Some Design Choices for IPv6 Networks
draft-ietf-v6ops-design-choices-05

Abstract

This document presents advice on certain routing-related design choices that arise when designing IPv6 networks (both dual-stack and IPv6-only). The intended audience is someone designing an IPv6 network who is knowledgeable about best current practices around IPv4 network design, and wishes to learn the corresponding practices for IPv6.

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 August 26, 2015.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust’s Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of
1. Introduction

This document discusses certain choices that arise when designing a IPv6-only or dual-stack network. The focus is on routing-related design choices that do not usually come up when designing an IPv4-only network. The document presents each choice and the alternatives, and then discusses the pros and cons of the alternatives in detail. Where consensus currently exists around the best practice, this is documented; otherwise the document simply summarizes the current state of the discussion. Thus this document serves to both document the reasoning behind best current practices for IPv6, and to allow a designer to make an informed choice where no such consensus exists.

This document does not present advice on strategies for adding IPv6 to a network, nor does it discuss transition mechanisms. For advice in these areas, see [RFC6180] for general advice, [RFC6782] for wireline service providers, [RFC6342] for mobile network providers, [RFC5963] for exchange point operators, [RFC6883] for content providers, and both [RFC4852] and [RFC7381] for enterprises. Nor
does this document discuss the particulars of creating an IPv6 addressing plan; for advice in this area, see [RFC5375] or [v6-addressing-plan]. The details of ULA usage is also not discussed; for this the reader is referred to [I-D.ietf-v6ops-ula-usage-recommendations].

Finally, this document focuses on unicast routing design only and does not cover multicast or the issues involved in running MPLS over IPv6 transport.

2. Design Choices

Each subsection below presents a design choice and discusses the pros and cons of the various options. If there is consensus in the industry for a particular option, then the consensus position is noted.

2.1. Interfaces

2.1.1. Mix IPv4 and IPv6 on the Same Layer-3 Interface?

If a network is going to carry both IPv4 and IPv6 traffic, as many networks do today, then a fundamental question arises: Should an operator mix IPv4 and IPv6 traffic or keep them separated? More specifically, should the design:

a. Mix IPv4 and IPv6 traffic on the same layer-3 interface, OR

b. Separate IPv4 and IPv6 by using separate interfaces (e.g., two physical links or two VLANs on the same link)?

Option (a) implies a single layer-3 interface at each end of the connection with both IPv4 and IPv6 addresses; while option (b) implies two layer-3 interfaces at each end, one for IPv4 addresses and one with IPv6 addresses.

The advantages of option (a) include:

- Requires only half as many layer 3 interfaces as option (b), thus providing better scaling;
- May require fewer physical ports, thus saving money;
- Can make the QoS implementation much easier (for example, rate-limiting the combined IPv4 and IPv6 traffic to or from a customer);
o Works well in practice, as any increase in IPv6 traffic is usually counter-balanced by a corresponding decrease in IPv4 traffic to or from the same host (ignoring the common pattern of an overall increase in Internet usage);

o And is generally conceptually simpler.

For these reasons, there is a relatively strong consensus in the operator community that option (a) is the preferred way to go. Most networks today use option (a) wherever possible.

However, there can be times when option (b) is the pragmatic choice. Most commonly, option (b) is used to work around limitations in network equipment. One big example is the generally poor level of support today for individual statistics on IPv4 traffic vs IPv6 traffic when option (a) is used. Other, device-specific, limitations exist as well. It is expected that these limitations will go away as support for IPv6 matures, making option (b) less and less attractive until the day that IPv4 is finally turned off.

2.1.2. Interfaces with Only Link-Local Addresses?

As noted in the introduction, this document does not cover the ins and outs around creating an IPv6 addressing plan. However, there is one fundamental question in this area that often arises: Should the interface:

a. Use only link-local addresses ("unnumbered"), OR

b. Have global and/or unique-local) addresses assigned in addition to link-locals?

There are two advantages of unnumbered interfaces. The first advantage is ease of configuration. In a network with a large number of unnumbered interfaces, the operator can just enable an IGP on each router, without going through the tedious process of assigning and tracking the addresses for each interface. The second advantage is security. Since packets with link-local destination addresses should not be routed, it is very difficult to attack the associated interfaces from an off-link device. This implies less effort around maintaining security ACLs.

Countering this advantage are various disadvantages to unnumbered interfaces in IPv6:

o It is not possible to ping an interface that has only a link-local address from a device that is not directly attached to the link. Thus, to troubleshoot, one must typically log into a device that
is directly attached to the device in question, and execute the
ping from there.

