Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: April 16, 2016

Q. Sun
China Telecom
M. Boucadair
France Telecom
S. Sivakumar
Cisco Systems
C. Zhou
Huawei Technologies
T. Tsou
Huawei Technologies (USA)
S. Perreault
Jive Communications
October 14, 2015

Port Control Protocol (PCP) Extension for Port Set Allocation
draft-ietf-pcp-port-set-11

Abstract

In some use cases, e.g.,Lightweight 4over6 (lw4o6) [RFC7596], the client may require not just one port, but a port set. This document defines an extension to the Port Control Protocol (PCP) allowing clients to manipulate sets of ports as a whole. This is accomplished by a new MAP option: PORT_SET.

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 http://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 April 16, 2016.

Copyright Notice

Table of Contents

1. Introduction

This document extends PCP [RFC6887] with the ability to retrieve a set of ports using a single request. It does so by defining a new PORT_SET option.

This section describes a few (and non-exhaustive) envisioned use cases. Note that the PCP extension defined in this document is generic and is expected to be applicable to other use cases.

1.1. Applications Using Port Sets

Some applications require not just one port, but a port set. One example is a Session Initiation Protocol (SIP) User Agent Server (UAS) [RFC3261] expecting to handle multiple concurrent calls, including media termination. When it receives a call, it needs to signal media port numbers to its peer. Generating individual PCP MAP requests for each of the media ports during call setup would introduce unwanted latency. Instead, the server can pre-allocate a set of ports such that no PCP exchange is needed during call setup.

1.2. Lightweight 4over6

In the Lightweight 4over6 (lw4o6) [RFC7596] architecture, shared global addresses can be allocated to customers. It allows moving the Network Address Translation (NAT) function, otherwise accomplished by a Carrier-Grade NAT (CGN) [RFC6888], to the Customer-Premises Equipment (CPE). This provides more control over the NAT function to the user, and more scalability to the ISP.

In the lw4o6 architecture, the PCP-controlled device corresponds to the Lightweight AFTR (lwAFTR), and the PCP client corresponds to the Lightweight B4 (lwB4). The PCP client sends a PCP MAP request containing a PORT_SET option to trigger shared address allocation on the Lightweight AFTR (lwAFTR). The PCP response contains the shared address information, including the port set allocated to the Lightweight B4 (lwB4).

1.3. Firewall Control

Port sets are often used in firewall rules. For example, defining a range for RTP [RFC3550] traffic is common practice. The MAP request can already be used for firewall control. The PORT_SET option brings the additional ability to manipulate firewall rules operating on port sets instead of single ports.

1.4. Discovering Stateless Port Set Mappings

A MAP request can be used to retrieve a mapping from a stateless device (i.e., one that does not establish any per-flow state, and simply rewrites the address and/or port in a purely algorithmic fashion, including no rewriting). Similarly, a MAP request with a PORT_SET request can be used to discover a port set mapping from a stateless device. See Section 5.2 for an example.

2. The need for PORT_SET

Multiple MAP requests can be used to manipulate a set of ports, having roughly the same effect as a single use of a MAP request with a PORT_SET option. However, use of the PORT_SET option is more efficient when considering the following aspects:

Network Traffic: A single request uses less network resources than multiple requests.

Latency: Even though MAP requests can be sent in parallel, we can expect the total processing time to be longer for multiple requests than a single one.

Server-side efficiency: Some PCP-controlled devices can allocate port sets in a manner such that data passing through the device is processed much more efficiently than the equivalent using individual port allocations. For example, a CGN having a "bulk" port allocation scheme (see [RFC6888] section 5) often has this property.

Server-side scalability: The number of state table entries in PCP-controlled devices is often a limiting factor. Allocating port sets in a single request can result in a single mapping entry being used, therefore allowing greater scalability.

Therefore, while it is functionally possible to obtain the same results using plain MAP, the extension proposed in this document allows greater efficiency, scalability, and simplicity, while lowering latency and necessary network traffic.

In addition, PORT_SET supports parity preservation. Some protocols (e.g. RTP [RFC3550]) assign meaning to a port number's parity. When mapping sets of ports for the purpose of using such kind of protocol, preserving parity can be necessary.

3. Terminology

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 [RFC2119].

4. The PORT_SET Option

Option Name: PORT_SET

Number: TBD

Purpose: To map sets of ports.

