The Internet Protocol Journal

June 2011 Volume 14, Number 2

A Quarterly Technical Publication for Internet and Intranet Professionals

In This Issue

You can download IPJ back issues and find subscription information at: `www.cisco.com/ipj`

ISSN 1944-1134

From The Editor

The process of adding security to various components of Internet architecture reminds me a little bit of the extensive seismic retrofitting that has been going on in California for decades. The process is slow, expensive, and occasionally intensified by a strong earthquake after which new lessons are learned. Over the past 13 years this journal has carried many articles about network security enhancements: *IP Security* (IPSec), *Secure Sockets Layer* (SSL), *Domain Name System Security Extensions* (DNSSEC), Wireless Network Security, and E-mail Security, to name but a few. In this issue we look at routing security again, specifically the efforts underway in the *Secure Inter-Domain Routing* (SIDR) Working Group of the IETF to provide a secure mechanism for route propagation in the *Border Gateway Protocol* (BGP). The article is by Geoff Huston and Randy Bush.

Our second article discusses *Site Multihoming* in IPv6. Multihoming is a fairly common technique in the IPv4 world, but as part of the development and deployment of IPv6, several new and improved solutions have been proposed. Fred Baker gives an overview of these solutions and discusses the implications of each proposal.

By all accounts, *World IPv6 Day* was a successful demonstration and an important step toward deployment of IPv6 in the global Internet. Several major sites left IPv6 connectivity in place after the event, an encouraging sign. Discussions are already underway for another similar event, this time perhaps lasting for as long as a week. Phil Roberts gives an overview of what happened on June 8 and provides pointers to some of the important lessons learned from this experiment.

I want to take a moment to mention the IPJ subscription renewal campaign. As you know, each subscriber is issued a unique subscription ID that, coupled with an e-mail address, gives access to the subscription database by means of a "magic URL." Unfortunately, sometimes the e-mail containing this URL may not arrive in the subscriber's mailbox, perhaps because of spam filtering. Additionally, readers change e-mail addresses as well as postal addresses. If your subscription has expired or you have changed e-mail, postal mail, or delivery preference, send an e-mail to `ipj@cisco.com` with the updated information and we will make sure your subscription is re-instated. The purpose of the renewal campaign is to ensure that we are sending copies of IPJ to the correct addresses and only to those who prefer paper copies. IPJ is always available via our website at `http://cisco.com/ipj`

—Ole J. Jacobsen, Editor and Publisher

`ole@cisco.com`

Securing BGP with BGPsec

by Geoff Huston, APNIC and Randy Bush, IIJ

For many years the fundamental elements of the Internet: *names* and *addresses*, were the source of basic structural vulnerabilities in the network. With the increasing momentum behind the deployment of *Domain Name System Security Extensions* (DNSSEC)[0], there is some cause for optimism that we have the elements of securing the name space now in hand, but what about addresses and routing? In this article we will look at current efforts within the *Internet Engineering Task Force* (IETF) to secure the use of addresses within the routing infrastructure of the Internet, and the status of current work of the *Secure Inter-Domain Routing* (SIDR) Working Group.

We will look at the approach the SIDR Working Group has taken, and examine the architecture and mechanisms that it has adopted as part of this study. This work was undertaken in three stages: the first concentrated on the mechanisms to support attestations relating to addresses and their use; the second looked at how to secure origination of routing announcements; and the third looked at how to secure the transitive part of *Border Gateway Protocol* (BGP) route propagation.

Supporting Attestations About Addresses Through the RPKI

Prior work in the area of securing the Internet routing system has focused on the operation of BGP in an effort to secure the operation of the protocol and validate, as far as is possible, the contents of *BGP Update* messages. Some notable contributions in more than a decade of study include *Secure-BGP* (S-BGP)[1, 16], *Secure Origin BGP* (soBGP)[2], *Pretty Secure BGP* (psBGP)[3], IRR[4], and the use of an *Autonomous System* (AS) *Resource Record* (RR) in the *Domain Name System* (DNS), signed by DNSSEC[5].

The common factor in this prior work was that they all required, as a primary input, a means of validating basic assertions relating to origination of a route into the interdomain routing system: that the IP address block and the AS numbers being used are valid and that the parties using these IP addresses and AS numbers in the context of routing advertisement are properly authorized to so do.

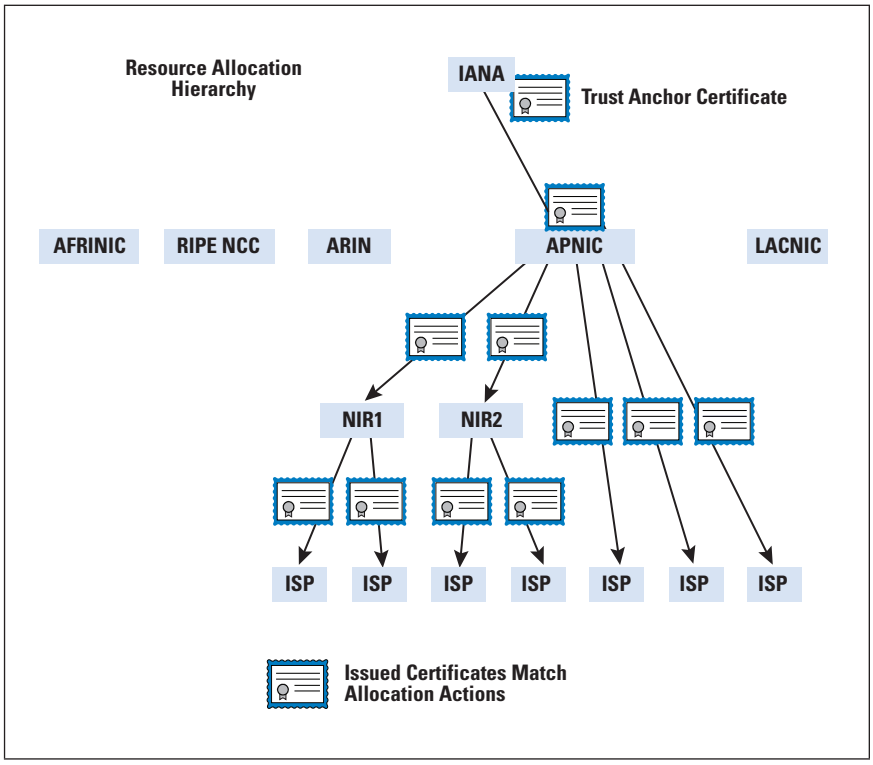
The approach adopted by SIDR for the way in which trust is formalized in the routing environment is through the use of *Resource Certificates*. These certificates are X.509 certificates that conform to the *Public-Key Infrastructure X.509* (PKIX) profile[6]. They also contain an extension field that lists a collection of IP resources (IPv4 addresses, IPv6 addresses, and AS Numbers)[7]. These certificates attest that the certificate issuer has granted to the certificate subject a unique "right-of-use" for the associated set of IP resources, by virtue of a resource allocation action.

