Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: April 24, 2014

Q. Sun
China Telecom
M. Boucadair
France Telecom
S. Sivakumar
Cisco Systems
C. Zhou
Huawei Technologies
T. Tsou
Huawei Technologies (USA)
S. Perreault
Viagenie
October 21, 2013

# Port Control Protocol (PCP) Extension for Port Set Allocation
# draft-ietf-pcp-port-set-03

Abstract

This document defines an extension to 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 24, 2014.

Table of Contents



## 1. Introduction

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. Lightweight 4over6

In the Lightweight 4over6 [I-D.ietf-softwire-lw4over6] 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 lwAFTR, and the PCP client corresponds to the lwB4. The client sends a PCP MAP request containing a PORT_SET option to trigger shared address allocation on the lwAFTR. The PCP response contains the shared address information, including the port set allocated to the lwB4.

## 1.2. 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.

Using PORT_SET, an application can manipulate port sets much more efficiently than with individual MAP requests.

Another example of an application using port sets is that of a busy back-to-back PCP server/client [I-D.cheshire-recursive-pcp], handling many requests per second. It could benefit from PORT_SET by obtaining ports from upstream in big chunks. Then it would manage those chunks like port pools from which it would allocate to downstream clients. That could be more efficient than obtaining ports from upstream with individual MAP requests.

## 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 discover a stateless mapping. Similarly, a MAP request with a PORT_SET request can be used to discover a stateless port set mapping. Hence, PORT_SET is applicable for port set mapping discovery in Stateless NAT44 [I-D.tsou-stateless-nat44].

## 2. 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].

## 3. 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.

Client-side simplicity: The logic that is necessary for maintaining a set of ports using a single port set entity is much simpler than that required for maintaining individual ports, especially when considering failures, retransmissions, lifetime expiration, and re-allocations.

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 mapping 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 a nutshell, PORT_SET is a necessary optimization.

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.

## 4. The PORT_SET Option

Option Name: PORT_SET

Number: TBD

Purpose: To map sets of ports.

Valid for Opcodes: MAP

Length: 3 bytes

May appear in: Both requests and responses

Maximum occurrences: 1

> NOTE TO IANA (to be removed prior to publication as an RFC): The number is to be assigned by IANA in the range 128-191 (i.e., optional to process and created via Standards Action).

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

The PORT_SET Option is formatted as shown in Figure 1.

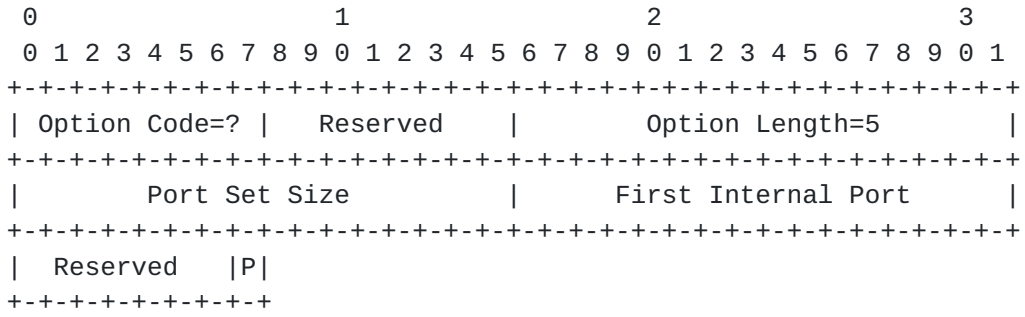


Figure 1: PORT_SET Option

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 client. In a response, this field indicates the first internal port of the port set mapped by the 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.

P: 1 if parity preservation is requested, 0 otherwise.

NOTE: In its current form, PORT_SET does not support allocating discontinuous port sets. That feature could be added in the future depending on input from the working group.

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 server).

### 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 server to preserve the port parity (i.e., the Assigned External Port and First Internal Port have the same parity). The PCP client MUST indicate a suggested Port Set Size. A non-null value MUST be used. The client MUST set the First Internal Port field equal to the Internal Port field in the MAP opcode.

