emphasis on seamless support of "server pooling", high availability, dynamic scalability, and close-to-real-time name translation.

In particular, ASAP can be characterized by:

- A) Seamless support of "server pooling" --- ASAP allows multiple servers to register under the same name. It also allows servers to be dynamically add to or removed from a server pool without any reconfiguration.
- B) Support automatic receiver "fail-over" --- when the chosen message receiver fails, ASAP, with pre-stated permission from its upper layer, can automatically re-direct the message to an alternative server under the same name if one exists.
- C) Transaction management by nickname or "association handle" --- this is to allow a continuous transaction or session consisting of multiple interactions be held between a client endpoint and and one particular server in a server pool.

Note: For details on ASAP please see [ASAP].

D) Fully distributed name space --- For achieving a high degree of fault tolerance and operation efficiency, the ENRP daemons which provide name translation service and the name space data are distributed across the operational scope of the network.

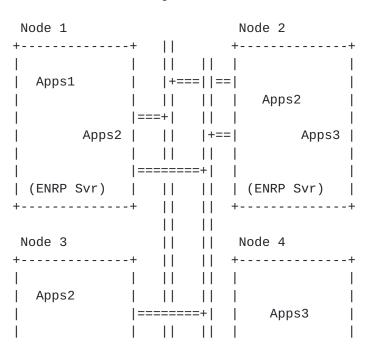
ENRP daemon servers can be added to or removed from the operation scope dynamically, without interrupting the current name translation service.

For example, a node may be originally configured to operate without a local ENRP server. When the load condition changes, one can start a new ENRP server on that node to increase the operation capacity. The new ENRP server will automatically integrate itself with the existing ENRP server(s) in the scope.

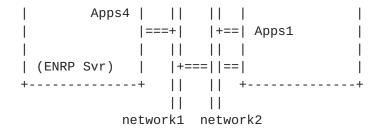
Similarly, when an ENRP server becomes unavailable for service (e.g., being intentionally shutdown, or suffered failure), its ASAP clients will be automatically taken-over by a remote ENRP server and continuously have ENRP services provided.

E) Network failure detection and automatic recovery --- In the case when a major network failure breaks the operation scope into isolated communication islands, the name translation service will survive and continue inside each island so long as there is one or more ENRP servers present in that island. Endpoints inside each island will still continue to be able to communicate with each other. Moreover, when the network recovers, the isolated ENRP servers will re-discover each other and re-integrate the name space back into its original form.

Figure 2. shows an example of distributed applications operating in a scope that is connected by a pair of redundant networks.







In this example, there are four nodes in the scope, as shown in Figure **2**. On Node 1, Node 2, and Node 3, there is an ENRP server running. On each of the nodes, there are also some applications running. Each application has a registered name in the name space collectively managed by the three ENRP servers. In the example, the registered names are "Apps1", "Apps2", "Apps3", and "Apps4". Some of the applications (Apps1, Apps2, and Apps3) are distributed as server pools.

When sending messages to each other, the sender application simply addresses the recipient by its registered name. The ASAP layer inside the sender application will query its home ENRP server to obtain the transport address(es) and load control information of the recipient before sending the message out.

Also note in the example, there is no ENRP server on Node 4. But the applications on Node 4 will be served by one of the ENRP servers on other nodes.

<u>1.4</u> Organization of this document

Chapter 2 we give the details of the ENRP interface. ENRP defines the messaging structure and relevant rules for communications between an ASAP endpoint and an ENRP server. This chapter discusses how ENRP maintains the name space in a high availablitity manner. Chapter 3 defines the message format and structures used by ENRP (in conjunction with those used by ASAP). Chapter 4 provides settable protocol values.

1.5 Scope of ENRP

The scope of the ASAP/ENRP is NOT Internet wide. The namespace is neither hierarchical nor arbitrarily large like DNS. We propose a flat peer-to-peer model. Pools of servers will exist in different administrative domains. For example, suppose I want to use ASAP/ENRP. First, the PU will use DNS to contact an ENRP server. Suppose a PU in North America and wish to contact the server pool in Japan instead of North America. The PU would use DNS to get the IP address of the Japanese server pool domain, that is, the address of an ENRP server('s) in Japan.

2. The ENRP interface

This section discusses the messages and procedures for communicating between the ASAP layer of a ASAP endpoint and an ENRP name space server, as well as that between peer ENRP servers.

2.1 Functional Summary

In this section, we discuss the functions defined by ENRP. The functions are divided into three groups, namely the basic ENRP operations, fault management, and control and maintenance functions.

Most of the ENRP operations involve message exchanges between an ENRP server and a ASAP endpoint, as well as message exchanges between the ENRP server and its peers. Some of the ENRP message formats are also found in [ASAP]

2.1.1 Basic ENRP Operations

2.1.1.1 Endpoint Registration

ENRP server <-> endpoint:

A ASAP endpoint shall send a REGISTRATION message, over the ENRP client channel, to its home ENRP server, in order to register itself with the name space.