This concept mirrors the resource allocation framework of the Internet *Assigned Numbers Authority* (IANA), the *Regional Internet Registries* (RIRs), operators, and others, and the certificate provides a means for a third party (relying party) to formally validate assertions related to resource allocations[8].

The hierarchy of the *Resource Public Key Infrastructure* (RPKI) is based on the administrative resource allocation hierarchy, where resources are distributed from the IANA to the RIRs, *Local Internet Registries* (LIRs), *National Internet Registries* (NIRs), and end users. The RPKI mirrors this allocation hierarchy with certificates that match current resource allocations (Figure 1).

The *Certification Authorities* (CAs) in this RPKI correspond to entities that have been allocated resources. Those entities are able to sign authorities and attestations, and to do so they use specific-purpose *End Entity* (EE) certificates. This additional level of indirection allows the entity to customize each issued authority for specific subsets of number resources that are administered by this entity. Through the use of single-use EE certificates, the issuer can control the validity of the signed authority through the ability to revoke the EE certificate used to sign the authority. As is often the case, a level of indirection comes in handy.

Figure 1: Hierarchy of the RPKI



Signed attestations relating to addresses and their use in routing are generated by selecting a subset of resources that will be the subject of the attestation, by generating an EE certificate that lists these resources, and by specifying validity dates in the EE certificate that correspond to the validity dates of the authority. The authority is published in the RPKI repository publication point of the entity. The RPKI makes conventional use of *Certificate Revocation Lists* (CRLs) to revoke certificates that have not expired but are no longer valid. Every Certification Authority in the RPKI regularly issues a CRL according to the declared CRL update cycle of the Certification Authority. A Certification Authority certificate may be revoked by an issuing authority for numerous reasons, including key rollover, the reduction in the resource set associated with the certificate subject, or termination of the resource allocation. To invalidate an object that can be verified by a given EE certificate, the Certification Authority that issued the EE certificate can revoke the corresponding EE certificate.

The RPKI uses a distributed publication framework, wherein each Certification Authority publishes its products (including EE certificates, CRLs, and signed objects) at a location of its choosing. The set of all such repositories forms a complete information space, and it is fundamental to the model of securing BGP in the public Internet that the entire RPKI information space be available to every *Relying Party* (RP). It is the role of each RP to maintain a local cache of the entire distributed repository collection by regularly synchronizing each element in the local cache against the original repository publication point. To assist RPs in the synchronization task, each RPKI publication point uses a *manifest*, a signed object that lists the names (and hash values) of all the objects published at that publication point. It is used to assist RPs to ensure that they have managed to synchronize against a complete copy of the material published at the Certification Authority publication point.

The utility of the RPKI lies in its ability to validate digitally signed information and, therefore, give relying parties some confidence in the validity of signed attestations about addresses and their use. The particular utility of the RPKI is not as a means of validation of attestations of an individual's identity or that individual's role, but as a means of validating that person's authority to use IP address resources. Although it is possible to digitally sign any digital object, it has been suggested that the RPKI system uses a very small number of standard signed objects that have particular meaning in the context of routing security.

Securing Route Origination

The approach adopted by SIDR to secure origination of routing information is one that uses a particular signed authority, a *Route Origination Authorization* (ROA)[10]. An ROA is an authority created by a prefix holder that authorizes an AS to originate one or more specific route advertisements into the interdomain routing system.

An ROA is a digital object formatted according to the *Cryptographic Message Syntax Specification* (CMS)[11] that contains a list of address prefixes and one AS number. The AS is the specific AS being authorized to originate route advertisements for one or more of the address prefixes in the ROA. The CMS object also includes the EE resource certificate for the key used to verify the ROA. The IP Address extension in this EE certificate must encompass the IP address prefixes listed in the ROA contents.

The ROA conveys a simple authority. It does not convey any further routing policy information, nor does it convey whether or not the AS holder has even consented to actually announce the prefix(es) into the routing system. The associated EE certificate is used to control the validity of the ROA, and the CMS wrapper is used to securely bind the ROA and the EE certificate within a single signed structure.

There is one special ROA, one that authorizes AS 0 to originate a route. Because AS 0 is a reserved AS that should never be used by a BGP speaker, this ROA is a "negative" authority, used to indicate that no AS has authority to originate a route for the address prefix(es) listed in the ROA.

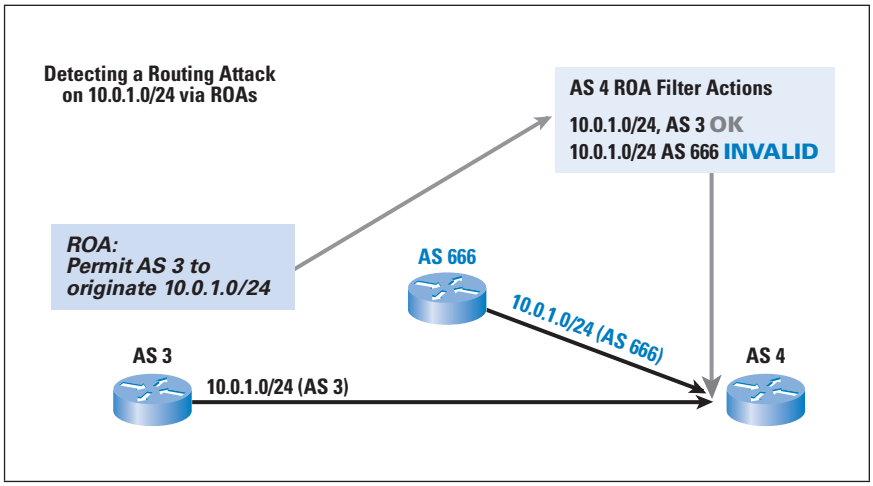
If the entire routing system were to be populated with ROAs, then identification of an invalid route advertisement would be directly related to detection of an invalid ROA or a missing ROA. However, in a more likely scenario of partial use of ROAs (such as when only some legitimate route originations are authorized in a ROA), the absence of an ROA cannot be interpreted simply as an unauthorized use of an address prefix. This scenario leads to the use of a tri-state validation process for routes, as follows.