Valid for Opcodes: MAP

Length: 5 bytes

May appear in: Both requests and responses

Maximum occurrences: 1

The PORT_SET Option indicates that the PCP client wishes to reserve a set of ports. The requested number of ports in that set is indicated in the option.

The maximum occurrences of the PORT_SET Option should be limited to 1. The reason is that the suggested external port set depends on the data contained in the MAP Opcode header. Having two PORT_SET options with a single MAP Opcode header would imply having two overlapping suggested external port sets.

Note that the option number is in the "optional to process" range (128-191), meaning that a MAP request with a PORT_SET option will be interpreted by a PCP server that does not support PORT_SET as a single-port MAP request, as if the PORT_SET option was absent.

The PORT_SET Option is formatted as shown in Figure 1.

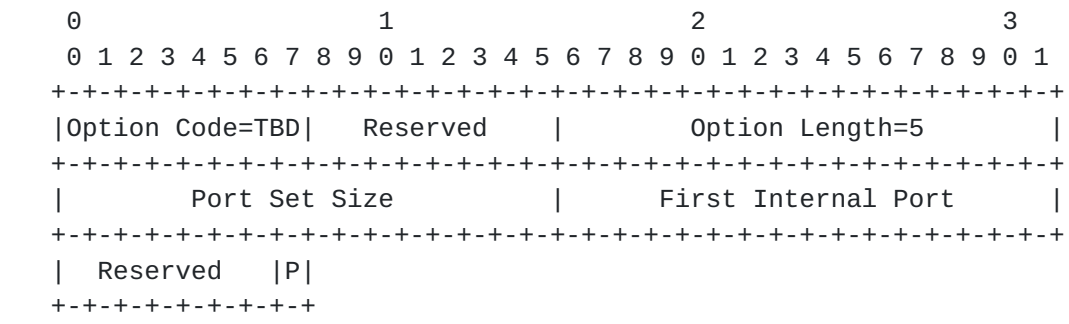

Figure 1: PORT_SET Option

The fields are as follows:

Port Set Size: Number of ports requested. MUST NOT be zero.

First Internal Port: In a request, this field MUST be set equal to the Internal Port field in the MAP opcode by the PCP client. In a response, this field indicates the first internal port of the port set mapped by the PCP server, which may differ from the value sent in the request. That is to be contrasted to the Internal Port field, which by necessity is always identical in matched requests and responses.

Reserved: MUST be set to zero when sending, MUST be ignored when receiving.

P: 1 if parity preservation is requested, 0 otherwise. See [RFC4787], Section 4.2.2.

The Internal Port Set is defined as being the range of Port Set Size ports starting from the First Internal Port. The External Port Set is respectively defined as being the range of Port Set Size ports starting from the Assigned External Port. The two ranges always have the same size (i.e., the Port Set Size returned by the PCP server).

The Suggested External Port corresponds to the first port in the Assigned External Port Set. Its purpose is for clients to be able to regenerate previous mappings after state loss. When such an event happens, clients may attempt to regenerate identical mappings by suggesting the same External Port Set as before the state loss. Note that there is no guarantee that the allocated External Port Set will be the one suggested by the client. In particular, the PREFER_FAILURE option MUST NOT be present in a request that contains a PORT_SET option.

4.1. Client Behavior

To retrieve a set of ports, the PCP client adds a PORT_SET option to its PCP MAP request. If port preservation is required, the PCP Client MUST set the parity bit (to 1) to ask the PCP server to preserve the port parity.

The PCP Client MUST NOT include more than one PORT_SET option in a MAP request. If several port sets are needed, the PCP client MUST issue separate MAP requests, each potentially including a PORT_SET option. These individual MAP requests MUST include distinct Internal Port.

If the PCP Client does not know the exact number of ports it requires, it MAY then set the Port Set Size to 0xffff, indicating that it is willing to accept as many ports as the PCP server can offer.

When the PCP-controlled device supports multiple port-sets delegation for a given PCP client, the PCP client MAY re-initiate a PCP request to get another port set when it has exhausted all the ports within the port-set.

4.2. Server Behavior

In addition to regular MAP request processing, the following checks are made upon receipt of a PORT_SET option with non-zero Requested Lifetime:

- If multiple PORT_SET options are present in a single MAP request, a MALFORMED_OPTION error is returned.

