Network Working Group   A. Kirkham
Internet-Draft   Palo Alto Networks
Obsoletes: None (if approved)   April 10, 2012
Intended status: Informational
Expires: October 12, 2012

Issues with Private IP Addressing in the Internet
draft-ietf-grow-private-ip-sp-cores-01

Abstract

The purpose of this document is to provide a discussion of the potential problems of using private, RFC1918, or non-globally-routable addressing within the core of an SP network. The discussion focuses on link addresses and to a small extent loopback addresses. While many of the issues are well recognised within the ISP community, there appears to be no document that collectively describes the issues.

Legal

Status of this Memo

Table of Contents

1. Introduction

In the mid to late 90's, some Internet Service Providers (ISPs) adopted the practice of utilising private (or non-globally unique) IP (i.e. RFC1918) addresses for the infrastructure links and in some cases the loopback interfaces within their networks. The reasons for this approach centered on conservation of address space (i.e. scarcity of public IPv4 address space), and security of the core network (also known as core hiding).

However, a number of technical and operational issues occurred as a result of using private (or non-globally unique) IP addresses, and virtually all these ISPs moved away from the practice. Tier 1 ISPs are considered the benchmark of the industry and as of the time of writing, there is no known tier 1 ISP that utilises the practice of private addressing within their core network.

The following sections will discuss the various issues associated with deploying private IP (i.e. RFC1918) addresses within ISP core networks.

The intent of this document is not to suggest that private IP can not be used with the core of an SP network as some providers use this practice and operate successfully. The intent is to outline the potential issues or effects of such a practice.

Note: The practice of ISPs using 'stolen' address space (also known as 'squat' space) has many of the same issues (or effects) as that of using private IP address space within core networks. The term "stolen IP address space" refers to the practice of an ISP using address space for its own infrastructure/core network addressing that has been officially allocated by an RIR to another provider, but that provider is not currently using or advertising within the Internet. Stolen addressing is not discussed further in this document. It is simply noted as an associated issue.

2. Conservation of Address Space

One of the original intents for the use of private IP addressing within an ISP core was the conservation of IP address space. When an ISP is allocated a block of public IP addresses (from a RIR), this address block was traditionally split in order to dedicate some portion for infrastructure use (i.e. for the core network), and the other portion for customer (subscriber) or other address pool use. Typically, the number of infrastructure addresses needed is relatively small in comparison to the total address count. So unless the ISP was only granted a small public block, dedicating some

portion to infrastructure links and loopback addresses (/32) is rarely a large enough issue to outweigh the problems that are potentially caused when private address space is used.

Additionally, specifications and equipment capability improvements now allow for the use of /31 subnets [RFC3021] for link addresses in place of the original /30 subnets - further minimising the impact of dedicating public addresses to infrastructure links by only using two (2) IP addresses per point to point link versus four (4) respectively.

The use of private addressing as a conservation technique within an Internet Service Provider (ISP) core can cause a number of technical and operational issues or effects. The main effects are described below.

3. Effects on Traceroute

The single biggest effect caused by the use of private (RFC1918) addressing within an Internet core is the fact that it can disrupt the operation of traceroute in some situations. This section provides some examples of the issues that can occur.

A first example illustrates the situation where the traceroute crosses an AS boundary and one of the networks has utilised private addressing. The following simple network is used to show the effects.

```
              AS64496                 EBGP                AS64497
                    IBGP Mesh <--------------->  IBGP Mesh

R1 Pool -                                                      R6 Pool -
203.0.113.0/26                                          203.0.113.64/26

                               198.51.100.8/30
                                             198.51.100.4/30
    10.1.1.0/30  10.1.1.4/30                             198.51.100.0/30
                               .9          .10
    .1       .2  .5       .6    ------------    .6      .5  .2      .1
  R1-----------R2-----------R3--|          |--R4----------R5----------R6


  R1 Loopback: 10.1.1.101                    R4 Loopback: 198.51.100.103
  R2 Loopback: 10.1.1.102                    R5 Loopback: 198.51.100.102
  R3 Loopback: 10.1.1.103                    R6 Loopback: 198.51.100.101
```


Using this example, performing the traceroute from AS64497 to AS64496, we can see the private addresses of the infrastructure links in AS64496 are returned.

```
R6#traceroute 203.0.113.1
Type escape sequence to abort.
Tracing the route to 203.0.113.1

  1 198.51.100.2 40 msec 20 msec 32 msec
  2 198.51.100.6 16 msec 20 msec 20 msec
  3 198.51.100.9 20 msec 20 msec 32 msec
  4 10.1.1.5 20 msec 20 msec 20 msec
  5 10.1.1.1 20 msec 20 msec 20 msec
R6#
```

This effect in itself is often not a problem. However, if anti-spoofing controls are applied at network perimeters, then responses returned from hops with private IP addresses will be dropped. Anti-spoofing refers to a security control where traffic with an invalid source address is discarded. Anti-spoofing is further described in BCP 38/RFC 2827.