If a given route matches exactly the information contained in an ROA whose EE certificate can be validated in the RPKI (a "valid" ROA), then the route can be regarded as a "valid" origination. Where the address prefix matches that in a valid ROA but the origination AS does not match the AS number in the ROA, and there are no other valid ROAs that explicitly validate the announcing AS, then the route can be considered to be "invalid." Also, where the address prefix is more specific than that of a valid ROA, and there are no other valid ROAs that match the prefix, then the route can also be considered "invalid." Where the prefix in a route is not described in any ROA and is not a more specific prefix of any ROA, the route has an "unknown" validation outcome.

These three potential outcomes can be considered a set of relative local preferences. Routes whose origins can be considered "valid" are generally proposed to be preferred over routes whose origins are unknown, which, in turn, can generally be preferred over routes whose origins are considered invalid. However, such relative preferences are a matter to be determined by local routing policy. Local policies may choose to adopt a stricter policy and, for example, discard routes with an invalid validation outcome[12].

The way in which ROAs are used to validate the origin of routes in BGP differs from many previous proposals for securing BGP. In this framework the ROAs are published in the RPKI distributed repository framework. Each RP can use the locally cached collection of valid ROAs to create a validation filter collection, with each element of the set containing an address, prefix size constraints, and an originating AS. It is this filter set—rather than the ROAs themselves—that are fed to the local routers[13]. An example of the way in which ROAs can be used to detect prefix hijack attempts is shown in Figure 2.

Figure 2: Use of ROAs to detect Unauthorized Route Origination



The model of injecting validation of origination into the BGP domain is an example of a highly modular and piecemeal deployment. There are no changes to the BGP protocol for this origin validation part of the secure routing framework.

The process of securing origination starts with the address holder, who generates local keys and requests certification of their address space from the entity from whom their addresses were allocated or assigned. With this Certification Authority resource certificate, the address holder is then in a position to generate an EE certificate and a ROA that assigns an authority for a nominated AS to advertise a route for an address prefix drawn from its address holdings. The one condition here is that if an address holder issues a ROA for an address prefix providing an authority for one AS to originate a route for this prefix, then the address holder is required to issue ROAs for all the ASs that have been similarly authorized to originate a route for this address prefix. The address holder publishes this ROA in its publication point in the distributed RPKI repository structure.

Relying parties can configure a locally managed cache of the distributed RPKI repository and collect the set of valid ROAs. They can then, with the dedicated RPKI cache-to-router protocol[13], maintain, on a set of "client" routers, the set of address prefix/originating AS authorities that are described in valid ROAs. The BGP-speaking router can use this information as an input to the local route decision process.

This model of operation supports piecemeal incremental deployment, wherein individual address holders may issue ROAs to authorized routing advertisements independent of the actions of other address holders. Also, ASs may deploy local validation of route origination independently of the actions of other ASs. And given that there are no changes to the operation of BGP, then there are no complex inter-dependencies that hinder piecemeal incremental deployment of this particular aspect of securing routing.

Securing Route Propagation: BGPsec

Origin validation as described earlier does not provide cryptographic assurance that the origin AS in a received BGP route was indeed the originating AS of this route. A malicious BGP speaker can synthesize a route as if it came from the authorized AS. Thus, it is very useful in detecting accidental misannouncements, but origination validation does little to prevent malicious routing attacks from a determined attacker.

In looking at the operation of the BGP protocol, some parts of the protocol interaction are strictly local between two BGP-speaking peers, such as advising a peer of local attributes. Another part of the BGP protocol is a "chained" interaction, in which each AS adds information to the protocol object. This attribute of a BGP update, the *AS Path,* is not only useful to detect and prevent routing loops, it is also used in the BGP best-path-selection algorithm.

A related routing security question concerns the validity of this "chained" information, namely the AS Path information contained in a route. Within the operation of the BGP protocol, each AS that propagates an update to its AS neighbors is required to add its AS number to the AS Path sequence. The inference is that at any stage in the propagation of a route through the interdomain routing system, the AS Path represents a viable AS transit sequence from the local AS to the AS originating the route. This AS Path attribute of a route is used for loop detection. Locally, the AS Path may also be used as input to a local route policy process, using the length of the AS Path as a route metric.

Attacks on the AS Path can be used to subvert the routing environment. A malicious BGP speaker may manipulate the AS Path to prevent an AS from accepting a route by adding its AS number to the AS Path, or it may attempt to make a particular route more likely to be selected by a remote AS by stripping out ASs from the AS Path. Accordingly, it is important to equip a secure BGP framework with the ability to validate the authenticity of the AS Path presented in a BGP update[14].

When attempting to validate an AS path, many potential validation questions must be addressed.