- A traceroute passing over the unnumbered link will return the
  loopback or system address of the router, rather than the address
  of the interface itself.

- In cases of parallel point to point links it is difficult to
determine which of the parallel links was taken when attempting to
troubleshoot unless one sends packets directly between the two
attached link-locals on the specific interfaces. Since many
network problems behave differently for traffic to/from a router
than for traffic through the router(s) in question, this can pose
a significant hurdle to some troubleshooting scenarios.

- On some routers, by default the link-layer address of the
  interface is derived from the MAC address assigned to interface.
  When this is done, swapping out the interface hardware (e.g.
  interface card) will cause the link-layer address to change. In
  some cases (peering config, ACLs, etc) this may require additional
  changes. However, many devices allow the link-layer address of an
  interface to be explicitly configured, which avoids this issue.
  This problem should fade away over time as more and more routers
  select interface identifiers according to the rules in [RFC7217].

- The practice of naming router interfaces using DNS names is
difficult and not recommended when using link-locals only. More
generally, it is not recommended to put link-local addresses into
DNS; see [RFC4472].

- It is often not possible to identify the interface or link (in a
database, email, etc) by giving just its address without also
  specifying the link in some manner.

It should be noted that it is quite possible for the same link-local
address to be assigned to multiple interfaces. This can happen
because the MAC address is duplicated (due to manufacturing process
defaults or the use of virtualization), because a device deliberately
re-uses automatically-assigned link-local addresses on different
links, or because an operator manually assigns the same easy-to-type
link-local address to multiple interfaces. All these are allowed in
IPv6 as long as the addresses are used on different links.

For more discussion on the pros and cons, see [RFC7404]. See also
[RFC5375] for IPv6 unicast address assignment considerations.

Today, most operators use numbered interfaces (option b).
2.2. Static Routes

2.2.1. Link-Local Next-Hop in a Static Route?

For the most part, the use of static routes in IPv6 parallels their use in IPv4. There is, however, one exception, which revolves around the choice of next-hop address in the static route. Specifically, should an operator:

a. Use the far-end’s link-local address as the next-hop address, OR

b. Use the far-end’s GUA/ULA address as the next-hop address?

Recall that the IPv6 specs for OSPF [RFC5340] and ISIS [RFC5308] dictate that they always use link-locals for next-hop addresses. For static routes, [RFC4861] section 8 says:

A router MUST be able to determine the link-local address for each of its neighboring routers in order to ensure that the target address in a Redirect message identifies the neighbor router by its link-local address. For static routing, this requirement implies that the next-hop router’s address should be specified using the link-local address of the router.

This implies that using a GUA or ULA as the next hop will prevent a router from sending Redirect messages for packets that "hit" this static route. All this argues for using a link-local as the next-hop address in a static route.

However, there are two cases where using a link-local address as the next-hop clearly does not work. One is when the static route is an indirect (or multi-hop) static route. The second is when the static route is redistributed into another routing protocol. In these cases, the above text from RFC 4861 notwithstanding, either a GUA or ULA must be used.

Furthermore, many network operators are concerned about the dependency of the default link-local address on an underlying MAC address, as described in the previous section.

Today most operators use GUAs as next-hop addresses.

2.3. IGPs
2.3.1. IGP Choice

One of the main decisions for an IPv6 implementer is the choice of IGP (Interior Gateway Protocol) within the network. The primary options are OSPF [RFC2328] [RFC5340] or IS-IS [RFC5120] [RFC5308], though some operators may consider RIP [RFC2080] or non-standardized protocols. Here we limit our discussion to the pros and cons of OSPF vs. IS-IS.

The discussion in this section revolves around the options in the table below:

<table>
<thead>
<tr>
<th>Option</th>
<th>IGP for IPv4</th>
<th>IGP for IPv6</th>
<th>Known to work well</th>
<th>Hard separation</th>
<th>Similar configuration possible</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>IS-IS</td>
<td>IS-IS</td>
<td>YES</td>
<td>-</td>
<td>YES</td>
</tr>
<tr>
<td>b</td>
<td>IS-IS</td>
<td>OSPFv3</td>
<td>-</td>
<td>YES</td>
<td>-</td>
</tr>
<tr>
<td>c</td>
<td>OSPFv2</td>
<td>IS-IS</td>
<td>YES</td>
<td>YES</td>
<td>-</td>
</tr>
<tr>
<td>d</td>
<td>OSPFv2</td>
<td>OSPFv3</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>e</td>
<td>OSPFv3</td>
<td>IS-IS</td>
<td>-</td>
<td>YES</td>
<td>-</td>
</tr>
<tr>
<td>f</td>
<td>OSPFv3</td>
<td>OSPFv3</td>
<td>-</td>
<td>-</td>
<td>YES</td>
</tr>
</tbody>
</table>

Three of the options above are marked as "Known to work well". These options have seen significant deployments and are generally considered to be good choices. The other options represent valid choices, but have not seen widespread use, so it is hard to offer comments on how well they work. In particular, options (e) and (f) use OSPFv3 to route IPv4 [RFC5838], which is still rather new and untested.