- If the Port Set Size is zero, a MALFORMED_OPTION error is returned.

PREFER_FAILURE MUST NOT appear in a request with PORT_SET option. As a reminder PREFER_FAILURE was specifically designed for the Universal Plug and Play (UPnP) Internet Gateway Device - Port Control Protocol Interworking Function (IGD-PCP IWF) [RFC6970]. The reasons for not recommending the use of PREFER_FAILURE are discussed in Section 13.2 of [RFC6887]. The PCP server MAY map fewer ports than the value of Port Set Size from the request. It MUST NOT map more ports than the PCP client asked for. Internal ports outside the range of Port Set Size ports starting from the Internal Port MUST NOT be mapped by the PCP server.

If the requested port set cannot be fully satisfied, the PCP server SHOULD map as many ports as possible, and SHOULD map at least one port (which is the same behavior as if Port Set Size is set to 1).

If the PCP server ends up mapping only a single port, for any reason, the PORT_SET option MUST NOT be present in the response.

If the port parity preservation is requested (P = 1), the PCP server MAY preserve port parity. In that case, the External Port is set to a value having the same parity as the First Internal Port.

If the mapping is successful, the MAP response's Assigned External Port is set to the first port in the External Port Set, and the PORT_SET option's Port Set Size is set to number of ports in the mapped port set. The First Internal Port field is set to the first port in the Internal Port Set.

4.3. Absence of Capability Discovery

A PCP client that wishes to make use of a port set unconditionally includes the PORT_SET option. If no PORT_SET option is present in the response, the PCP client cannot conclude that the PCP server does not support the PORT_SET option. It may just be that the PCP server does support PORT_SET but decided to allocate only a single port, for reasons that are its own. If the client wishes to obtain more ports, it MAY send additional MAP requests (see Section 6.4), which the PCP server may or may not grant according to local policy.

If port set capability is added to or removed from a running PCP server, the server MAY reset its Epoch time and send an ANNOUNCE message as described in the PCP specification ([RFC6887], Section 14.1). This causes PCP clients to re-try, and those using PORT_SET will now receive a different response.

4.4. Port Set Renewal and Deletion

Port set mappings are renewed and deleted as a single entity. That is, the lifetime of all port mappings in the set is set to the Assigned Lifetime at once.

A PCP client attempting to refresh or delete a port set mapping MUST include the PORT_SET option in its request. A PCP client MUST NOT send a PORT_SET option for single-port refreshes.

4.4.1. Overlap Conditions

Port set map requests can overlap with existing single port or port set mappings. This can happen either by mistake or after a PCP client becomes out of sync with server state.

If a PCP server receives a MAP request, with or without a PORT_SET option, that tries to map one or more internal ports or port sets belonging to already existing mappings, then the request is considered to be a refresh request applying those mappings. Each of the matching port or port set mappings is processed independently, as if a separate refresh request had been received. The processing is as described in Section 15 of [RFC6887]. The PCP server sends a Mapping Update message for each of the mappings.

5. Examples

5.1. Simple Request on NAT44

An application requires a range of 100 IPv4 UDP ports to be mapped to itself. The application running on the host has created sockets bound to IPv4 UDP ports 50,000 to 50,099 for this purpose. It does not care about which external port numbers are allocated. The PCP client sends a PCP request with the following parameters over IPv4:

- MAP opcode

  Mapping Nonce: <a random nonce>

  Protocol: 17

  Internal Port: 50,000

  Suggested External Port: 0

  Suggested External IP Address: ::ffff:0.0.0.0

- PORT_SET Option

  Port Set Size: 100

  First Internal Port: 50,000

  P: 0

The PCP server is unable to fulfill the request fully: it is configured by local policy to only allocate 32 ports per user. Since the PREFER_FAILURE option is absent from the request, it decides to

map UDP ports 37,056 to 37,087 on external address 192.0.2.3 to internal ports 50,000 to 50,031. After setting up the mapping in the NAT44 device it controls, it replies with the following PCP response:

- MAP opcode

  Mapping Nonce: <copied from the request>

  Protocol: 17

  Internal Port: 50,000

  Assigned External Port: 37,056

  Assigned External IP Address: ::ffff:192.0.2.3

- PORT_SET Option

  Port Set Size: 32

  First Internal Port: 50,000

  P: 0

Upon receiving this response, the host decides that 32 ports is good enough for its purposes. It closes sockets bound to ports 50,032 to 50,099, sets up a refresh timer, and starts using the port range it has just been assigned.