In the REGISTRATION message, the endpoint shall indicate its name in the form of a character string, network access information (e.g., a list of valid transport addresses with which the endpoint can be reached), and load control information.

The ENRP server shall handle the REGISTRATION message following the rules listed below:

- o If the name does not exist in the name space, the ENRP server shall create the name and add the new endpoint under that name.
- o If the name already exists in the name space, the requesting endpoint shall be added under the same name and be made a member of the named group.
- o If both the name and the requesting endpoint already exist in the name space, i.e., a case of duplicated registration, the ENRP server shall grant the request without taking any further actions.
- o The ENRP server may reject the registration due to reasons such as invalid values, lack of resource, etc.

In all the above cases, if the REGISTRATION request is granted, the ENRP server shall assume the ownership of the requesting endpoint.

In response, the home ENRP server shall reply to the requesting endpoint with a REGISTRATION_RESPONSE message, and shall indicate in the message body whether the registration is granted or rejected.

ENRP server <-> peers:

If the registration request is not a duplicate and is granted, the home ENRP server shall take the name space modification action described in <u>section 3.1.1.8</u>??. Otherwise, no message shall be exchanged with its peers.

2.1.1.2 Endpoint De-registration

ENRP server <-> endpoint:

A ASAP endpoint shall send a DEREGISTRATION message, over the ENRP client channel, to its home ENRP server in order to remove itself from the name space.

If the endpoint is the last one under that name in the name space the home ENRP server shall remove the name from its space as well.

The ENRP server may reject the de-registration request due to reasons such as invalid parameters, etc.

In response, the home ENRP server shall send a REGISTRATION_RESPONSE message to the endpoint, and shall indicate in the message body whether the request is granted or rejected.

It should be noted that de-registration does not stop the ASAP endpoint from sending or receiving messages. It only means that other ASAP endpoints will no longer be able to send message to that endpoint by name.

ENRP server <-> peers:

Once the de-registration request is granted and the endpoint removed from its local copy of the name space, the home ENRP server shall take the name space modification action described in <u>section 2.1.1.9</u>.

2.1.1.3 Name Translation

ENRP server <-> endpoint:

An endpoint shall send a NAME_REQUEST messages to its home ENRP server to get a name translation service. In the NAME_REQUEST message, the endpoint shall include the name it wants to be translated.

If the name exits in the name space, the ENRP server shall send back a

NAME_INFORMATION message that shall carry all information of the ASAP endpoint(s) currently registered under that name, including current load control policy of the group, policy value of each endpoint in the group, and a list of transport addresses for each endpoint in the group with which the endpoint can be reached, etc.

If the name does not exist in the name space, the ENRP server shall respond with a NAME_UNKNOWN message.

ENRP server <-> peers:

None.

2.1.1.4 Server Name Space Update

This includes a set of update operations used by an ENRP server to inform its peer(s) the addition a new ASAP endpoint, removal of an existing ASAP endpoint, change property of a named group, etc.

2.1.1.4.1 Addition of a New ASAP Endpoint

When a new ASAP endpoint is granted registration to the name space, the home ENRP server uses this procedure to inform all its peers.

ENRP server <-> endpoint:

None:

ENRP server <-> peers:

An ENRP server shall multicast over the ENRP server channel a PEER_NAME_UPDATE message with the appropriate flag set to indicate to its peers about the addition of the new endpoint to the name space.

Upon the reception of this PEER_NAME_UPDATE message, each of the peer ENRP servers shall take the following actions:

- o If the name does not exist in its local copy of the name space, the peer ENRP server shall create the name and add the new endpoint under that name in its local name space copy, along with other attributes about the endpoint carried in the message.
- o If the name already exists in the peer server's local copy of the name space, the new endpoint endpoint shall be added as a new member of the named group.
- o If both the same ASAP endpoint already exists in the named group in the local copy of the name space of the peer, the peer ENRP server shall take no actions.

After adding the endpoint into its local copy of name space, the peer ENRP server shall check if this endpoint is located on the same host as the peer ENRP server itself does. If so, the peer ENRP server shall assume the ownership of the endpoint, and take the ?? actions described in section 2.1.1.12??.

2.1.1.4.2 Removal of a ASAP Endpoint

This procedure is used by an ENRP server to inform its peer(s) to remove an existing ASAP endpoint, regardless of the ownership of the endpoint.

ENRP server <-> endpoint:

None:

ENRP server <-> peers:

The ENRP server shall multicast over the ENRP server channel a PEER_NAME_UPDATE message with the appropriate flag set to instruct its peers to remove of the endpoint from their local copy of the name space.

On the reception of this PEER_NAME_UPDATE message, each peer ENRP server shall find and remove the ASAP endpoint from its local copy of the name space regardless whether or not it has ownership on the endpoint.

2.1.1.4.3 Removal of a ASAP Endpoint with no Ownership

This operation is used by an ENRP server to instruct its peers to remove an existing ASAP endpoint which the peer does not have an ownership on.