A number of options are marked "Hard separation". These options use a different IGP for IPv4 vs IPv6. With these options, a problem with routing IPv6 is unlikely to affect IPv4 or visa-versa.
Three options are marked "Similar configuration possible". This means it is possible (but not required) to use very similar IGP configuration for IPv4 and IPv6: for example, the same area boundaries, area numbering, link costing, etc. If you are happy with your IPv4 IGP design, then this will likely be a consideration. By contrast, the options that use IS-IS for one IP version and OSPF for the other version will require considerably different configuration, and will also require the operations staff to become familiar with the difference between the two protocols.

With option (a), there is an additional choice of whether to run IS-IS in single-topology mode (where IPv4 and IPv6 share a single topology and a single set of link costs[RFC5308]) or multi-topology mode (where IPv4 and IPv6 have separate topologies and potentially different link costs[RFC5120]). A big problem with single-topology mode is that it cannot easily accommodate devices that support IPv4-only or IPv6-only. Thus, today there is general agreement that multi-topology is the right choice as this gives the greatest flexibility in network design.

It should be noted that a number of ISPs have run OSPF as their IPv4 IGP for quite a few years, but have selected IS-IS as their IPv6 IGP. However, there are very few (none?) that have made the reverse choice. This is, in part, because routers generally support more nodes in an IS-IS area than in the corresponding OSPF area, and because IS-IS is seen as more secure because it runs at layer 2.

2.4. BGP

2.4.1. Which Transport for Which Routes?

BGP these days is multi-protocol. It can carry routes from many different families, and it can do this when the BGP session, or more accurately the underlying TCP connection, runs over either IPv4 or IPv6 (here referred to as either "IPv4 transport" or "IPv6 transport"). Given this flexibility, one of the biggest questions when deploying BGP in a dual-stack network is the question of which routes should be carried over sessions using IPv4 transport and which should be carried over sessions using IPv6 transport.

To answer this question, consider the following table:
<table>
<thead>
<tr>
<th>Route Family</th>
<th>Transport</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unlabeled IPv4</td>
<td>IPv4</td>
<td>Works well</td>
</tr>
<tr>
<td>Unlabeled IPv4</td>
<td>IPv6</td>
<td>Next-hop issues</td>
</tr>
<tr>
<td>Unlabeled IPv6</td>
<td>IPv4</td>
<td>Next-hop issues</td>
</tr>
<tr>
<td>Unlabeled IPv6</td>
<td>IPv6</td>
<td>Works well</td>
</tr>
<tr>
<td>Labeled IPv4</td>
<td>IPv4</td>
<td>Works well</td>
</tr>
<tr>
<td>Labeled IPv4</td>
<td>IPv6</td>
<td>Next-hop issues</td>
</tr>
<tr>
<td>Labeled IPv6</td>
<td>IPv4</td>
<td>(6PE) Works well</td>
</tr>
<tr>
<td>Labeled IPv6</td>
<td>IPv6</td>
<td>Needs MPLS over IPv6</td>
</tr>
<tr>
<td>VPN IPv4</td>
<td>IPv4</td>
<td>Works well</td>
</tr>
<tr>
<td>VPN IPv4</td>
<td>IPv6</td>
<td>Next-hop issues</td>
</tr>
<tr>
<td>VPN IPv6</td>
<td>IPv4</td>
<td>(6VPE) Works well</td>
</tr>
<tr>
<td>VPN IPv6</td>
<td>IPv6</td>
<td>Needs MPLS over IPv6</td>
</tr>
</tbody>
</table>

The first column in this table lists various route families, where "unlabeled" means SAFI 1, "labeled" means the routes carry an MPLS label (SAFI 4, see [RFC3107]), and "VPN" means the routes are normally associated with a layer-3 VPN (SAFI 128, see [RFC4364]). The second column lists the protocol used to transport the BGP session, frequently specified by giving either an IPv4 or IPv6 address in the "neighbor" statement.

The third column comments on the combination in the first two columns:

- For combinations marked "Works well", these combinations are widely supported and are generally recommended.
For combinations marked "Next-hop issues", these combinations are less-widely supported and when supported, often have next-hop issues. That is, the next-hop address is typically a v4-mapped IPv6 address, which is based on some IPv4 address on the sending router. This v4-mapped IPv6 address is often not reachable by default using IPv6 routing. One common solution to this problem is to use routing policy to change the next-hop to a different IPv6 address.