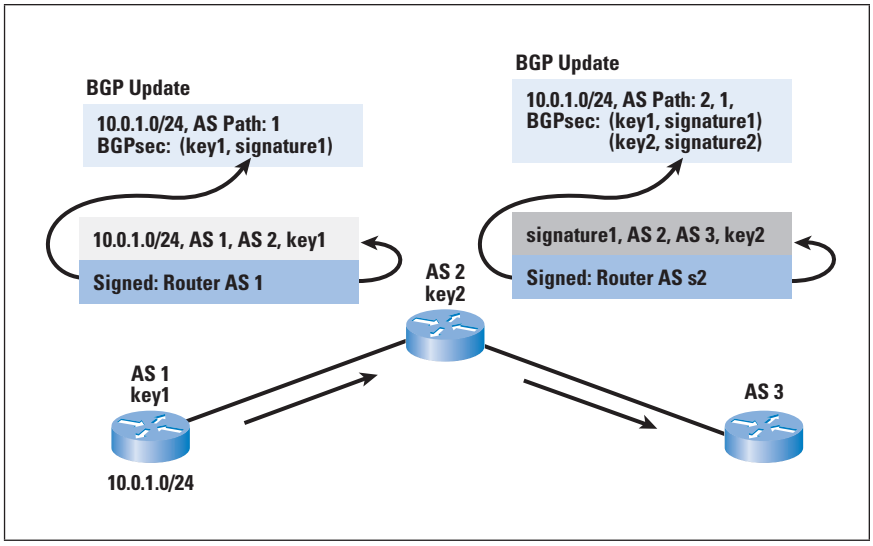
- The first and weakest question is: Are all ASs in the AS Path valid ASs?
- A slightly stronger validation question is: Do all the AS pairs in the AS Path represent valid AS adjacencies (where both ASs in the pair-wise association are willing to attest to their mutual adjacency in BGP)?
- A even stronger question is: Does the sequence of ASs in the AS Path represent the actual propagation path of the BGP route object?

This last question forms the basis for the SIDR activity in defining an AS Path validation framework, BGPsec. This attempt is to assure a BGP speaker that the operation of the BGP protocol is operating correctly and that the content of a BGP update correctly represents the inter-AS propagation path of the update from the point of origination to the receiver of the route. This tool is not the same as a policy validation tool and it does not necessarily assure the receiver of the route that this update conforms to the routing policies of neighboring BGP speakers. This route also does not necessarily reflect the policy intent of the originator of the route. The BGPsec framework proposed for securing the AS Path also uses a local RPKI cache, but it includes an additional element of certification. The additional element of the security credentials used here is an extension to the certification of AS numbers with a set of operational keys and their associated certificates used for signing update messages on *External Border Gateway Protocol* (eBGP) routers in the AS. These "router certificates" can sign BGP update attributes in the routing infrastructure, and the signature can be interpreted as being a signature made "in the name of" an AS number.

In the BGPsec framework, eBGP-speaking routers within the AS have the ability to "sign" a BGP update before sending it. In this case, the added signature "covers" the signature of the received BGP update, the local AS number, the AS number to which the update is being sent, as well as a hash of the public key part of the router key pair used to sign route updates.

The couplet of the public key hash and the signature itself are added to the BGP protocol update as BGPsec update attributes. As the update traverses a sequence of transit ASs, each eBGP speaker at the egress of each AS adds its own public key hash and digital signature to the BGPsec attribute sequence (Figure 3).

Figure 3: BGPsec AS Path Protection



This interlocking of signatures allows a receiver of a BGP update to use the interlocking chain of digital signatures to validate (for each AS in the AS Path) that the corresponding signature was correctly generated "in the name of" that AS in the AS Path, and that the next AS in the path matches the next AS in the signed material. The "forward signing" that includes the AS to which the update is being sent prevents a man-in-the-middle attack of the form of taking a legitimate outbound route announcement destined for one neighbor AS and redirecting it to another AS. But this signing of the AS Path is not quite enough to secure the route update, because the AS Path needs to be coupled to the actual address prefix by the route originator. The route originator needs to sign across not only the local AS and the AS to whom the route update is being sent, but also the address prefix and the expiry time of the route. This action allows the path to be "bound" to the prefix and prevents a man-in-the-middle from splicing a signed path or signed-path fragment against a different prefix.

If the signatures that "span" the AS Path in the BGP update can all be validated, then the receiver of the BGP update can validate, in a cryptographic sense, the currency of the routing update. It can also validate that the route update was propagated across the inter-AS routing space in a manner that is faithfully represented in the AS Path of the route.

The expiry time of the EE certificates used in conjunction with signed route updates introduces a new behavior into BGPsec. In the context of BGP, an announced route remains current until it is explicitly withdrawn or until the peer session that announced the route goes down. This property of BGP introduces the possibility of "ghost-route" attacks in BGP, wherein a BGP speaker fails to propagate a withdrawal in order to divert the consequent misdirected traffic from its peers.

In BGPsec, all route advertisements are given an expiry time by the originator of the route. This expiry time corresponds to the "notAfter" time of the EE certificate used to sign the protocol update, after which time the route is considered invalid. The implication is that a route originator is required to readvertise the route, and refresh the implicit expiry timer of the associated digital signature at regular intervals.

This approach to route-update validation is not quite the "light-touch" of origination validation. In this case the mechanism requires the use of a new BGP attribute and negotiation of a new BGP capability between eBGP peers, in turn meaning that the model of incremental deployment is one that is more "viral" than truly piecemeal. By "viral" we mean that this model is one of incremental deployment in which direct eBGP peers of a BGPsec-speaking AS will be able to speak BGPsec between themselves in a meaningful way. In turn these adjacent ASs can offer to speak BGPsec with their eBGP peers, and so on. This reality does not imply that BGPsec deployment must necessarily start from a single AS, but it does imply that communities of interconnected ASs all speaking BGPsec will be able to provide assurance via BGPsec on those routes originated and propagated within that community of interconnected ASs. It also implies that the greatest level of benefit to adopters of secure BGP will be realized by ASs that adopt BGPsec as a connected community of ASs.

Other changes to the behavior of BGP are implied by this mechanism. BGP conventionally permits "update packing," where numerous address prefixes can be placed in a single update message if they share a common collection of attributes, including the AS Path. At this stage it appears that such update packing would not be supported in secure BGP, and each update in secure BGP would refer to a single prefix. Obviously this situation would have some effect on the level of BGP traffic, but early experiments suggest not at an unreasonable cost.

There are further effects on BGP that have not been fully quantified in studies to date. The addition of a compound attribute of a signature and a public key identifier for every AS in the AS Path has size implications on the amount of local storage a secure BGP speaker will need to store these additional per-prefix per-peer attributes. It also has broader implications if used in conjunction with current proposals for multipath BGP where multiple paths, in addition to the "best" path, are propagated to eBGP peers. Also, the computational load of validation of signatures in secure BGP is significantly higher in terms of the number of cryptographic operations that are required to validate a BGP update.