ENRP server <-> endpoint:

None:

ENRP server <-> peers:

An ENRP server shall multicast over the ENRP server channel a PEER_NAME_UPDATE message with the appropriate flag set to instruct its peers to remove the specified endpoint from its local copy of the name space IF the peer does not have ownership on the endpoint.

On the reception of this PEER_NAME_UPDATE message, a peer ENRP server shall find and remove the endpoint from its local copy of the name space only if the peer server does not own this endpoint.

2.1.1.4.4 Update Endpoint Attributes

This operation is used by an ENRP server to inform its peers to update the attributes of an existing ASAP endpoint.

ENRP server <-> endpoint:

None:

ENRP server <-> peers:

An ENRP server shall multicast over the ENRP server channel a PEER_NAME_UPDATE message with the appropriate flag set to instruct its peers to replace the attributes of an existing ASAP endpoint in its local copy of the name space.

On the reception of this PEER_NAME_UPDATE message, a peer ENRP server shall replace the attributes of the existing endpoint with the new information carried in the message if the endpoint exists in its local copy of the name space. If the specified endpoint is not found in its local name space copy, the peer server shall add the endpoint following the steps in <u>Section 2.1.1.4.1</u>??.

2.1.1.4.5 Claim Endpoint Ownership

This operation is used by an ENRP server to claim the ownership on a specific endpoint and to inform its peers about its claim.

ENRP server <-> endpoint:

An ENRP server shall send an ENDPOINT_KEEP_ALIVE message to the endpoint. This message will cause the endpoint to adopt this ENRP server as its new home ENRP server (see <u>Section 2.5.3</u>).

ENRP server <-> peers:

An ENRP server shall multicast over the ENRP server channel a PEER_NAME_UPDATE message with the appropriate flag set to inform its peers that it has taken the ownership of the specified endpoint.

Upon the reception of this PEER_NAME_UPDATE message, a peer server shall check whether it is the current owner of the endpoint. If so, this peer server shall relinquish its ownership on that endpoint. Otherwise, no action is needed.

2.1.1.4.6 Report Endpoint Failure

This operation is used by an ENRP server to warn its peers that it has noticed a potentially unreachable endpoint that the server does not have ownership on. ENRP server <-> endpoint:

None:

ENRP server <-> peers:

An ENRP server shall multicast over the ENRP server channel a PEER_NAME_UPDATE message with the appropriate flag set to indicate that the specified ASAP endpoint is potentially unreachable.

On the reception of this message, each peer ENRP server shall check whether it owns the specified endpoint. If it does, the peer server shall increase the <endpoint-report-failures> counter of the specified endpoint by 1. If the value of the <endpoint-report-failures> counter has exceeded the protocol parameter Max-Endpoint-Report-Failures, the peer server shall remove the endpoint from its local name space and take actions described in <u>Section 2.1.1.4.3</u>. If the peer server does not own the specified endpoint, it shall take no action.

2.1.1.5 Endpoint Change Policy Value

A ASAP endpoint can modify its policy value at any time. Depending on the current number of members in the named group and the server pooling policy, this operation allows the ASAP endpoint to control its share of inbound messages received within the named group dynamically (also see <u>Section 2.1.5.1</u> for more on load control).

ENRP server <-> endpoint:

A ASAP endpoint shall send an UPDATE_POLICY_VALUE message over the ENRP client channel to its home ENRP server in order to modify its policy value. The new policy value shall be indicated in the message.

Upon the reception of this UPDATE_POLICY_VALUE message, the home ENRP server shall replace the policy value of that endpoint in its local copy of the name space with the new value indicated in the message.

ENRP server <-> peers:

If the update on its local copy of the name space is successful, the home ENRP server shall take the Server Name Space Update actions as described in <u>Section 2.1.1.4.4</u>.

2.1.1.6 Server Down Load Name Space from a Peer

This operation allows an ENRP server to request and receive a copy of a specific portion of the name space from one of its peer ENRP servers. This is useful for a newly started ENRP server to initiate its local copy of the name space, or for correcting name space inconsistency.

ENRP server <-> endpoint:

None.

ENRP server <-> peers:

An ENRP server shall first send a PEER_NAME_TABLE_REQUEST message directly to one of its peers. In the message, it shall indicate which portion of the name space is requested.

Upon the reception of this message, the peer server shall initiate a download session in which the requested portion of the name space shall be sent to the requesting ENRP server in one or more PEER_NAME_TABLE_RESPONSE messages.

If the sending ENRP server determines that multiple PEER_NAME_TABLE_RESPONSE messages are needed for the session, it shall set the appropriate flag in each PEER_NAME_TABLE_RESPONSE message to inform the receiving ENRP server whether or not the data in this message is the last piece of the transfer.