5.2. Stateless Mapping Discovery

A host wants to discover a stateless NAT44 mapping pointing to it. To do so, it sends the following request over IPv4:

- MAP opcode

  Mapping Nonce: <a random nonce>

  Protocol: 0

  Internal Port: 1

  Suggested External Port: 0

  Suggested External IP Address: ::ffff:0.0.0.0

- PORT_SET Option

Port Set Size: 65,535

First Internal Port: 1

P: 0

The PCP server sends the following response:

- MAP opcode

  Mapping Nonce: <copied from the request>

  Protocol: 0

  Internal Port: 1

  Assigned External Port: 26,624

  Assigned External IP Address: ::ffff:192.0.2.5

- PORT_SET Option

  Port Set Size: 2048

  First Internal Port: 26,624

  P: 0

From this response, the host understands that a 2048-port stateless mapping is pointing to itself, starting from port 26,624 on external IP address 192.0.2.5.

5.3. Resolving Overlap

This example relates to Section 4.4.1.

Suppose internal port 100 is mapped to external port 100 and port set 101-199 is mapped to external port set 201-299. The PCP server receives a MAP request with Internal Port = 100, External Port = 0, and a PORT_SET option with Port Set Size = 100. The request's Mapping Nonce is equal to those of the existing single port and port set mappings. This request is therefore treated as two refresh requests, the first one applying to the single port mapping and the second one applying to the port set mapping. The PCP server updates both mapping's lifetimes as usual then sends two responses: the first one contains Internal Port = 100, External Port = 100, and no PORT_SET option, while the second one contains Internal Port = 101, External Port = 201, and a PORT_SET option with Port Set Size = 99.

6. Operational Considerations

6.1. Limits and Quotas

It is up to the PCP server to determine the port-set quota, if any, for each PCP client.

If the PCP server is configured to allocate multiple port-set allocations for one subscriber, the same Assigned External IP Address SHOULD be assigned to the subscriber in multiple port-set responses.

To optimize the number of mapping entries maintained by the PCP server, it is RECOMMENDED to configure the PCP server to assign the maximum allowed port set size in a single response. This policy SHOULD be configurable.

6.2. High Availability

The failover mechanism in MAP [section 14 in [RFC6887]] can also be applied to port sets.

6.3. Idempotence

A core, desirable property of the PCP protocol is idempotence. In a nutshell, requests produce the same results whether they are executed once or multiple times. This property is preserved with the PORT_SET attribute, with the following caveat: the order in which the PCP server receives requests with overlapping Internal Port Sets will affect the mappings being created and the responses received.

For example suppose these two requests are sent by a PCP client:

Request A: Internal Port Set 1-10

Request B: Internal Port Set 5-14

The PCP server's actions will depend on which request is received first. Suppose that A is received before B:

Upon reception of A: Internal ports 1-10 are mapped. A success response containing the following fields is sent:

Internal Port: 1

First Internal Port: 1

Port Set Size: 10

Upon reception of B: The request matches mapping A. The request is interpreted as a refresh request for mapping A, and a response containing the following fields is sent:

Internal Port: 5

First Internal Port: 1

Port Set Size: 10

If the order of reception is reversed (B before A), the created mapping will be different, and the First Internal Port in both responses would then be 5.

To avoid surprises, PCP clients MUST ensure that port set mapping requests do not inadvertently overlap. For example, a host's operating system could include a central PCP client process through which port set mapping requests would be arbitrated. Alternatively, individual PCP clients running on the same host would be required to acquire the internal ports from the operating system (e.g., a call to the bind() function from the BSD API) before trying to map them with PCP.

6.4. What should a PCP client do when it receives fewer ports than requested?

Suppose a PCP client asks for 16 ports and receives 8. What should it do? Should it consider this a final answer? Should it try a second request, asking for 8 more ports? Should it fall back to 8 individual MAP requests? This document leaves the answers to be implementation-specific, but describes issues to be considered when answering them.

First, the PCP server has decided to allocate 8 ports for some reason. It may be that allocation sizes have been limited by the PCP server's administrator. It may be that the PCP client has reached a quota. It may be that these 8 ports were the last contiguous ones available. Depending on the reason, asking for more ports may or may not be likely to actually yield more ports. However, the PCP client has no way of knowing.

