Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: 1 October 2025

T. Dreibholz
SimulaMet
X. Zhou
Hainan University

30 March 2025

Definition of a Delay Measurement Infrastructure and Delay-Sensitive Least-Used Policy for Reliable Server Pooling
draft-dreibholz-rserpool-delay-35

Abstract

This document contains the definition of a delay measurement infrastructure and a delay-sensitive Least-Used policy for Reliable Server Pooling.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on 1 October 2025.

Copyright Notice

Table of Contents

1. Introduction

Reliable Server Pooling defines protocols for providing highly available services. PEs of a pool may be distributed over a large geographical area, in order to provide redundancy in case of localized disasters. But the current pool policies defined in [8] do not incorporate the fact of distances (i.e. delay) between PU and PE. This leads to a low performance for delay-sensitive applications.

1.1. Scope

This draft defines a delay measurement infrastructure for ENRP servers to add delay information into the handlespace. Furthermore, a delay-sensitive Least-Used policy is defined. Performance evaluations can be found in [13].

1.2. Terminology

The terms are commonly identified in related work and can be found in the Aggregate Server Access Protocol and Endpoint Handlespace Redundancy Protocol Common Parameters document [6].

1.3. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [1].

2. Delay-Measurement Infrastructure

This section describes the necessary delay measurement infrastructure for the policy later defined in Section 3. It has to be provided as part of the ENRP servers.

2.1. Quantification of Distance

Measuring delay for SCTP associations is easy: the SCTP protocol [2] already calculates a smoothed round-trip time (RTT) for the primary path. This RTT only has to be queried via the standard SCTP API as defined in [9]. By default, the calculated RTT has a small restriction: a SCTP endpoint waits up to 200ms before acknowledging a packet, in order to piggyback the acknowledgement chunk with payload data. In this case, the RTT would include this latency. By using the option SCTP_DELAYED_SACK (see [9]), the maximum delay before acknowledging a packet can be set to 0ms (i.e. "acknowledge as soon as possible"). After that, the RTT approximately consists of the network latency only. Then, using the RTT, the end-to-end delay between two associated components is approximately 0.5*RTT.

In real networks, there may be negligible delay differences: for example, the delay between a PU and PE #1 is 5ms and the latency between the PU and PE #2 is 6ms. From the service user’s perspective, such minor delay differences may be ignored and are furthermore unavoidable in Internet scenarios. Therefore, the distance parameter between two components A and B is defined as follows:

Distance = DistanceStep * round((0.5*RTT) / DistanceStep)

That is, the distance parameter is defined as the nearest integer multiple of the constant DistanceStep for the measured delay (i.e. 0.5*RTT).

2.2. Distance Measurement Environment

In order to define a distance-aware policy, it is first necessary to define a basic rule: PEs and PUs choose "nearby" ENRP servers. Since the operation scope of RSerPool is restricted to a single organization, this condition can be met easily by appropriately locating ENRP servers.

* A Home ENRP server can measure the delay of the ASAP associations to its PE. As part of its ENRP updates to other ENRP servers, it can report this measured delay together with the PE information.

* A non-Home-ENRP server receiving such an update simply adds the delay of the ENRP association with the Home ENRP server to the PE's reported delay.

Now, each ENRP server can approximate the distance to every PE in the operation scope using the equation in Section 2.1.

Note, that delay changes are propagated to all ENRP servers upon PE re-registrations, i.e. the delay information (and the approximated distance) dynamically adapts to the state of the network.

3. Distance-Sensitive Least-Used Policy

In this section, a distance-sensitive Least Used policy is defined, based on the delay-measurement infrastructure introduced in Section 2.

3.1. Description

The Least Used with Distance Penalty Factor (LU-DPF) policy uses load information provided by the pool elements to select the lowest-loaded pool elements within the pool. If there are multiple elements having lowest load, the nearest PE should be chosen.