Every time the requesting ENRP server receives a PEER_NAME_TABLE_RESPONSE message, it shall transfer the data entries carried in the message into its local name space database, and then check whether or not the data in this message is the last piece to be transfered. If more data transfer is indicated, the requesting ENRP server shall send another PEER_NAME_TABLE_REQUEST message to the same peer to prompt for the next piece.

When transferring the data entries from the PEER_NAME_TABLE_RESPONSE message into its local name space database, the requesting ENRP server shall follow the same procedures as described in <u>section 2.1.1.4.1</u> when parsing through the endpoints carrying in the message one by one.

2.1.1.7 Server Monitor Peer Status

ENRP server <-> endpoint:

None:

ENRP server <-> peers:

An ENRP server shall keep a record on the status of each of its peers.

If a message of any type is received from a peer, the server shall update that peers status as <active>. If a message of any type is received from a peer previously unknown to this server, i.e., a new peer, the server shall create a record for the new peer and mark the new peer as <active>.

2.1.1.8 Server Down Load Peer List

This operation allows an ENRP server to request from a peer server a copy of its internal peer list. This is useful for a new ENRP server to initiate its own peer list at startup.

ENRP server <-> endpoint:

None.

ENRP server <-> peers:

An ENRP server shall send a PEER_LIST_REQUEST message to a peer to request a copy of its peer list.

Upon the reception of this message, the peer server shall reply with a PEER_LIST_RESPONSE message and include in the message body a copy of its internal peer list, if the peer itself is in operational state.

If the peer itself is in the process of startup, it shall response with a PEER_LIST_RESPONSE message but set the appropriate flag to indicate that it can not grant the PEER_LIST_REQUEST. In such a case, the requesting ENRP server shall select another peer and repeat the peer list request with the new peer at a later time.

2.1.1.9 Endpoint Initialization

At startup, a ASAP endpoint shall always assume the existence of a local ENRP server on the local host and mark it as its home ENRP server, and initiate the registration procedure described in 2.1.1.1??.

2.1.1.10 Server Initialization

At startup, before getting into service, an ENRP server (initiating server) shall multicast a PEER_PRESENCE message with reply required flag set over the ENRP server channel, in order to inform any other active peers in the operation scope about its presence.

Upon the reception of this message, a peer shall send a PEER_PRESENCE without reply required flag back to the initiating server, in order to help the initiating server to build its peer list.

If no response to its PEER_PRESENCE message are received, the initiating server shall assume that it is alone in the operation scope and shall mark the initialization process as completed.

If there are responses to its PEER_PRESENCE message, the initiating server shall then take the actions described in 2.1.1.8 to request a

peer list from one of the peers that have responded.

Upon the reception of the PEER_LIST_RESPONSE message from that peer, the initiating server shall use the information carried in the message to build a complete peer list, including both active and inactive peers in the operation scope.

Then, the initiating server shall perform a name database download, as described in 2.1.1.6, with each of the active peers on the peer list, indicating that the portion of the name database to download shall only include the endpoints owned by that peer.

Moreover, the initiating server shall also pick one of the active peer and request to that peer for a download of the table of remote endpoints.

2.1.2 Fault Management Operations

The following operations are used to detect and recover from various system faults.

2.1.2.1 Detect and Report Unreachable Endpoint

Two mechanisms exist to detect and report an unreachable ASAP endpoint:

1) Home ENRP server periodic sanity check

An ENRP server shall send, in every <Endpoint-sanity-cycle> seconds, an ENDPOINT_KEEP_ALIVE message to each of the endpoints it owns, and shall keep the number of consecutive failed send attempts in the counter <Endpoint-sanity-failures> of that endpoint.

If the value of <Endpoint-sanity-failures> of an endpoint exceeds the pre-set threshold Max-endpoint-sanity-failures, the home ENRP server shall remove the endpoint from its copy of the name database and take the actions described in <u>section 2.1.1.4.3</u> to inform its peers.

The handling of the ENDPOINT_KEEP_ALIVE message by the endpoint is described in <u>Section 2.5.3</u>??.

2) Detection by peer endpoints

Whenever a ASAP endpoint finds a peer unreachable (e.g., via an SCTP SEND.FAILURE Notification, see <u>Section 2.2.5</u>??), the endpoint shall send an ENDPOINT_UNREACHABLE message over the ENRP client channel to its home ENRP server. The message shall contain one of the transport addresses of the unreachable peer and have the severity flag set to NORMAL_REPORT.

Upon the reception of this message, the home ENRP server shall first check whether it owns the unreachable endpoint. If not, the server shall take the actions described in <u>section 2.1.1.4.6</u>??.

Otherwise, the server shall increase the <Endpoint-report-failures> counter of the unreachable endpoint by 1. If the value of the <Endpoint-report-failures> counter has exceeded Max-endpoint-report-failures, the server shall remove the endpoint from its name database and take actions described in 2.1.1.4.3??.