The effects are illustrated in a second example below. The same network as example 1 is used, but with the addition of anti-spoofing deployed at the ingress of R4 on the R3-R4 interface (ip address 198.51.100.10).

```
R6#traceroute 203.0.113.1

Type escape sequence to abort.
Tracing the route to 203.0.113.1

  1 198.51.100.2 24 msec 20 msec 20 msec
  2 198.51.100.6 20 msec 52 msec 44 msec
  3 198.51.100.9 44 msec 20 msec 32 msec
  4  *  *  *
  5  *  *  *
  6  *  *  *
  7  *  *  *
  8  *  *  *
  9  *  *  *
 10  *  *  *
 11  *  *  *
 12  *  *  *
```

In a third example, a similar effect is caused. If a traceroute is initiated from a router with a private (source) IP address, located

in AS64496 and the destination is outside of the ISPs AS (AS64497), then in this situation the traceroute will fail completely beyond the AS boundary.

```
R1# traceroute 203.0.113.65
Type escape sequence to abort.
Tracing the route to 203.0.113.65

  1 10.1.1.2 20 msec 20 msec 20 msec
  2 10.1.1.6 52 msec 24 msec 40 msec
  3  *  *  *
  4  *  *  *
  5  *  *  *
  6  *  *  *
R1#
```

While it is completely unreasonable to expect a packet with a private source address to be successfully returned in a typical SP environment, the case is included to show the effect as it can have implications for troubleshooting. This case will be referenced in a later section.

In a complex topology, with multiple paths and exit points, the provider will lose their ability to trace paths originating within their own AS, through their network, to destinations within other ASs. Such a situation could be a severe troubleshooting impediment.

For completeness, a fourth example is included to show that a successful traceroute can be achieved by specifying a public source address as the source address of the traceroute. Such an approach can be used in many operational situations if the router initiating the traceroute has at least one public address configured. However, the approach is more cumbersome.


```
R1#traceroute
Protocol [ip]:
Target IP address: 203.0.113.65
Source address: 203.0.113.1
Numeric display [n]:
Timeout in seconds [3]:
Probe count [3]:
Minimum Time to Live [1]:
```

```
Maximum Time to Live [30]: 10
Port Number [33434]:
Loose, Strict, Record, Timestamp, Verbose[none]:
Type escape sequence to abort.
Tracing the route to 203.0.113.65

  1 10.1.1.2 0 msec 4 msec 0 msec
  2 10.1.1.6 0 msec 4 msec 0 msec
  3 198.51.100.10 [AS 64497] 0 msec 4 msec 0 msec
  4 198.51.100.5 [AS 64497] 0 msec 0 msec 4 msec
  5 198.51.100.1 [AS 64497] 0 msec 0 msec 4 msec
R1#
```

It should be noted that some solutions to this problem have been proposed in RFC 5837 which provides extensions to ICMP to allow the identification of interfaces and their components by any combination of the following: ifIndex, IPv4 address, IPv6 address, name, and MTU. However at the time of writing, little or no deployment was known to be in place.

4. Effects on Path MTU Discovery

The Path MTU Discovery (PMTUD) process was designed to allow hosts to make an accurate assessment of the maximum packet size that can be sent across a path without fragmentation. Path MTU Discovery is supported for TCP (and other protocols that support PMTUD such as GRE and IPsec) and works as follows:

o When a router attempts to forward an IP datagram with the Do Not Fragment (DF) bit set out a link that has a lower MTU than the size of the packet, the router MUST drop the packet and return an Internet Control Message Protocol (ICMP) 'destination unreachable - fragmentation needed and DF set (type 3, code 4)' message to the source of the IP datagram. This message includes the MTU of that next-hop network. As a result, the source station which receives the ICMP message, will lower the send Maximum Segment Size (MSS).