For combinations marked as "Needs MPLS over IPv6", these require MPLS over IPv6 for full support, though special policy configuration may allow them to be used with MPLS over IPv4.

Also, it is important to note that changing the set of address families being carried over a BGP session requires the BGP session to be reset (unless something like [I-D.ietf-idr-dynamic-cap] or [I-D.ietf-idr-bgp-multisession] is in use). This is generally more of an issue with eBGP sessions than iBGP sessions: for iBGP sessions it is common practice for a router to have two iBGP sessions, one to each member of a route reflector pair, so one can change the set of address families on first one of the sessions and then the other.

The following subsections discuss specific scenarios in more detail.

### 2.4.1.1. BGP Sessions for Unlabeled Routes

Unlabeled routes are commonly carried on eBGP sessions, as well as on iBGP sessions in networks where Internet traffic is carried unlabeled across the network. In these scenarios, operators today most commonly use two BGP sessions: one session is transported over IPv4 and carries the unlabeled IPv4 routes, while the second session is transported over IPv6 and carries the unlabeled IPv6 routes.

There are several reasons for this choice:

- It gives a clean separation between IPv4 and IPv6. This can be especially useful when first deploying IPv6 and troubleshooting resulting problems.

- This avoids the next-hop problem described in note 1 above.

- The status of the routes follows the status of the underlying transport. If, for example, the IPv6 data path between the two BGP speakers fails, then the IPv6 session between the two speakers will fail and the IPv6 routes will be withdrawn, which will allow the traffic to be re-routed elsewhere. By contrast, if the IPv6 routes were transported over IPv4, then the failure of the IPv6 data path might leave a working IPv4 data path, so the BGP session...
would remain up and the IPv6 routes would not be withdrawn, and thus the IPv6 traffic would be sent into a black hole.

- It avoids resetting the BGP session when adding IPv6 to an existing session, or when removing IPv4 from an existing session.

### 2.4.1.2. BGP sessions for Labeled or VPN Routes

In these scenarios, it is most common today to carry both the IPv4 and IPv6 routes over sessions transported over IPv4. This can be done with either: (a) one session carrying both route families, or (b) two sessions, one for each family.

Using a single session is usually appropriate for an iBGP session going to a route reflector handling both route families. Using a single session here usually means that the BGP session will reset when changing the set of address families, but as noted above, this is usually not a problem when redundant route reflectors are involved.

In eBGP situations, two sessions are usually more appropriate.

### 2.4.2. eBGP Endpoints: Global or Link-Local Addresses?

When running eBGP over IPv6, there are two options for the addresses to use at each end of the eBGP session (or more properly, the underlying TCP session):

a. Use link-local addresses for the eBGP session, OR

b. Use global addresses for the eBGP session.

Note that the choice here is the addresses to use for the eBGP sessions, and not whether the link itself has global (or unique-local) addresses. In particular, it is quite possible for the eBGP session to use link-local addresses even when the link has global addresses.

The big attraction for option (a) is security: an eBGP session using link-local addresses is extremely difficult to attack from a device that is off-link. This provides very strong protection against TCP RST and similar attacks. Though there are other ways to get an equivalent level of security (e.g. GTSM [RFC5082], MD5 [RFC5925], or ACLs), these other ways require additional configuration which can be forgotten or potentially mis-configured.

However, there are a number of small disadvantages to using link-local addresses:
o Using link-local addresses only works for single-hop eBGP sessions; it does not work for multi-hop sessions.

o One must use "next-hop self" at both endpoints, otherwise re-advertising routes learned via eBGP into iBGP will not work. (Some products enable "next-hop self" in this situation automatically).

o Operators and their tools are used to referring to eBGP sessions by address only, something that is not possible with link-local addresses.

o If one is configuring parallel eBGP sessions for IPv4 and IPv6 routes, then using link-local addresses for the IPv6 session introduces extra operational differences between the two sessions which could otherwise be avoided.

o On some products, an eBGP session using a link-local address is more complex to configure than a session that uses a global address.

o If hardware or other issues cause one to move the cable to a different local interface, then reconfiguration is required at both ends: at the local end because the interface has changed (and with link-local addresses, the interface must always be specified along with the address), and at the remote end because the link-local address has likely changed. (Contrast this with using global addresses, where less re-configuration is required at the local end, and no reconfiguration is required at the remote end).

o Finally, a strict application of [RFC2545] forbids running eBGP between link-local addresses, as [RFC2545] requires the BGP next-hop field to contain at least a global address.