2.1.2.2 ENRP Server Heartbeat

An ENRP server shall multicast, in every <Peer-heartbeat-cycle> seconds, a PEER_PRESENCE message over the ENRP server channel to inform its peers that it is still operational. In the PEER_PRESENCE message, the sending ENRP server shall set the appropriate flag to indicate that no reply is required.

>From time to time, an ENRP server may also send a point-to-point PEER_PRESENCE message to a specific peer server, with the flag setting in the message indicates that a reply is required. In such a case, the peer server shall immediately respond to the sender with its own point-to-point PEER_PRESENCE message, and shall indicate in the message that no reply is required.

2.1.2.3 ENRP Server Hunt

An endpoint shall initiate the following home server hunt procedure if it fails to send to, or times out on a service request with its current home server.

In the home server hunt procedure, the endpoint shall multicast a SERVER_HUNT message over the ENRP client channel, and shall repeat sending this message every <Timeout-server-hunt> seconds until a SERVER_HUNT_RESPONSE message is received from an ENRP server. Each time the 'Timeout-server-hunt' timer expires the criticality should be raised (initially criticality should be set to LOW_CRITICALITY).

Then the endpoint shall pick one of the servers that have responded as its new home server, and continue the service request with that server.

Upon the reception of the SERVER_HUNT message, a server shall reply to the endpoint with a SERVER_HUNT_RESPONSE message, unless:

- its peer status information indicates that there is a caretaker server other than itself for the node where the endpoint is from, AND
- 2) the criticality flag in the SERVER_HUNT message is not

HIGH_CRITICALITY.

2.1.2.4 ENRP Server Detect and Take-over Inactive Peer

An ENRP server shall keep track the time when the last message (multicast or point-to-point) was received from each known peer.

If a peer has not been heard for more than Max-time-last-heard, the ENRP server shall send a point-to-point PEER_PRESENCE with reply request to that peer.

If the send fails or the peer does not reply after Max-time-no-response seconds, the ENRP server shall initiate the following server take-over procedures:

1) Initiate Server Take-over Arbitration

The ENRP server (the initiating server) shall initiate a take-over arbitration on the inactive peer (the target server) by multicasting a TAKEOVER_INITIATE message over the ENRP server channel. In the message, the initiating server shall specify the identification of the target server.

After multicasting the TAKEOVER_INITIATE message, the initiating server shall wait for a TAKEOVER_INITIATE_RESPONSE message from each of its active peers.

Upon the reception of this message, other peer servers shall take the following actions accordingly:

- o If the peer server finds that itself is the target server indicated in the TAKEOVER_INITIATE message, it shall immediately multicast a PEER_PRESENCE message over the ENRP server channel in an attempt of stopping the take-over process.
- o If the peer server finds that itself has also initiated a take-over process on the same target server and its IP address is smaller in value than that of the sender of the TAKEOVER_INITIATE message, it shall abort its own take-over process.
- o Peers other than the target peer and the peer that is taking-over shall mark the target server as <inactive> and mark the initiating server as the caretaker of the target server and reply to the initiating server with a TAKEOVER_INITIATE_RESPONSE message.

Once it has received TAKEOVER_INITIATE_RESPONSE message from all of its active peers, the initiating server shall consider it won the arbitration and shall then take the actions in 2) in order to complete the take-over.

However, if it receives a PEER_PRESENCE from the target server at any

point of the take-over, the initiating server shall immediately abort the take-over process and re-mark the target server as <active>.

2) Take-over the target peer server

An ENRP server shall multicast a TAKEOVER_PEER_SERVER message over the ENRP server channel in order to inform all its peers about the take-over. In the message, identification of the inactive peer server targeted for the take-over shall be included.

The server shall mark the target server as <inactive> and mark itself as the caretaker of the target server.

Then it shall assume ownership on each of the endpoints originally owned by the target server.

The server shall also check whether there are any other inactive peers which has designated the target server as their caretaker. The server shall perform the above take-over procedure on each one of those inactive peers as well.

2.1.2.5 Register Homeless Endpoints

When an ENRP server receives a REGISTRATION message from an endpoint located on a remote node, it shall always accept and grant the registration, unless its peer status information indicates that the peer on that node is inactive and a caretaker other than itself exists for that node. In that case, the server shall reject the registration and take no further actions.

If the server has no record about the peer on that node, the server shall grant the registration and then create a record about that peer, mark it as inactive, and initiate a take-over procedure on it, as described in 2.1.2.4??.

2.1.3 Maintenance Operations

The following operations are used by an ENRP maintenance client to monitor the name space data and perform maintenances on ENRP servers in an operation scope.

2.1.3.1 Forceful Removal of Endpoint

A maintenance endpoint shall send a ENDPOINT_UNREACHABLE message to an ENRP server, in order to force the removal of another endpoint from the name space. The message shall contain one of the transport addresses of the target endpoint and have the severity flag set to FINAL_REPORT.

Upon the reception of this message, the ENRP server shall immediately remove the target endpoint from its copy of the name database and take actions described in <u>Section 2.1.1.4.2</u>.