It is obviously desirable that packets be sent between two communicating hosts without fragmentation as this process imposes

extra load on the fragmenting router (process of fragmentation), intermediate routers (forwarding additional packets), as well as the receiving host (reassembly of the fragmented packets). Additionally, many applications, including some web servers, set the DF (do not fragment) bit causing undesirable interactions if the path MTU is insufficient. Other TCP implementations may set an MTU size of 576 bytes if PMTUD is unavailable. In addition, IPsec and other tunneling protocols will often require MTUs greater than 1500 bytes and often rely on PMTUD.

While it is uncommon these days for core SP networks not to support path MTUs in excess of 1500 bytes (with 4470 or greater being common), the situation of 1500 byte path MTUs is still common in many ethernet edge or aggregation networks.

The issue is as follows:

o When an ICMP Type 3 Code 4 message is issued from an infrastructure link that uses a private (RFC1918) address, it must be routed back to the originating host. As the originating host will typically be a globally routable IP address, its source address is used as the destination address of the returned ICMP Type 3 packet. At this point there are normally no problems.

o As the returned packet will have an RFC1918 source address, problems can occur when the returned packet passes through an anti-spoofing security control (such as Unicast RPF (uRPF)), other anti-spoofing ACLs, or virtually any perimeter firewall. These devices will typically drop packets with an RFC1918 source address, breaking the successful operation of PMTUD.

As a result, the potential for application level issues may be created.

5. Unexpected interactions with some NAT implementations

Private addressing is legitimately used within many enterprise, corporate or government networks for internal network addressing. When users on the inside of the network require Internet access, they will typically connect through a perimeter router, firewall, or network proxy, that provides Network Address Translation (NAT) or Network Address Port Translation (NAPT) services to a public interface.

Scarcity of public IPv4 addresses, and the transition to IPv6, is forcing many service providers to make use of NAT. CGN (Carrier Grade NAT) will enable service providers to assign private RFC 1918

IPv4 addresses to their customers rather than public, globally unique IPv4 addresses. NAT444 will make use of a double NAT process.

Unpredictable or confusing interactions could occur if traffic such as traceroute, PMTUD and possibly other applications were launched from the NAT IPv4 'inside address' and it passed over the same address range in the public IP core. While such a situation would be unlikely to occur if the NAT pools and the private infrastructure addressing were under the same administration, such a situation could occur in the more typical situation of a NAT'ed corporate network connecting to an ISP. For example, say if 10.1.1.0/24 is used to internally number the corporate network. A traceroute or PMTUD request is initiated inside the corporate network from say 10.1.1.1. The packet passes through a NAT (or NAPT) gateway, then over an ISP core numbered from the same range. When the responses are delivered back to the originator, the returned packets from the privately addressed part of the ISP core could have an identical source and destination address of 10.1.1.1.

```
          NAT Pool -
         203.0.113.0/24

  10.1.1.0/30                 10.1.1.0/30               198.51.100.0/30
            198.51.100.12/30              198.51.100.4/30

  .1       .2  .14     .13  .1           .2  .6      .5  .2      .1
R1-----------R2-----------R3---------------R4----------R5----------R6
             NAT
                                                      R6 Loopback:
                                                      198.51.100.100
```

```
R1#traceroute 198.51.100.100

Type escape sequence to abort.
Tracing the route to 198.51.100.100

  1 10.1.1.2 0 msec 0 msec 0 msec
  2 198.51.100.13 0 msec 4 msec 0 msec
  3 10.1.1.2 0 msec 4 msec 0 msec        <<<<
  4 198.51.100.5 4 msec 0 msec 4 msec
  5 198.51.100.1 0 msec 0 msec 0 msec
R1#
```

This example has been included to illustrate an effect. Whether that effect would be problematic would depend on both the deployment

scenario and the application in use.

Certainly a scenario where the same RFC1918 address space becomes utilised on both the inside and outside interfaces of a NAT/NAPT device can be problematic. For example, the same private address range is assigned by both the administrator of a corporate network and their ISP. Some applications discover the outside address of their local CPE to determine if that address is reserver for special use. Application behavior may then be based on this determination. [weil-shared-transition-space-request] provides further analysis of this situation.