However, BGPsec is not intended to "tunnel" across those parts of the interdomain routing space that do not support BGPsec capabilities. When an update leaves a BGPsec realm, the BGPsec signature attributes of the route are stripped out, so the storage overheads of BGPsec are not seen by other BGP speakers.

Similarly, the periodic updates that result from the expiry timer should not propagate beyond the BGPsec realm. If the boundary is prepared to perform BGP update packing to non-BGPsec peers, then even the unpacked update overhead is not carried outside of the BGPsec realm.

It is also noted that the "full" load of BGPsec would only necessarily be carried by "transit" ASs; that is, those ASs that propagate routes on behalf of other ASs. Historically we see some 15 percent of ASs are "transit" ASs, while all other ASs behave as "stub" ASs that only originate routes and do not appear to transit routes for others. Such stub ASs can support a "lightweight" simplex version of BGPsec that can either point a default route to its upstream AS provider or trust its upstream ASs to perform BGPsec validation. In this case the stub AS needs to provide BGPsec signed originated routes to its upstream ASs, but no more.

Conclusion

The work on the specification of the RPKI itself and the specification of origin validation is nearing a point of logical completion of the first phase of standardization within the IETF, and the working draft documents are being passed from the working group into the review process leading to their publication as proposed standard RFCs. The RIRs are in the process of launching their RPKI services based on these specifications, and the initial deployment of working code has been made by numerous parties, who are also working on integration of origination validation in BGP implementations.

The work on securing the AS Path is at an earlier phase in the development process, and the SIDR Working Group is considering the initial design material. It is expected to take a similar path of further review and refinement in light of developing experience and study of the proposed approach.

The RPKI has been designed as a robust and simple framework. As far as possible, existing standards, technologies, and processes have been exploited, reflecting the conservatism of the routing community and the difficulty in securing rapid, widespread adoption of novel technologies.

Acknowledgements

The work described here is the outcome of the efforts of many individuals who have contributed to securing BGP over a period that now spans two decades, and certainly too many to ensure that all the contributors are recognized here. Instead, the authors would like to acknowledge their work and trust that the mechanisms described here are a faithful representation of the cumulative sum of their various contributions.

References

[0] Miek Gieben, "DNSSEC: The Protocol, Deployment, and a Bit of Development," *The Internet Protocol Journal,* Volume 7, No. 2, June 2004.

[1] Stephen Kent, Charlie Lynn, and Karen Seo, "Secure Border Gateway Protocol (S-BGP)," *IEEE Journal on Selected Areas in Communications,* Volume 18, No. 4, pp 582–592, April 2000.

[2] Russ White, "Securing BGP through secure origin BGP," *The Internet Protocol Journal,* Volume 6, No. 3, September 2003.

[3] Paul van Oorschot, Tao Wan, and Evangelos Kranakis, "On Inter-domain Routing Security and Pretty Secure BGP (psBGP)," *ACM Transactions on Information and System Security,* Volume 10, No. 3, July 2007.

[4] Geoffrey Goodell, William Aiello, Timothy Griffin, John Ioannidis, and Patrick D. McDaniel, "Working Around BGP: An Incremental Approach to Improving Security and Accuracy of Interdomain Routing," *Proceedings of Internet Society Symposium on Network and Distributed System Security* (NDSS '03), February 2003.

[5] Tony Bates, Randy Bush, Tony Li, and Yakov Rekhter, "DNS-based NLRI origin AS verification in BGP," Internet Draft, Work in Progress, July 1998.

[6] David Cooper et al., "Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile," RFC 5280, May 2008.

[7] Charlie Lynn, Stephen Kent, and Karen Seo, "X.509 Extensions for IP Addresses and AS Identifiers," RFC 3779, June 2004.

[8] Matt Lepinski and Stephen Kent, "An Infrastructure to Support Secure Internet Routing," Internet Draft, Work in Progress, February 2008.

[9] Geoff Huston, George Michaelson, and Robert Loomans, "A Profile for X.509 PKIX Resource Certificates," Internet Draft, Work in Progress, September 2008.

[10] Matt Lepinski, Stephen Kent, and Derrick Kong, "A Profile for Route Origin Authorizations (ROAs)," Internet Draft, Work in Progress, July 2008.

[11] Russ Housley, "Cryptographic Message Syntax (CMS)," RFC 3852, July 2004.

[12] Geoff Huston and George Michaelson, "Validation of Route Origination using the Resource Certificate PKI and ROAs," Internet Draft, Work in Progress, November 2010.

[13] Randy Bush and Rob Austein, "The RPKI/Router Protocol," Internet Draft, Work in Progress, March 2011.

[14] Kim Zetter, "Revealed: The Internet's Biggest Security Hole," *WIRED,* August 2008, **http://www.wired.com/threatlevel/2008/08/revealed-the-in/**

[15] Geoff Huston, "Resource Certification," *The Internet Protocol Journal,* Volume 12, No. 1, March 2009.

[16] Stephen Kent, "Securing the Border Gateway Protocol," *The Internet Protocol Journal,* Volume 6, No. 3, September 2003.

Ed.: A version of this article also appeared in *The IETF Journal,* Volume 7, Issue 1, July 2011. *The IETF Journal* can be obtained from: **http://isoc.org/ietfjournal/**

GEOFF HUSTON, B.Sc., M.Sc., is the Chief Scientist at APNIC, the Regional Internet Registry serving the Asia Pacific region. He has been closely involved with the development of the Internet for many years, particularly within Australia, where he was responsible for the initial build of the Internet within the Australian academic and research sector. He is author of numerous Internet-related books, and was a member of the Internet Architecture Board from 1999 until 2005; he served on the Board of Trustees of the Internet Society from 1992 until 2001. E-mail: **gih@apnic.net**

RANDY BUSH is a Research Fellow and Network Operator at Internet Initiative Japan (IIJ), Japan's first commercial ISP. He specializes in network measurement, especially routing, network security, routing protocols, and IPv6 deployment. Randy has been in computing for 45 years, and has a few decades of Internet operations experience. He was the engineering founder of Verio, which is now NTT/Verio. He has been heavily involved in transferring Internet technologies to developing economies for more than 20 years. E-mail: **randy@psg.com**

Views of IPv6 Site Multihoming

by Fred Baker, Cisco Systems

In today's Internet, *site multihoming*—an edge network configuration that has more than one service provider but does not provide transit communication between them—is relatively common. Per the statistics at **www.potaroo.net**, almost 40,000 *Autonomous Systems* are in the network, of which about 5,000 seem to offer transit services to one or more customers. The rest are in terminal positions, possibly meaning three things. They could be access networks, broadband providers offering Internet access to small companies and residential customers; they could be multihomed edge networks; or they might be networks that intend to multihome at some point in the future. The vast majority, on the order of 75 percent, are multihomed or intend to multihome. That is but one measure; you do not have to use *Border Gateway Protocol* (BGP) routing to have multiple upstream networks. Current estimates suggest that there is one multihomed entity per 50,000 people worldwide, and one per 18,000 in the United States.