Second, not all PCP clients asking for N ports actually need all N ports to function correctly. For example, a DNS resolver could ask for N ports to be used for source port randomization. If fewer than N ports are received, the DNS resolver will still work correctly, but source port randomization will be slightly less efficient, having fewer bits to play with. In that case, it would not make much sense to ask for more ports.

Finally, asking for more ports could be considered abuse. External ports are a resource that is to be shared among multiple PCP clients. A PCP client trying to obtain more than its fair share could trigger countermeasures according to local policy.

In conclusion, it is expected that for most applications, asking for more ports would not yield benefits justifying the additional costs.

7. Security Considerations

The security considerations discussed in [RFC6887] apply to this extension.

As described in Section 4.4.1, a single PCP request using the PORT_SET option may result in multiple responses. For this to happen it is necessary that the request contain the nonce associated to multiple mappings on the server. Therefore, an on-path attacker could use an eavesdropped nonce to mount an amplification attack. Use of PCP authentication ([RFC6887], Section 18) eliminates this attack vector.

8. IANA Considerations

IANA has allocated value TBD (note to IANA: to be allocated from the range 128-191) in the "PCP Options" registry at http://www.iana.org/assignments/pcp-parameters for the new PCP option defined in Section 4.

9. Contributors

The following are extended authors who contributed to the effort:

Yunqing Chen

China Telecom

Room 502, No.118, Xizhimennei Street

Beijing 100035

P.R.China

Chongfeng Xie

China Telecom

Room 502, No.118, Xizhimennei Street

Beijing 100035

P.R.China

Yong Cui

Tsinghua University

Beijing 100084

P.R.China

Phone: +86-10-62603059

Email: yong@csnet1.cs.tsinghua.edu.cn

Qi Sun

Tsinghua University

Beijing 100084

P.R.China

Phone: +86-10-62785822

Email: sunqibupt@gmail.com

Gabor Bajko

Nokia

Email: gabor.bajko@nokia.com

Xiaohong Deng

France Telecom

Email: xiaohong.deng@orange-ftgroup.com

10. Acknowledgements

The authors would like to show sincere appreciation to Alain Durand, Cong Liu, Dan Wing, Dave Thaler, Peter Koch, Reinaldo Penno, Sam Hartman, Stuart Cheshire, Ted Lemon, and Yoshihiro Ohba, for their useful comments and suggestions.

11. References

11.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC6887] Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P. Selkirk, "Port Control Protocol (PCP)", RFC 6887, April 2013.

11.2. Informative References

[RFC3261] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., and E. Schooler, "SIP: Session Initiation Protocol", RFC 3261, June 2002.

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.

[RFC4787] Audet, F. and C. Jennings, "Network Address Translation (NAT) Behavioral Requirements for Unicast UDP", BCP 127, RFC 4787, January 2007.

[RFC6888] Perreault, S., Yamagata, I., Miyakawa, S., Nakagawa, A., and H. Ashida, "Common Requirements for Carrier-Grade NATs (CGNs)", BCP 127, RFC 6888, April 2013.

[RFC7596] Cui, Y., Qiong, Q., Boucadair, M., Tsou, T., Lee, Y., and I. Farrer, "Lightweight 4over6: An Extension to the DS-Lite Architecture", RFC 7596, July 2015.

Authors' Addresses

Qiong Sun
China Telecom
P.R.China

Phone: 86 10 58552936
Email: sunqiong@ctbri.com.cn

Mohamed Boucadair
France Telecom
Rennes 35000
France

Email: mohamed.boucadair@orange.com

Senthil Sivakumar
Cisco Systems
7100-8 Kit Creek Road
Research Triangle Park, North Carolina 27709
USA

Phone: +1 919 392 5158
Email: ssenthil@cisco.com

Cathy Zhou
Huawei Technologies
Bantian, Longgang District
Shenzhen 518129
P.R. China

Email: cathy.zhou@huawei.com

Tina Tsou
Huawei Technologies (USA)
2330 Central Expressway
Santa Clara, CA 95050
USA

Phone: +1 408 330 4424
Email: Tina.Tsou.Zouting@huawei.com

Simon Perreault
Jive Communications
Quebec, QC
Canada

Email: sperreault@jive.com