PCP working group
Internet-Draft
Intended status: Standards Track
Expires: September 11, 2013

S. Cheshire
Apple

Mar 10, 2013

Recursive PCP
draft-cheshire-recursive-pcp-02

Abstract

The Port Control Protocol (PCP) allows clients to request explicit dynamic inbound and outbound port mappings in their closest on-path NAT, firewall, or other middlebox. However, in today's world, there may be more than one NAT on the path between a client and the public Internet. This document describes how the closest on-path middlebox generates a corresponding upstream PCP request to the next closest on-path middlebox, to request an appropriate explicit dynamic port mapping in that middlebox too. Applied recursively, this generates the necessary chain of port mappings in any number of middleboxes on the path between the client and the public Internet.

1. Introduction

When NAT Port Mapping Protocol [NAT-PMP] was first created in 2004, a common network configuration was that a residential customer received a single public routable IPv4 address from their ISP, and had a single NAT gateway serving multiple computers in their home. Consequently, creating appropriate mappings in that single NAT gateway was sufficient to provide full Internet connectivity.

In today's world, with public routable IPv4 addresses becoming less readily available, it is increasingly common for customers to receive a private address from their ISP, and the ISP uses a NAT gateway of its own to translate those packets before sending them out onto the public Internet. This means that there is likely to be more than on NAT on the path between client machines and the public Internet:

- If a residential customer receives a translated address from their ISP, and then installs their own residential NAT gateway to share that address between multiple client devices in their home, then there are at least two NAT gateways on the path between client devices and the public Internet.

- If a mobile phone customer receives a translated address from their mobile phone carrier, and uses "Personal Hotspot" or "Internet Sharing" software on their mobile phone to make Wi-Fi Internet access available to other client devices, then there are at least two NAT gateways on the path between those client devices and the public Internet.

- If a hotel guest connects a portable Wi-Fi gateway, such as an Apple AirPort Express, to their hotel room Ethernet port to share their room's Internet connection between their phone, their iPad, and their laptop computer, then packets from the client devices may traverse the hotel guest's portable NAT, the hotel network's NAT, and the ISP's NAT before reaching the public Internet.

While it is possible, in theory, that client devices could somehow discover all the NATs on the path, and communicate with each one separately using Port Control Protocol [PCP] (NAT-PMP's IETF Standards Track successor), in practice it's not clear how client devices would reliably learn this information. Since the NAT

gateways are installed and operated by different individuals and organizations, no single entity has knowledge of all the NATs on the path. Also, even if a client device could somehow know all the NATs on the path, requiring a client device to communicate separately with all of them imposes unreasonable complexity on PCP clients, many of which are expected to be simple low-cost devices.

In addition, this goes against the spirit of NAT gateways. The main purpose of a NAT gateway is to make multiple downstream client devices making outgoing TCP connections to appear, from the point of view of everything upstream of the NAT gateway, to be a single client device making outgoing TCP connections. In the same spirit, it makes sense for a PCP-capable NAT gateway to make multiple downstream client devices requesting port mappings to appear, from the point of view of everything upstream of the NAT gateway, to be a single client device requesting port mappings.

This document specifies how a PCP-capable NAT gateway uses Recursive PCP to create the appearance of being a single device, from the point of view of the upstream network.

1.1. Conventions and Terminology Used in this Document

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 "Key words for use in RFCs to Indicate Requirement Levels" [RFC2119].

Where this document uses the terms "upstream" and "downstream", the term "upstream" refers to the direction outbound packets travel towards the public Internet, and the term "downstream" refers to the direction inbound packets travel from the public Internet towards client systems. Typically when a home user views a web site, their computer sends an outbound TCP SYN packet upstream towards the public Internet, and an inbound downstream TCP SYN ACK reply comes back from the public Internet.

1.2. Recursive Application

The protocol specified is described as "recursive" because of the following properties:

- When the text refers to the upstream PCP server as if it were the final outermost NAT gateway, in fact that upstream PCP server could itself be another Recursive PCP server making requests to its own upstream PCP server, and relaying back the corresponding replies. That distinction is invisible to the PCP client making the request.

- When the text refers to an incoming PCP request being received from a downstream PCP client, that downstream PCP client could itself be a Recursive PCP server relaying a request on behalf of one of its own downstream PCP clients (which could itself be another Recursive PCP server, and so on). The fact that the Recursive PCP server receiving the request does not need to be aware of this or take any special action, is an important simplifying property of the protocol. The purpose of a NAT gateway is to make many downstream client devices appear to be a single client device, and the purpose of a Recursive PCP server is to make many downstream client devices making PCP requests appear to be a single client device making PCP requests.

This recursive operation is an important simplifying property of the design.

When a PCP client talks to a PCP server, that PCP server behaves *exactly* as if it were the one and only NAT gateway on the path to the public Internet. If the PCP server is not in fact the final outermost NAT gateway, it is the PCP server's responsibility to hide that fact. The client should never have to be aware of the difference between talking to a single NAT gateway, and talking to a NAT gateway which is itself behind one or more other NAT gateways. This simplifying property applies both when the PCP client is a simple end-host client, and when the PCP client is itself the client face of a Recursive PCP server.