2.1.3.2 Dump Home Endpoint List

A maintenance endpoint shall send a SERVER_DUMP message with type flag set to HOME_LIST to a server, in order to require a copy of the information on all the endpoints owned by that server.

Upon receiving this message, the server shall response with a SERVER_DUMP_RESPONSE message with the type flag set to HOME_LIST to the maintenance endpoint. In the message body, the server shall include information on all the endpoints the server currently owns.

2.1.3.3 Dump Remote Endpoint List

A maintenance endpoint shall send a SERVER_DUMP message with type flag set to REMOTE_LIST to a server, in order to require a copy of the information on all the endpoints NOT owned by that server.

Upon receiving this message, the server shall response with a SERVER_DUMP_RESPONSE message with the type flag set to REMOTE_LIST to the maintenance endpoint. In the message body, the server shall include information on all the endpoints the server currently does NOT owns.

2.1.3.4 Dump Peer Server List

A maintenance endpoint shall send a SERVER_DUMP message with the type flag set to PEER_SERVER_LIST to a server, in order to require a copy of the peer list known to that server.

Upon receiving this message, the server shall response with a SERVER_DUMP_RESPONSE message with the type flag set to PEER_SERVER_LIST to the maintenance endpoint. In the message body, the server shall include information on all the peers that server currently knows.

<u>3</u> Message Summary

All messages as well as their fields described below shall be in Network Byte Order during transmission. For fields with a length bigger than 4 octets, a number in a pair of parentheses may follow the filed name to indicate the length of the field in number of octets.

3.1 Endpoint Entry

This field is used to represent a ASAP endpoint and the associated information, such as its transport address(es), load control, and other operational status information.

The field is defined to support endpoint with up to 8 different transport addresses.

0 1			2	3	3
0 1 2 3 4 5 6 7 8 9 0	12345	6 7 8 9	01234	567890	91
+-	+-+-+-+-	+-+-+	-+-+-+-+-	+ - + - + - + - + - +	+-+-+
	IP addr	ess #0			
+		+			+
	IP addr	ess #1			1
+		+			+
\backslash	\		Λ		\
/	/		/		/
\backslash	\		Λ		\
+		+			+
	IP addr	ess #7			1
+		+			· +
SCTP Port		1	Padding		1
+		+			+
Server Pooling Poli	cv	1	Policy Val	.ue	I
++	-	+	-		· +

The size of the endpoint entry is 40 octets.

3.2 PEER_NAME_TABLE_REQUEST message

Θ	1	:	2	3
0123	4 5 6 7 8 9 0 1 2 3 4 9	56789	0 1 2 3 4 5 6 7	8901
+-+-+-	-+	+ - + - + - + - + - •	+ - + - + - + - + - + - + - + -	+-+-+-+-+
	ENRP server message i	dentifier :	$\#1 = 0 \times 27047729$	
+		+		+
	ENRP server message i	dentifier :	$#2 = 0 \times 53829149$	
+		+		+
	Type =	0x102		
+		+		+
	sending serve	er's IP ad	dress	I
+		+		+
	sender's SCTP port		padding	
+		+		+
	receiving serv	ver's IP a	ddress	I
+		+		+
l r	receiver's SCTP port		padding	I
+		+		+
	Table ty	ype = (see	be⊥ow)	I
+		+		+

Note, the receiver's IP address and port do not need to be filled in if the message is being multicasted.

The requested table type shall take one of the following values: 0x1 --- HOME_LIST: endpoints owned by the server. 0x2 --- REMOTE_LIST: endpoint NOT owned by the server.

3.3 PEER_NAME_TABLE_RESPONSE message

0 1 2 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 ENRP server message identifier #1 = 0x27047729+-----+ ENRP server message identifier #2 = 0x53829149 -----+ Type = 0×103 +-----+ sender's IP address +----+ I sender's SCTP port padding +-----+ receiver's IP address +-----+ receiver's SCTP port padding +----+ Table type = (see below) +----+ More to send = (see below) -----+ number of names = n -----+ Name entry 1 (see below) / +----+ Name entry n (see below)

'Table type' shall take one of the following values: 0x1 --- HOME_LIST: endpoints owned by the server. 0x2 --- REMOTE_LIST: endpoint NOT owned by the server. 'More to send' flag shall be set to 0x1 if there are more name entries to be sent for the requested table type. Otherwise, it shall be set to 0x0.

Each 'Name entry' represents an endpoint and shall consist of the following:

3.4 PEER_LIST_REQUEST message

0 1 2 3 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 ENRP server message identifier #1 = 0x27047729 ENRP server message identifier #2 = 0x53829149 +-----Type = $0 \times 10b$ -----+ sender's IP address +----+ sender's SCTP port padding -----+ receiver's IP address -----+ receiver's SCTP port | padding ----+

The receiver's IP address and port do not need to be filled in if the message is being multicasted.

3.5 PEER_LIST_RESPONSE message

This message shall contain all the peer information of the sending server.