We also expect site multihoming to become more common. A current proposal in Japan suggests that each home might be multihomed; it would have one upstream connection for Internet TV, and one or more other connections provided by *Internet Service Providers* (ISPs), operating over a common *Digital Subscriber Line* (DSL) or fiber-optic infrastructure. That scenario has one multihomed entity for every four people.

Why do edge networks multihome? Reasons vary. In the Japanese case just propounded, it is a fact of life—users have no other option. In many cases, it is a result of a work arrangement, or a strategy for achieving network reliability through redundancy.

For present purposes, this article considers scaling targets derived from a world of 10 billion people (circa 2050), and a ratio of one multihomed entity per thousand people—on the order of 10,000,000 multihomed entities at the edge of the Internet. Those estimates may not be accurate 40 years from now, but given current trends they seem like reasonable guesses.

RFC 1726[1], the technical criteria considered in the selection of what at the time was called *IP Next Generation* (IPng), did not mention multihoming per se. Even so, among the requirements are scalable and flexible routing, of which multihoming is a special case. When IPv6 was selected as the "next generation," multihoming was one of the topics discussed. The Internet community has complained that this particular goal was not fulfilled. Several proposals have been proffered; unfortunately, each has benefits, and each has concerns. No single perfect solution is universally accepted.

In this article, I would like to look at the alternatives proposed and consider the effects they have. In this context, the goals set forth in RFC 3582[2] are important; many people tried to state what they would like from a multihoming architecture, and the result was a set of goals that solutions only asymptotically approach.

The proposals considered in this article include:

- *Provider Independent Addressing,* also known as *BGP Multihoming*
- *Exchange-Based Addressing*
- *Shim6,* also known as *Level 3 Multihoming*
- *Identifier-Locator Network Protocol* (ILNP)
- *Network Prefix Translation,* also known as *NAT66*

BGP Multihoming

BGP Multihoming involves a mechanism relatively common in the IPv4 Internet; the edge network either becomes a member of a *Regional Internet Registry* (RIR) [APNIC, RIPE, LACNIC, AFRINIC, ARIN] and from that source obtains a *Provider-Independent* (PI) prefix, or obtains a *Provider-Allocated* (PA) prefix from one provider and negotiates contracts with others using the same prefix. In any case, it advertises the prefix in BGP, meaning that all ISPs—including in the PA case—the provider that allocated it, must carry it as a separate route in their routing tables.

The benefit to the edge is easily explained, and in the case of large organizations it is substantial. Consider the case of Cisco Systems, whose internal network rivals medium-sized ISPs for size and complexity. With about 30 *Points of Attachment* (PoAs) to the global Internet, and at least as many service providers, Cisco has an IPv6 /32 PI prefix, and hundreds of offices to interconnect using it. One possible way to enumerate the Cisco network would be to use the next five bits of its address (32 /37 prefixes) at its PoAs, and allocate prefixes to its offices by the rule that if their default route is to a given PoA, their addresses are derived from that PoA. By advertising the PoAs /37 and a backup /32 into the Internet core at each PoA, Cisco could obtain effective global routing. It would also obtain relative simplicity for its internal network—only one subnet is needed on any given *Local-Area Network* (LAN) regardless of provider count or addressing, and routing can be optimized independently from the outside world.

The problem that arises with PI addressing, if taken to its logical extreme, is that the size of the routing table explodes. If every edge network obtains a PI prefix—neglecting for the moment both BGP traffic engineering and the kind of de-aggregation suggested in Cisco's case—the logical outcome of enumerating the edge is a routing table with on the order of 107 routes. The memory required to store the routing table, and in the *Secure Interdomain Routing* (SIDR) case the certificates that secure it, is one of the factors in the cost of equipment. The volume of information also affects the time it takes to advertise a full routing table, and in the end the amount of power that a router uses, the heat it produces, and a switching center's air conditioning requirements. Thus both the capital cost of equipment used in transit networks and the cost of operations would be affected. In effect, the Internet becomes the "poster child" for the *Tragedy of the Commons*.

Exchange-Based Addressing

Steve Deering proposed the concept of exchange-based addressing at the IETF meeting in Stockholm in 1995, under the name *Metropolitan Addressing*. In this model, prefixes do not map to companies, but to Internet exchange consortia, likely regional. One organizing principle might be to associate an Internet exchange with each commercial airport worldwide, about 4000 total, resulting in a global routing table on the same order of magnitude in size. Edge networks, including residential networks, within that domain obtain their prefix from the exchange, and they are used by any or all ISPs in the region. Routes advertized to other regions, even within the same ISP, are aggregated to the consortium prefix.

The benefits to the edge network in exchange-based addressing are similar to the benefits of PI addressing for a large corporation. In effect, the edge networks served by an exchange consortium behave like the "departments" of a "user consortium," and they enjoy great independence from their upstream providers. They can multihome or move between providers without changing their addressing, and on a global scale the routing table is contained to a small multiple of the number of such consortia.

However, the benefit to users is in most cases a detriment to their ISPs; the ISPs are forced to maintain routes to each user network served by the consortium—or at least routes for their own customers and a default route to the exchange. Thus, the complexity of routing is moved from the transit core to the access networks serving regional consortia. In addition, if there is no impediment to a user flitting among ISPs, users can be expected to flit, imposing business costs.

The biggest short-term effect on the ISP might well be the reengineering of its transit contracts. In today's Internet, a datagram sent by users to their ISPs is quickly shuttled to the destination's ISPs, which then carry it over the long haul. In an exchange-based network, there is no way to remotely determine which local ISP or ISP instance is serving a given customer.