Similarly, when a PCP server receives a request from a PCP client, that PCP client behaves exactly as if it were a simple end-host PCP client requesting mappings for itself. If the client is not in fact a simple end-host PCP client, it is the PCP client's responsibility to hide that fact. The server should never have to be aware of the difference between talking to an end-host PCP client, and talking to the client face of a Recursive PCP server that is requesting mappings on behalf of its own downstream clients. If the PCP client is a firewall device, and it chooses to use the PCP THIRD_PARTY Option to make mappings on behalf of its downstream clients, then it should still behave like any other PCP client using the THIRD_PARTY Option.

2. Operation of Recursive PCP

Upon receipt of a PCP mapping-creation request from a downstream PCP client, a Recursive PCP server first examines its local mapping table to see if it already has a valid active mapping matching the Internal Address and Internal Port (and in the case of PEER requests, remote peer) given in the request.

If the Recursive PCP server does not already have a valid active mapping for this mapping-creation request, then it allocates an available port on its external interface. We assume for the sake of this description that the address of its external interface is itself a private address, subject to translation by an upstream NAT. The Recursive PCP server then constructs an appropriate corresponding PCP request of its own (described below), and sends it to its upstream NAT, and the newly-created local mapping is considered temporary until a confirming reply is received from the upstream PCP server.

If the Recursive PCP server does already have a valid active mapping for this mapping-creation request, and the lifetime remaining on the local mapping is at least 3/4 of the lifetime requested by the PCP client, then the Recursive PCP server SHOULD send an immediate reply giving the outermost External Address and Port (previously learned using Recursive PCP, as described below), and the actual lifetime remaining for this mapping. If the lifetime remaining on the local mapping is less than 3/4 of the lifetime requested by the PCP client, then the Recursive PCP server MUST generate an upstream request as described below.

For mapping-deletion requests (Lifetime = 0), the local mapping, if any, is deleted, and then (regardless of whether a local mapping existed) a corresponding upstream request is generated.

How the Recursive PCP server knows the destination IP address for its upstream PCP request is outside the scope of this document, but this may be achieved in a zero-configuration manner using PCP Anycast [Anycast]. In the upstream PCP request:

- The PCP Client's IP Address and Internal Port are the Recursive PCP server's own external address and port just allocated for this mapping.

- The Suggested External Address and Port in the upstream PCP request SHOULD be copied from the original PCP request.

- The Requested Lifetime is as requested by the client if it falls within the acceptable range for this PCP server; otherwise it SHOULD be capped to appropriate minimum and maximum values configured for this PCP server.

- The Mapping Nonce is copied from the original PCP request.

- For PEER requests, the Remote Peer IP Address and Port are copied from the original PCP request.

- Any options in the original PCP request are handled or rejected locally. No options are blindly copied from the original PCP request to the upstream PCP request. Options in the original PCP request pertain to the transaction between the client and its Recursive PCP server. In the new upstream PCP request PCP options may also be used if necessary to create the desired mapping, but they are best thought of as new options pertaining to the transaction between the Recursive PCP server and its upstream PCP server, rather than as pre-existing options that were "copied" from the original PCP request (even if, in some cases, the content of those new options may be similar or identical to the options in the original PCP request).

Upon receipt of a PCP reply giving the outermost (i.e. publicly routable) External Address, Port and Lifetime, the Recursive PCP server records this information in its own mapping table and relays the information to the requesting downstream PCP client in a PCP reply. The Recursive PCP server therefore records, among other things, the following information in its mapping table:

- Client's Internal Address and Port.

- External Address and Port allocated by this Recursive PCP server.

- Outermost External Address and Port allocated by the upstream PCP server.

- Mapping lifetime (also dictated by the upstream PCP server).

- Mapping nonce.

In the downstream PCP reply:

- The Lifetime is as granted by the upstream PCP server, or less, if the granted lifetime exceeds the maximum lifetime this PCP server is configured to grant. If the downstream Lifetime is more than the Lifetime granted by the upstream PCP server (which is NOT RECOMMENDED) then this Recursive PCP server MUST take responsibility for renewing the upstream mapping itself.

- The Epoch Time is *this* Recursive PCP server's Epoch Time, not the Epoch Time of the upstream PCP server. Each PCP server has its own independent Epoch Time. However, if the Epoch Time received from the upstream PCP server indicates a loss of state in that PCP server, the Recursive PCP server can either recreate the lost mappings itself, or it can reset its own Epoch Time to cause its downstream clients to perform such state repairs themselves. A Recursive PCP server MUST NOT simply copy the upstream PCP

server's Epoch Time into its downstream PCP replies, since if it suffers its own state loss it needs the ability to communicate that state loss to clients. Thus each PCP server has its own independent Epoch Time. However, as a convenience, a downstream Recursive PCP server may simply choose to reset its own Epoch Time whenever it detects that its upstream PCP server has lost state. Thus, in this case, the Recursive PCP server's Epoch Time always resets whenever its upstream PCP server loses state; it may also reset at other times too.