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 as many MAP requests each of them include 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 server can offer.

If the PORT_SET option is not supported by the server, the PCP client will receive a response with no PORT_SET option, which is the operation of a PCP server that does not support PORT_SET. That response will map one port. To map its other desired ports, the PCP client will then have to issue individual MAP requests with no PORT_SET option to achieve similar functionality.

### 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.

If the PREFER_FAILURE option is present and the server is unable to map all ports in the requested External Port Set or is unable to preserve parity (P = 1), the CANNOT_PROVIDE_EXTERNAL error is returned.

If the PREFER_FAILURE option is absent, the server MAY map fewer ports than the value of Port Set Size from the request. It MUST NOT map more ports than the client asked for. Internal ports outside the range of Port Set Size ports starting from the Internal Port MUST NOT be mapped by the 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 same behavior as if PORT_SET is set to 1).

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

If the PREFER_FAILURE option is absent and port parity preservation is requested (P = 1), the 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. 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 client attempting to refresh or delete a port set mapping MUST include the PORT_SET option in its request.

### 4.3.1. 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 client becomes out of sync with server state.

If a 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 a if a separate refresh request had been received. The processing is as described in Section 15 of [RFC6887], with the updated nonce check behavior described in Section 3 of [I-D.cheshire-pcp-unsupp-family]. The server sends a Mapping Update message for each of the mappings.

## 5. Examples

### 5.1. Simple Request on NAT44

A host requires a range of 100 IPv4 UDP ports to be mapped it 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 what external port numbers are allocated. The host 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 4.3.1.

Suppose internal port 100 is mapped to external port 100 and port set 101-199 is mapped to external port set 201-299. The 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 a two refresh requests, the first one applying to the single port mapping and the second one applying to the port set mapping. The server updates both mapping's lifetimes as usual then sends two Mapping Update messages: 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

It is totally up to the PCP server to determine the port-set quota for each PCP client. In addition, 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.

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

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

The failover mechanism in MAP [section 14 in [RFC6887]] and [I-D.boucadair-pcp-failure] can also be applied to port sets.

## 7. Security Considerations

It is believed that no additional security considerations beyond those discussed in [RFC6887] apply to this extension.

## 8. IANA Considerations

IANA shall allocate a code in the range 1-63 for the new PCP option defined in Section 4.

## 9. Authors List

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, Dan Wing, Dave Thaler, Reinaldo Penno, Sam Hartman, Stuart Cheshire, and Yoshihiro Ohba, for their useful comments and suggestions.

## 11. References

### 11.1. Normative References

[I-D.cheshire-pcp-unsupp-family]
Cheshire, S. and S. Perreault, "Update to the PCP specification", draft-cheshire-pcp-unsupp-family-04 (work in progress), June 2013.

[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

[I-D.boucadair-pcp-failure]
Boucadair, M. and R. Penno, "Analysis of Port Control Protocol (PCP) Failure Scenarios", draft-boucadair-pcp-failure-06 (work in progress), May 2013.

[I-D.cheshire-recursive-pcp]
Cheshire, S., "Recursive PCP", draft-cheshire-recursive-pcp-02 (work in progress), March 2013.

[I-D.ietf-softwire-lw4over6]
Cui, Y., Sun, Q., Boucadair, M., Tsou, T., Lee, Y., and I. Farrer, "Lightweight 4over6: An Extension to the DS-Lite Architecture", draft-ietf-softwire-lw4over6-00 (work in progress), April 2013.

[I-D.tsou-stateless-nat44]
Tsou, T., Liu, W., Perreault, S., Penno, R., and M. Chen, "Stateless IPv4 Network Address Translation", draft-tsou-stateless-nat44-02 (work in progress), October 2012.

[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.

[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.

## 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
Viagenie
246 Aberdeen
Quebec, QC G1R 2E1
Canada

Phone: +1 418 656 9254
Email: simon.perreault@viagenie.ca