2 0 1 3 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 ENRP server message identifier #1 = 0x27047729 +-----+ ENRP server message identifier #2 = 0x53829149 +-----+ Type = $0 \times 10c$ -----+ sender's IP address +----+ padding senders SCTP port +----+ receiver's IP address +-----+ receiver's SCTP port padding +----+ responseIndication (see below) +-----+ number of peers = n-----+ Peer entry 1 (see below) +----+ Peer entry n (see below) +----+

The 'responseIndication' flag shall be set to 0x2 to indicate a rejection to the request, and no 'Peer entry' shall be attached if the request is rejected.

Otherwise, the 'responseIndication' flag shall be set to 0x1 and n 'Peer entries' attached.

Each 'Peer entry' shall consist of the following fields:

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1

	Peer's	IP address	
	Peer's SCTP port	padding	
		r's IP address	
+	caretaker's SCTP port	padding	

The peer's IP address and port number serve as the identification of that peer. If the peer is inactive, its caretaker's IP address and port number shall be filled in. Otherwise, the caretaker IP and port fields shall be set to zeros.

<u>3.6</u> PEER_NAME_UPDATE message

0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 +-+-+-++-++++++++++++++++++++++++++++			
<pre>++ ENRP server message identifier #2 = 0x53829149 ++</pre>	+ +		
Type = 0x104	I		
sender's IP address	+		
sender's SCTP port padding			
receiver's IP address	+ +		
receiver's SCTP port padding			
Endpoint name (32)			
 endpoint entry (40) 	 		
Update action (see below)	 +		

The receiver's IP address and port do not need to be filled in if the message is being multicasted.

- 0x1 --- DELETE_ENDPOINT: delete the named endpoint from the name database, if the receiving server owns the endpoint.
- 0x2 --- REMOVE_ENDPOINT: remove the named endpoint from the name database, regardless who owns the endpoint.
- 0x3 --- UPDATE_ENDPOINT: replace the endpoint's attributes with the new information carried in this message.
- 0x4 --- ENDPOINT_FAILURE: warn the receiver that the named endpoint is potentially unreachable.
- 0x5 --- CLAIM_ENDPOINT: inform the receiver that the sender has taken the ownership of the specified endpoint.

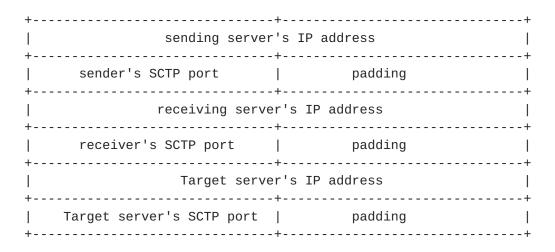
3.7 PEER_PRESENCE message

0 1 2 3 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 ENRP server message identifier #1 = 0x27047729+-----ENRP server message identifier #2 = 0x53829149+----+ Type = 0×100 +-----+ sender's IP address +----+ padding sender's SCTP port +-----+ receiver's IP address +-----+ receiver's SCTP port | padding +-----+ Reply required +-----+

The receiving server's IP address and port do not need to be filled in if the message is being multicasted.

'Reply required' shall be set to 0x1 if response to this message is required, otherwise set to 0x0.

3.8 TAKEOVER_INITIATE message



The receiving server's address and port do not need to be filled in if the message is being multicasted.

'Target server's IP address and port number must be supplied.

3.9 TAKEOVER_INITIATE_RESPONSE message

0 1	2	3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6	678901234567	8901
+-	+ - + - + - + - + - + - + - + - + - + -	+-+-+-+-+
ENRP server message ide)
+	+	+
ENRP server message ide)
+		+
Type = 0		I
	+	+
sending server	's IP address	
+	+	+
sender's SCTP port		
receiving server	•	+
+	+	۱ +
receiver's SCTP port	padding	
	+	· +
Target serve	r's IP address	1
+	+	· +
Target server's SCTP port	padding	1
+		·+

The receiving server's address and port do not need to be filled in if the message is being multicasted.

'Target server's IP address and port number must be supplied.

3.10 TAKEOVER_PEER_SERVER message

0 1 2 3 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 ENRP server message identifier #1 = 0x27047729+----+ ENRP server message identifier #2 = 0x53829149 +----+ Type = 0x108-----+ sending server's IP address sender's SCTP port padding +----+ receiving server's IP address -----+ receiver's SCTP port padding -----+ Target server's IP address -----+ Target server's SCTP port | padding +-----+

The receiving server's address and port do not need to be filled in if the message is being multicasted.

'Target server's IP address and port number must be supplied.

3.11 SERVER_DUMP message

0 2 1 3 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 ENRP endpoint message identifier #1 = 0x18038688 +----+ ENRP endpoint message identifier #2 = 0x77734683+----+ Type = 0x7-----+ Endpoint name (32) -----+ Dump Type (see below) -----+

The 'Dump Type' field shall take one of the following values:

- 0x0 --- HOME_LIST: dump a copy of the home endpoint portion of the name database of the server (i.e., endpoints owned by the server).
- 0x1 --- REMOTE_LIST: dump a copy of the remote endpoint portion of the

name database of the server (i.e., endpoints NOT owned by the server).