To address this scenario and others, at the time of writing, work was in progress to obtain a dedicated /10 address block for the purpose of Shared CGN (Carrier Grade NAT) Address Space. Please refer to [weil-shared-transition-space-request] for details. The purpose of Shared CGN Address Space is to number CPE (Customer Premise Equipment) interfaces that connect to CGN devices. As explained in [weil-shared-transition-space-request], RFC1918 addressing has issues when used in this deployment scenario.

6. Interactions with edge anti-spoofing techniques

Denial of service attacks and distributed denial of attacks can make use of spoofed source IP addresses in an attempt to obfuscate the source of an attack. RFC2827 (Network Ingress Filtering) strongly recommends that providers of Internet connectivity implement filtering to prevent packets using source addresses outside of their legitimately assigned and advertised prefix ranges. Such filtering should also prevent packets with private source addresses from egressing the AS.

Best security practices for ISPs also strongly recommend that packets with illegitimate source addresses should be dropped at the AS perimeter. Illegitimate source addresses includes private IP (RFC1918) addresses, addresses within the provider's assigned prefix ranges, and bogons (legitimate but unassigned IP addresses). Additionally, packets with private IP destination addresses should also be dropped at the AS perimeter.

If such filtering is properly deployed, then traffic either sourced from, or destined for privately addressed portions of the network should be dropped. Hence the negative consequences on traceroute, PMTUD and regular ping type traffic.

7. Peering using loopbacks

Although not a common technique, some ISPs use the loopback addresses of border routers (ASBRs) for peering, in particular where multiple connections or exchange points exist between the two ISPs. Such a technique is used by some ISPs as the foundation of fine grained traffic engineering and load balancing through the combination of IGP metrics and multi-hop BGP. When private or non-globally reachable addresses are used as loopback addresses, this technique is either not possible, or considerably more complex to implement.

8. DNS Interaction

Many ISPs utilise their DNS to perform both forward and reverse resolution for the infrastructure devices and infrastructure addresses. With a privately numbered core, the ISP itself will still have the capability to perform name resolution of their own infrastructure. However others outside of the autonomous system will not have this capability. At best, they will get a number of unidentified RFC1918 IP addresses returned from a traceroute.

It is also worth noting that in some cases the reverse resolution requests may leak outside of the AS. Such a situation can add load to public DNS servers. Further information on this problem is documented in the internet draft "AS112 Nameserver Operations".

9. Operational and Troubleshooting issues

Previous sections of the document have noted issues relating to network operations and troubleshooting. In particular when private

IP addressing within an ISP core is used, the ability to easily troubleshoot across the AS boundary may be limited. In some cases this may be a serious troubleshooting impediment. In other cases, it may be solved through the use of alternative troubleshooting techniques.

The key point is that the flexibility of initiating an outbound ping or traceroute from a privately numbered section of the network is lost. In a complex topology, with multiple paths and exit points from the AS, the provider may be restricted in their ability to trace paths through the network to other ASs. Such a situation could be a severe troubleshooting impediment.

For users outside of the AS, the loss of the ability to use a traceroute for troubleshooting is very often a serious issue. As soon as many of these people see a row of "* * *" in a traceroute

they often incorrectly assume that a large part of the network is down or inaccessible (e.g. behind a firewall). Operational experience in many large providers has shown that significant confusion can result.

10. Security Considerations

One of the arguments often put forward for the use of private addressing within an ISP is an improvement in the network security. It has been argued that if private addressing is used within the core, the network infrastructure becomes unreachable from outside the providers autonomous system, hence protecting the infrastructure. There is legitimacy to this argument. Certainly if the core is privately numbered and unreachable, it potentially provides a level of isolation in addition to what can be achieved with other techniques, such as infrastructure ACLs, on their own. This is especially true in the event of an ACL misconfiguration, something that does commonly occur as the result of human error.

There are three key security gaps that exist in a privately addressed IP core.

The approach does not protect against reflection attacks if edge anti-spoofing is not deployed. For example, if a packet with spoofed source address corresponding to the networks infrastructure address range, is sent to a host (or other device) attached to the network, that host will send its response directly to the infrastructure address. If such an attack was performed across a large number of hosts, then a successful large scale denial of service attack on the infrastructure could be achieved. This is not to say that a publicly numbered core will protect from the same attack, it won't. The key point is that a reflection attack does get around the apparent security offered in a privately addressed core.