Hence, the sender's ISP carries the datagram until it reaches the remote consortium, whence it switches to the access network serving the destination. One could argue that a "sender-pays" model might have benefits, but it is very different from the present model.

The edge network has problems, too. If the edge network is sufficiently distributed, it will have services in several exchange consortia, and therefore several prefixes. Although there is nothing inherently bad about that, it may not fit the way a cloud computing environment wants to move virtual hosts around, or miss other requirements.

Level 3 Multihoming: Shim6

The IETF's *shim6* model[9] starts from the premise that edge networks obtain their prefixes from their upstream ISPs—PA Addressing. If a typical residential or small business does so, there is no question of advertising its individual route everywhere; the ISP can route internally as its needs to, but globally, the number of ISPs directs the size of the routing table. If that is, as `potaroo` suggests, on the order of 10,000, the size of the routing table will be on the same order of magnitude.

The benefit to the ISP should be obvious; it does not have to change its transit contracts, and although there will be other concerns, it does not have the routing table ballooning memory costs or route exchange latencies.

However, as exchange-based addressing moves operational complexity from the transit core to the access network, shim6 moves such complexities to the edge network itself and to the host in it. If a network has multiple upstream providers, each LAN in it will carry a subnet from each of those providers—not one subnet per LAN, but as many as the providers of the host's LAN will use. At this point, the ingress filtering of RFC 3704[21] at the provider becomes a problem at the edge; the host must select a reasonable address for any session it opens, and must do so in the absence of specific knowledge of network routing. A wrong guess can have dramatic effects; a session routed to the wrong provider may not work at all, and an unfortunate address choice can change end-to-end latency from tens of milliseconds to hundreds or worse by virtue of backbone routing.

Application layer referrals and other application uses of addresses also have difficulties. Although the address a session is using will work both within and without the network, if a host has more than one address, one of the other addresses may be more appropriate to a given use. Hence, the application that really wants to use addresses is saddled with finding all of the addresses that its own host or a peer host might have.

There is also an opportunity. TCP today associates sessions with their source and destination addresses. The shim6 model, implemented in the *Stream Control Transmission Protocol* (SCTP)[17] and *Multipath TCP* (MPTCP)[16], allows a session to change its addresses, meaning that a session can survive a service provider outage. Doing the same in TCP requires the insertion of a shim protocol between IP and TCP; at the Internet layer, the address might change, but the shim tracks the addresses for TCP.

There are, of course, ways to solve the outstanding problems. For simple cases, RFC 3484[3, 4] describes an address-selection algorithm that has some promise. In the Japanese case, a residential host might use link-local addresses within its own network, addresses appropriate to the television service on its TV and set-top box, and an ISP's prefix for everything else. If there is more than one router in the residential LAN serving more than one ISP, exit routing can be accomplished by having the host send data using an ISP's source address to the router from which it learned the prefix. When the network becomes more complex, though, we are looking at new routing protocols that can route based on a combination of the source and the destination addresses, and we are looking at network management methodologies that make address management simpler than it is today, adding and dropping subnets on LANs—and as a result renumbering networks—without difficulty. It also implies a change to the typical host implementing the shim protocol. Those technologies either do not exist or are not widely implemented today.

Identifier-Locator Network Protocol

The concept of separating a host's identity from its location has been intrinsic to numerous protocol suites, including the *Xerox Network Systems* (XNS), *Internetwork Packet Exchange* (IPX), and *Connectionless Network Service* (CLNS) models. In the IP community, it was first proposed in Saltzer's ruminations on naming and binding, RFC 1498[5], and in Noel Chiappa's NIMROD routing architecture, RFC 1992[6]. In short, a host (or a set of applications running on a host, or a set of sessions it participates in) has an identifier independent of its network topology, and sessions can change network paths by simply changing the topological locations of their endpoints. Mike O'Dell, in Internet Drafts in 1996 and 1997 called *8+8* and *GSE,* suggested an implementation of this scenario using the prefix in the IPv6 address as a locator and the interface identifier as an identifier. One implication of the GSE model is the use of a network prefix translation between an edge network and its upstream provider whatever prefix the edge network uses internally, in the transit backbone, the locator appears to be a PA prefix allocated by the ISP in question. As a result, the routing table, as in shim6, enumerates the ISPs in the network—on the order of 10,000.

The *Identifier-Locator Network Protocol* (ILNP) takes the solution to fruition, operating on that basic model and adding a *Domain Name System* (DNS) Resource Record and a random number nonce to mitigate on-path attacks that result from the fact that the *IPv6 Interface Identifier* (IID) is not globally unique.

As compared to the operational complexities and costs of PI Addressing, Exchange-Based Addressing, and shim6, ILNP has the advantage of being operationally simple. Each LAN has one subnet, when adding or changing providers no edge network renumbering is required, and, as noted, the cost of the global routing table does not increase. Additionally, it is trivial to load-share traffic across points of attachment to multiple ISPs, because the locator is irrelevant above the network layer. And unlike IPv4/IPv4 *Network Address Port Translation* (NAPT), the translation is stateless; as a result, sessions using *IP Security* (IPsec) *Encapsulation Security Protocol* (ESP) encryption can cross it.

In this case, the complexities of the network are transferred to the application itself, and to its transport. The application must, in some sense, know all of its "outside" addresses. It can learn them, of course, by using its domain name in referrals and other uses of the address; in some cases however, the application really wants to know the address itself. If it is communicating those addresses to other applications—the usual usage—the assumption that its view of its address is meaningful to its remote peer is, in the words of RFC 3582[2], *Unilateral Self-Address Fixing* (UNSAF), and the concerns raised in RFC 2993[7] are the result. To mitigate those concerns, ILNP excludes the locator from the TCP and *User Datagram Protocol* (UDP) pseudo-headers (and as a result from the checksum).

The implication of ILNP is, as a result, that TCP and UDP must be either changed or exchanged for other protocols such as *Stream Control Transmission Protocol* (SCTP) or *Multipath TCP* (MPTCP), and that applications must either use DNS names when referring to themselves or other systems in their network—sharply dividing between the application and network layers—or devise a means by which they can determine the full set of their "outside" addresses.