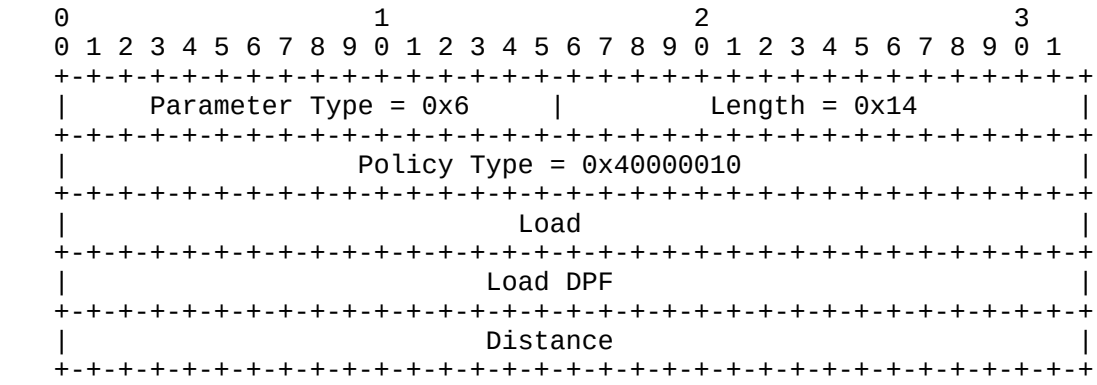
3.2. ENRP Server Considerations

The ENRP server SHOULD select at most the requested number of pool elements. Their load values SHOULD be the lowest possible ones within the pool and their distances also SHOULD be lowest. Each element MUST NOT be reported more than once to the pool user. If there is a choice of equal-loaded and equal-distanced pool elements, round robin selection SHOULD be made among these elements. The returned list of pool elements MUST be sorted by load value in ascending order (1st key) and distance in ascending order (2nd key).

3.3. Pool User Considerations

The pool user should try to use the pool elements returned from the list in the order returned by the ENRP server. A subsequent call for handle resolution may result in the same list. Therefore, it is RECOMMENDED for a pool user to request multiple entries in order to have a sufficient amount of feasible backup entries available.

3.4. Pool Member Selection Policy Parameter



* Load: Current load of the pool element.

* Load DPF: The LoadDPF setting of the PE.

* Distance: The approximated distance in milliseconds.

  - Between PE and Home ENRP server: The distance SHOULD be set to 0.

  - Between Non-Home ENRP server and Home ENRP server: The delay measured on the ASAP association between Home ENRP server and PE.

  - Between ENRP server and PU: The sums of the measured delays on the ASAP association and the ENRP association to the Home ENRP server.

4. Reference Implementation

The RSerPool reference implementation RSPLIB can be found at [15]. It supports the functionalities defined by [3], [4], [5], [6] and [8] as well as the options [10], [11] and of course the option defined by this document. An introduction to this implementation is provided in [12].

5. Testbed Platform

A large-scale and realistic Internet testbed platform with support for the multi-homing feature of the underlying SCTP protocol is NorNet. A description of NorNet is provided in [14], some further information can be found on the project website [16].

6. Security Considerations

Security considerations for RSerPool systems are described by [7].

7. IANA Considerations

This document does not require additional IANA actions beyond those already identified in the ENRP and ASAP protocol specifications.

8. References

8.1. Normative References

[1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[2] Stewart, R., Ed., "Stream Control Transmission Protocol", RFC 4960, DOI 10.17487/RFC4960, September 2007, <https://www.rfc-editor.org/info/rfc4960>.

[3] Lei, P., Ong, L., Tuexen, M., and T. Dreibholz, "An Overview of Reliable Server Pooling Protocols", RFC 5351, DOI 10.17487/RFC5351, September 2008, <https://www.rfc-editor.org/info/rfc5351>.

[4] Stewart, R., Xie, Q., Stillman, M., and M. Tuexen, "Aggregate Server Access Protocol (ASAP)", RFC 5352, DOI 10.17487/RFC5352, September 2008, <https://www.rfc-editor.org/info/rfc5352>.