For these reasons, most operators today choose to have their eBGP sessions use global addresses.

3. General Observations

There are two themes that run though many of the design choices in this document. This section presents some general discussion on these two themes.

3.1. Use of Link-Local Addresses

The proper use of link-local addresses is a common theme in the IPv6 network design choices. Link-layer addresses are, of course, always
present in an IPv6 network, but current network design practice mostly ignores them, despite efforts such as [RFC7404].

There are three main reasons for this current practice:

- Network operators are concerned about the volatility of link-local addresses based on MAC addresses, despite the fact that this concern can be overcome by manually-configuring link-local addresses;

- It is very difficult to impossible to ping a link-local address from a device that is not on the same subnet. This is a troubleshooting disadvantage, though it can also be viewed as a security advantage.

- Most operators are currently running networks that carry both IPv4 and IPv6 traffic, and wish to harmonize their IPv4 and IPv6 design and operational practices where possible.

3.2. Separation of IPv4 and IPv6

Currently, most operators are running or planning to run networks that carry both IPv4 and IPv6 traffic. Hence the question: To what degree should IPv4 and IPv6 be kept separate? As can be seen above, this breaks into two sub-questions: To what degree should IPv4 and IPv6 traffic be kept separate, and to what degree should IPv4 and IPv6 routing information be kept separate?

The general consensus around the first question is that IPv4 and IPv6 traffic should generally be mixed together. This recommendation is driven by the operational simplicity of mixing the traffic, plus the general observation that the service being offered to the end user is Internet connectivity and most users do not know or care about the differences between IPv4 and IPv6. Thus it is very desirable to mix IPv4 and IPv6 on the same link to the end user. On other links, separation is possible but more operationally complex, though it does occasionally allow the operator to work around limitations on network devices. The situation here is roughly comparable to IP and MPLS traffic: many networks mix the two traffic types on the same links without issues.

By contrast, there is more of an argument for carrying IPv6 routing information over IPv6 transport, while leaving IPv4 routing information on IPv4 transport. By doing this, one gets fate-sharing between the control and data plane for each IP protocol version: if the data plane fails for some reason, then often the control plane will too.
4. IANA Considerations

This document makes no requests of IANA.

5. Security Considerations

This document introduces no new security considerations that are not already documented elsewhere.

The following is a brief list of pointers to documents related to the topics covered above that the reader may wish to review for security considerations.

For general IPv6 security, [RFC4942] provides guidance on security considerations around IPv6 transition and coexistence.

For OSPFv3, the base protocol specification [RFC5340] has a short security considerations section which notes that the fundamental mechanism for protecting OSPFv3 from attacks is the mechanism described in [RFC4552].

For IS-IS, [RFC5308] notes that ISIS for IPv6 raises no new security considerations over ISIS for IPv4 over those documented in [ISO10589] and [RFC5304].

For BGP, [RFC2545] notes that BGP for IPv6 raises no new security considerations over those present in BGP for IPv4. However, there has been much discussion of BGP security recently, and the interested reader is referred to the documents of the IETF’s SIDR working group.

6. Acknowledgements

Many, many people in the V6OPS working group provided comments and suggestions that made their way into this document. A partial list includes: Rajiv Asati, Fred Baker, Michael Behringer, Marc Blanchet, Ron Bonica, Randy Bush, Cameron Byrne, Brian Carpenter, KK Chittimaneni, Tim Chown, Lorenzo Colitti, Gert Doering, Francis Dupont, Bill Fenner, Kedar K Gaonkar, Chris Grundemann, Steinar Haug, Ray Hunter, Joel Jaeggli, Victor Kuarsingh, Jen Linkova, Ivan Pepelnjak, Alexandru Petrescu, Rob Shakir, Mark Smith, Jean-Francois Tremblay, Dave Thaler, Tina Tsou, Eric Vyncke, Dan York, and Xuxiaohu.

The authors would also like to thank Pradeep Jain and Alastair Johnson for helpful comments on a very preliminary version of this document.
7. Informative References

[I-D.ietf-idr-bgp-multisession]

[I-D.ietf-idr-dynamic-cap]

[I-D.ietf-v6ops-ula-usage-recommendations]

[ISO10589]


Authors' Addresses

Philip Matthews
Alcatel-Lucent
600 March Road
Ottawa, Ontario K2K 2E6
Canada

Phone: +1 613-784-3139
Email: philip_matthews@magma.ca

Victor Kuarsingh
Dyn
150 Dow Street
Manchester, NH 03101
USA

Email: victor@jvknet.com