0x2 --- PEER_LIST: dump a copy of a list containing all the peers known to the server.

3.12 SERVER_DUMP_RESPONSE message

0 1 2 3 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 ENRP endpoint message identifier #1 = 0x18038688 -----+ ENRP endpoint message identifier #2 = 0x77734683 +-----+ Type = 0x8Endpoint name (32) Dump Type (see below) -----+ Number of Entries = n (see below) -----+ Dump entry 1 (see below) +-----+ Υ. -----+ Dump entry n (see below) ----+

The 'Dump Type' fields shall take the same values as defined in <u>Section 3.2.29</u>??.

If 'Dump Type' is HOME_LIST, or REMOTE_LIST, the 'Number of Entries' field shall be the number of endpoint entries carried in the message, and each 'Dump entry' field shall contain an endpoint entry and shall be defined as:

If 'Dump Type' is PEER_LIST, the 'Number of Entries' field shall be the number of peer entries carried in the message, and each 'Dump entry' field shall contain a peer entry and shall be defined as:

Θ 1 2 3 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 Peer's IP address +-----+ Peer's SCTP port padding +----+ Caretaker's IP address +----+ caretaker's SCTP port padding +----+

In a peer entry, the peer's IP address and port number serve as the identification of that peer. If the peer is inactive, its caretaker's IP address and port number shall be filled in. Otherwise, the caretaker IP and port fields shall be zeroed out.

4. Variables, Time Values, and Thresholds

The following is a summary of the variables, time values, and pre-set thresholds used in ASAP and ENRP protocol.

4.1 Variables

Endpoint-report-failures --- per endpoint; keeps the number of endpoint-unreachable reports concerning this endpoint.

Endpoint-sanity-failures --- per endpoint; keeps the number of failed sanity message send attempts concerning this endpoint.

Peer-server-last-heard --- per peer server; a time stamp on when the last message was received from this peer server.

4.2 Time values

Endpoint-sanity-cycle --- the period for a home ENRP server to start a new round of endpoint sanity check.

Peer-heartbeat-cycle ---the period for an ENRP server to send out a heart heat message.

T1-ENRPrequest - A timer started when a request is sent by ASAP to the ENRP server (providing application information is queued). Normally set to 15 seconds.

T2-registration - A timer started when sending a registration request to the local ENRP server, normally set to 30 seconds.

T3-registration-reattempt - If the registration cycle does not complete this timer is begun to restart the registration process. Normal value for this timer is 10 minutes.

T4-reregistration - This timer is started after successful registration into the ASAP name space and is used to cause a re-registration at a periodic interval. This timer is normally set to 10 minutes.

4.3 Thresholds

Max-endpoint-sanity-failures --- pre-set threshold for Endpoint-sanity-failures.

Max-endpoint-report-failures --- pre-set threshold for Endpoint-report-failures.

Max-time-last-heard --- pre-set threshold for Peer-last-heard.

Max-time-no-response --- pre-set threshold for a peer server to answer a PEER_PRESENCE message with reply required.

Timeout-registration --- pre-set threshold; how long an endpoint will wait for the REGISTRATION_RESPONSE from its home ENRP server.

Timeout-server-hunt --- pre-set threshold; how long an endpoint will wait for the REGISTRATION_RESPONSE from its home ENRP server.

num-of-serverhunts - The current count of server hunt messages that have been transmitted.

registration-count - The current count of attempted registrations.

max-reg-attempt - The maximum number of registration attempts to be made before a server hunt is issued.

max-request-retransmit - The maximum number of attempts to be made when requesting information from the local ENRP server before a server hunt is issued.

5. References

[SCTP] R. R. Stewart, Q. Xie, K. Morneault, C. Sharp, H. J. Schwarzbauer, <u>T. Taylor, I. Rytina, M. Kalla, L. Zhang, and, V. Paxson, "Stream</u> Control Transmission Protocol," <<u>RFC 2960</u>>, October 2000.

[ASAP] Q. Xie, R. R. Stewart "Aggregate Server Access Protocol", <u>draft-stewart-rserpool-asap-00.txt</u>, work in progress.

6. Acknowledgements

The authors wish to thank John Loughney, Lyndon Ong, and Maureen Stillman and many others for their invaluable comments.

7. Authors' Addresses

Randall R. Stewart 24 Burning Bush Trail. Crystal Lake, IL 60012 USA

Phone: +1-815-477-2127 EMail: rrs@cisco.com

Qiaobing Xie Motorola, Inc. 1501 W. Shure Drive, #2309 Arlington Heights, IL 60004 USA

Phone: +1-847-632-3028 EMail: qxie1@email.mot.com