[5] Xie, Q., Stewart, R., Stillman, M., Tuexen, M., and A. Silverton, "Endpoint Handlespace Redundancy Protocol (ENRP)", RFC 5353, DOI 10.17487/RFC5353, September 2008, <https://www.rfc-editor.org/info/rfc5353>.

[6] Stewart, R., Xie, Q., Stillman, M., and M. Tuexen, "Aggregate Server Access Protocol (ASAP) and Endpoint Handlespace Redundancy Protocol (ENRP) Parameters", RFC 5354, DOI 10.17487/RFC5354, September 2008, <https://www.rfc-editor.org/info/rfc5354>.

[7] Stillman, M., Ed., Gopal, R., Guttman, E., Sengodan, S., and M. Holdrege, "Threats Introduced by Reliable Server Pooling (RSerPool) and Requirements for Security in Response to Threats", RFC 5355, DOI 10.17487/RFC5355, September 2008, <https://www.rfc-editor.org/info/rfc5355>.

[8] Dreibholz, T. and M. Tuexen, "Reliable Server Pooling Policies", RFC 5356, DOI 10.17487/RFC5356, September 2008, <https://www.rfc-editor.org/info/rfc5356>.

[9] Stewart, R., Tuexen, M., Poon, K., Lei, P., and V. Yasevich, "Sockets API Extensions for the Stream Control Transmission Protocol (SCTP)", RFC 6458, DOI 10.17487/RFC6458, December 2011, <https://www.rfc-editor.org/info/rfc6458>.

[10] Dreibholz, T., "Handle Resolution Option for ASAP", Work in Progress, Internet-Draft, draft-dreibholz-rserpool-asap-hropt-29, 6 September 2021, <https://www.ietf.org/archive/id/draft-dreibholz-rserpool-asap-hropt-29.txt>.

[11] Dreibholz, T. and X. Zhou, "Takeover Suggestion Flag for the ENRP Handle Update Message", Work in Progress, Internet-Draft, draft-dreibholz-rserpool-enrp-takeover-26, 6 September 2021, <https://www.ietf.org/archive/id/draft-dreibholz-rserpool-enrp-takeover-26.txt>.

8.2. Informative References

[12] Dreibholz, T., "Reliable Server Pooling Evaluation, Optimization and Extension of a Novel IETF Architecture", 7 March 2007, <https://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-16326/Dre2006_final.pdf>.

[13] Dreibholz, T. and E. P. Rathgeb, "On Improving the Performance of Reliable Server Pooling Systems for Distance-Sensitive Distributed Applications", Proceedings of the 15. ITG/GI Fachtagung Kommunikation in Verteilten Systemen (KiVS) Pages 39-50, ISBN 978-3-540-69962-0, DOI 10.1007/978-3-540-69962-0_4, 28 February 2007, <https://www.wiwi.uni-due.de/fileadmin/fileupload/I-TDR/ReliableServer/Publications/KiVS2007.pdf>.

[14] Dreibholz, T. and E. G. Gran, "Design and Implementation of the NorNet Core Research Testbed for Multi-Homed Systems", Proceedings of the 3nd International Workshop on Protocols and Applications with Multi-Homing Support (PAMS) Pages 1094-1100, ISBN 978-0-7695-4952-1, DOI 10.1109/WAINA.2013.71, 27 March 2013, <https://www.simula.no/file/threfereedinproceedingsreference2012-12-207643198512pdf/download>.

[15] Dreibholz, T., "Thomas Dreibholz's RSerPool Page", 2022, <https://www.nntb.no/~dreibh/rserpool/>.

[16] Dreibholz, T., "NorNet A Real-World, Large-Scale Multi-Homing Testbed", 2022, <https://www.nntb.no/>.

Authors' Addresses

Thomas Dreibholz
Simula Metropolitan Centre for Digital Engineering
Stensberggata 27
0170 Oslo
Norway
Email: dreibh@simula.no
URI: https://www.simula.no/people/dreibh

Xing Zhou
Hainan University, College of Information Science and Technology
Renmin Avenue 58
570228 Haikou
Hainan,
China
Phone: +86-898-66279141
Email: zhouxing@hainanu.edu.cn