- The Mapping Nonce is copied from the reply received from the upstream PCP server.

- The Assigned External Port and Assigned External IP Address are copied from the reply received from the upstream PCP server. (I.e. they are the outermost External IP Address and Port, not the locally-assigned external address and port.)

- For PEER requests, the Remote Peer IP Address and Port are copied from the reply received from the upstream PCP server.

- Any options in the reply received from the upstream PCP server are handled locally as appropriate to the options in question. No options are blindly copied from the upstream PCP reply to the downstream PCP reply. If the original PCP request contained options which necessitate a corresponding option in the reply, then appropriate reply options should be generated and inserted into the downstream PCP reply by the Recursive PCP server. These downstream reply options are best thought of as data pertaining to the transaction between the Recursive PCP server and its downstream client, rather than as pre-existing options that were "copied" from the upstream PCP reply into the downstream PCP reply (even if, in some cases, the content of those new options in the downstream PCP reply may be similar or identical to the options received in the reply from the upstream PCP server).

2.1. Optimized Hairpin Routing

A Recursive PCP server SHOULD implement Optimized Hairpin Routing. What this means is the following:

- If a Recursive PCP server observes an outgoing packet arriving on its internal interface that is addressed to an External Address and Port appearing in the NAT gateway's own mapping table, then the NAT gateway SHOULD (after creating a new outbound mapping if one does not already exist) rewrite the packet appropriately and deliver it to the internal client currently allocated that External Address and Port.

- If a Recursive PCP server observes an outgoing packet arriving on its internal interface which is addressed to an Outermost External Address and Port appearing in the NAT gateway's own mapping table, then the NAT gateway SHOULD do likewise: create a new outbound mapping if one does not already exist, and then rewrite the packet appropriately and deliver it to the internal client currently allocated that Outermost External Address and Port. This is not necessary for successful communication, but for efficiency. Without this Optimized Hairpin Routing, the packet will be delivered all the way to the outermost NAT gateway, which will then perform standard hairpin translation and send it back. Using knowledge of the Outermost External Address and Port, this rewriting can be anticipated and performed locally, which will typically offer higher throughput and lower latency than sending it all the way to the outermost NAT gateway and back.

2.2. Termination of Recursion

Any recursive algorithm needs a mechanism to terminate the recursion at the appropriate point. This termination of recursion can be achieved in a variety of ways:

- An ISP's NAT gateway could be configured to know that it is the outermost NAT gateway, and consequently does not need to relay PCP requests upstream. In fact, it may be the case that many large-scale NATs of the kind used by ISPs may simply not implement Recursive PCP, thereby naturally terminating the recursion at that point.

- A NAT gateway could determine automatically that if its external address is not one of the known private addresses [RFC1918][RFC6598] then its external address is a public routable IP address, and consequently it does not need to relay PCP requests upstream.

- A NAT gateway could attempt sending PCP requests upstream, and upon failing to receive any positive reply (e.g. receiving ICMP host unreachable, ICMP port unreachable, or a timeout) conclude that it does not need to relay PCP requests upstream.

2.3. Recursive PCP with Firewalls

When a Recursive PCP server is a NAT gateway, it sends out upstream PCP requests using its own external IP address. When a Recursive PCP server is a firewall, it still needs to install upstream mappings on behalf of its downstream clients. It should do this either by using the downstream client's IP address as the source IP address in its upstream PCP request, or by using the PCP THIRD_PARTY Option in its

upstream PCP request.

3. IANA Considerations

No IANA actions are required by this document.

4. Security Considerations

No new security concerns are raised by use of Recursive PCP. Since the purpose of a NAT gateway is to enable multiple client devices to appear as a single client device to the upstream network, a NAT gateway implementing Recursive PCP maintains this property, appearing to the upstream network to be a single client device using PCP to request port mappings for itself. Whether those port mappings are for multiple processes running on multiple CPUs connected via an internal bus in a single computer, or multiple processes running on multiple CPUs connected via an IP network, is transparent to the external network.

5. References

5.1. Normative References

[PCP] Wing, D., Cheshire, S., Boucadair, M., Penno, R., and P. Selkirk, "Port Control Protocol (PCP)", draft-ietf-pcp-base-29 (work in progress), November 2012.

[RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.

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

[RFC6598] Weil, J., Kuarsingh, V., Donley, C., Liljenstolpe, C., and M. Azinger, "IANA-Reserved IPv4 Prefix for Shared Address Space", BCP 153, RFC 6598, April 2012.

5.2. Informative References

[Anycast] Cheshire, S., "PCP Anycast Address", draft-cheshire-pcp-anycast-00 (work in progress), February 2013.

[NAT-PMP] Cheshire, S., "NAT Port Mapping Protocol (NAT-PMP)",

draft-cheshire-nat-pmp-07 (work in progress),
January 2013.

Author's Address

Stuart Cheshire
Apple Inc.
1 Infinite Loop
Cupertino, California 95014
USA

Phone: +1 408 974 3207
Email: cheshire@apple.com