Even if anti-spoofing is deployed at the AS boundary, the border routers will potentially carry routing information for the privately addressed network infrastructure. This can mean that packets with spoofed addresses, corresponding to the private infrastructure addressing, may be considered legitimate by edge anti-spoofing techniques such as Unicast Reverse Path Forwarding - Loose Mode, and forwarded. To avoid this situation, an edge anti-spoofing algorithm such as Unicast Reverse Path Forwarding - Strict Mode, would be required. Strict approaches can be problematic in some environments or where asymmetric traffic paths exist.

The approach on its own does not protect the network infrastructure from directly connected customers (i.e. within the same AS). Unless other security controls, such as access control lists (ACLs), are deployed at the ingress point of the network, customer devices will normally be able to reach, and potentially attack, both core and edge infrastructure devices.

11. Alternate approaches to core network security

Today, hardware-based ACLs, which have minimal to no performance impact, are now widespread. Applying an ACL at the AS perimeter to prevent access to the network core may be a far simpler approach and provide comparable protection to using private addressing, Such a technique is known as an infrastructure ACL (iACL).

In concept, iACLs provide filtering at the edge network which allows traffic to cross the network core, but not to terminate on infrastructure addresses within the core. Proper iACL deployment will normally allow required network management traffic to be passed, such that traceroutes and PMTUD can still operate successfully. For an iACL deployment to be practical, the core network needs to have been addressed with a relatively small number of contiguous address blocks. For this reason, the technique may or may not be practical.

A second approach to preventing external access to the core is IS-IS core hiding. This technique makes use of a fundamental property of the IS-IS protocol which allows link addresses to be removed from the routing table while still allowing loopback addresses to be resolved as next hops for BGP. The technique prevents parties outside the AS from being able to route to infrastructure addresses, while still allowing traceroutes to operate successfully. IS-IS core hiding does not have the same practical requirement for the core to be addressed from a small number of contiguous address blocks as with iACLs. From an operational and troubleshooting perspective, care must be taken to ensure that pings and traceroutes are using source and destination addresses that exist in the routing tables of all routers in the path. i.e. Not hidden link addresses.

A third approach is the use of either an MPLS based IP VPN, or an MPLS based IP Core where the 'P' routers (or Label Switch Routers) do not carry global routing information. As the core 'P' routers (or Label Switch Routers) are only switching labeled traffic, they are effectively not reachable from outside of the MPLS domain. The 'P' routers can optionally be hidden such they do not appear in a traceroute. While this approach isolates the 'P' routers from directed attacks, it does not protect the edge routers - being either a 'PE' router or a Label Edge Router (LER). Obviously there are

numerous other engineering considerations in such an approach, we simply note it as an option.

These techniques may not be suitable for every network, however, there are many circumstances where they can be used successfully without the associated effects of a privately addressing the core.

12. Normative References

[RFC1191] Mogul, J. and S. Deering, "Path MTU Discovery", November 1990.

[RFC1393] Malkin, G., "Traceroute Using an IP Option", January 1993.

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

[RFC2728] Ferguson, P. and D. Senie , "RFC 2827 Network Ingress Filtering, BCP 38", May 2000.

[RFC3021] Retana, A., White, R., Fuller, V., and D. McPherson, "Using 31-Bit Prefixes on IPv4 Point-to-Point Links", December 2000.

[RFC6304] Abley, J. and W. Maton, "AS112 Nameserver Operations", July 2011.

[RFC792] Postel, J., "RFC792 Internet Control Message Protocol", September 1981.

[weil-shared-transition-space-request]
Weil, J., Kuarsingh, V., Donley, C., Liljenstolpe, C., and M. Azinger, "IANA Reserved IPv4 Prefix for Shared CGN Space".

Appendix A. Acknowledgments

The author would like to thank the following people for their input and review - Dan Wing (Cisco Systems), Roland Dobbins (Arbor Networks), Philip Smith (APNIC), Barry Greene (ISC), Anton Ivanov (kot-begemot.co.uk), Ryan Mcdowell (Cisco Systems), Russ White (Cisco Systems), Gregg Schudel (Cisco Systems), Michael Behringer (Cisco Systems), Stephan Millet (Cisco Systems), Tom Petch (BT Connect), Wes George (Time Warner Cable).

The author would also like to acknowledge the use of a variety of NANOG mail archives as references.

Index

Author's Address

Anthony Kirkham
Palo Alto Networks
Level 32, 101 Miller St
North Sydney, New South Wales 2060
Australia

Phone: +61 7 33530902
Email: tkirkham@paloaltonetworks.com