Network Prefix Translation, Also Known as NAT66

Like ILNP, *Network Prefix Translation* (NPTv6) derives from and can be considered a descendant of the GSE model. It differs from ILNP in that it defines no DNS Resource Record, defines no end-to-end nonce, and requires no change to the host, especially its TCP/UDP stacks. To achieve that, the translator updates the TCP/UDP checksum in the source and destination addresses.

If the ISP prefix is a /48 prefix, this prefix allows for load sharing of sessions across translators leading to multiple ISPs; if the ISP prefix is longer, such as a /56 or /60, the checksum update must be done in the IID, and as a result load sharing can be accomplished only across translators between the same two networks. Like ILNP and unlike IPv4/IPv4 NAPT, the translation is stateless; as a result, sessions using IPsec ESP encryption can cross it.

The complexities of the network are again transferred to the application itself, but not to its transport. The application must, in some sense, know all of its "outside" addresses. Using its domain name in referrals and other uses of the address can determine these addresses; in some cases, however, the application really wants to know the address itself. If it is communicating those addresses to other applications—the usual usage—the assumption that its view of its address is meaningful to its remote peer is, again in the words of RFC 3582[2], "UNSAF," and some of the concerns raised in RFC 2993[7] result.

The implication of NPTv6 is that applications must either use DNS names when referring to themselves or other systems in their network—sharply dividing between the application and network layers—or devise a means by which they can determine the full set of their "outside" addresses. However, the IPv6 goal of enabling any system in the network to communicate with any other given administrative support is retained.

Ways Forward

From the perspective of this author, the choice of multihoming technology will in the end be an operational choice. The practice of multihoming is proliferating and will continue to do so. There is a place for provider-independent addressing; it may not in reality make sense for 40,000 companies, but it probably does for the largest edge networks. At the other extreme, shim6-style multihoming makes sense in residential networks with a single LAN; as described earlier, there are simple approaches to making that work through reasonable policy approaches.

For the vast majority of networks in between, policy suggestions that do not substantially benefit the network or users who implement them do not have a good track record. Hence, while Exchange-Based Addressing materially assists in edge network problems, there is no substantive reason to believe that the transit backbone will implement it. Similarly, although shim6 materially helps with the capital and operational expenses of operating the transit backbone, it is not likely that edge networks will implement it.

We also have a poor track record in changing host software. For example, SCTP is in many respects a superior transport protocol to TCP—it allows for multiple streams, it is divorced from network layer addressing, and it allows endpoints to change their addresses midsession.

In a 2009 "Train Wreck" workshop at Stanford University, in which various researchers argued all day in favor of the development of a new transport with requirements much like those of SCTP, the research community acted as if ignorant of it when the protocol was brought up in conversation.

NPTv6 is not a perfect solution, but this author suspects that it will be operationally simple enough to deploy and manage and close enough to the requirements of edge networks and applications that it will, in fact, address the topic of multihoming.

References

[1] Craig Partridge and Frank Kastenholz, "Technical Criteria for Choosing IP The Next Generation (IPng)," RFC 1726, December 1994.

[2] Joe Abley, Benjamin Black, and Vijay Gill, "Goals for IPv6 Site-Multihoming Architectures," RFC 3582, August 2003.

[3] Richard Draves, "Default Address Selection for Internet Protocol version 6 (IPv6)," RFC 3484, February 2003.

[4] Arifumi Matsumoto, Jun-ya Kato, and Tomohiro Fujisaki, "Update to RFC 3484 Default Address Selection for IPv6," Internet Draft, Work in Progress, March 2011, **http://tools.ietf.org/html/draft-ietf-6man-rfc3484-revise**

[5] Jerome Saltzer, "On the Naming and Binding of Network Destinations," RFC 1498, August 1993.

[6] Isidro Castineyra, Noel Chiappa, and Martha Steenstrup, "The Nimrod Routing Architecture," RFC 1992, August 1996.

[7] Tony Hain, "Architectural Implications of NAT," RFC 2993, November 2000.

[8] Leslie Daigle, Ed., IAB "IAB Considerations for UNilateral Self-Address Fixing (UNSAF) Across Network Address Translation," RFC 3424, November 2002.

[9] Erik Nordmark and Marcelo Bagnulo, "Shim6: Level 3 Multihoming Shim Protocol for IPv6," RFC 5533, June 2009.

[10] Ole Troan, David Miles, Satoru Matsushima, Tadahisa Okimoto, and Dan Wing, "IPv6 Multihoming without Network Address Translation," Internet Draft, Work in Progress, **http://tools.ietf.org/html/draft-ietf-v6ops-ipv6-multihoming-without-ipv6nat**

[11] Margaret Wasserman and Fred Baker, "IPv6-to-IPv6 Network Prefix Translation,"
Internet Draft, Work in Progress, http://tools.ietf.org/html/draft-mrw-nat66

[12] Ran Atkinson and Scott Rose, "DNS Resource Records for ILNP," Internet Draft, Work in Progress,
http://tools.ietf.org/html/draft-rja-ilnp-dns

[13] Ran Atkinson, "ICMP Locator Update message," Internet Draft, Work in Progress,
http://tools.ietf.org/html/draft-rja-ilnp-icmp

[14] Ran Atkinson, "ILNP Concept of Operations," Internet Draft, Work in Progress,
http://tools.ietf.org/html/draft-rja-ilnp-intro

[15] Ran Atkinson, "ILNP Nonce Destination Option," Internet Draft, Work in Progress,
http://tools.ietf.org/html/draft-rja-ilnp-nonce

[16] Alan Ford, Costin Raiciu, Mark Handley, and Olivier Bonaventure, "TCP Extensions for Multipath Operation with Multiple Addresses," Internet Draft, Work in Progress,
http://tools.ietf.org/html/draft-ietf-mptcp-multiaddressed

[17] Randall Stewart, Ed., "Stream Control Transmission Protocol," RFC 4960, September 2007.

[18] Randall Stewart, Qiaobing Xie, Michael Tuexen, Shin Maruyama, and Masahiro Kozuka, "Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration," RFC 5061, September 2007.

[19] Jon Postel, "User Datagram Protocol," RFC 768, August 1980.

[20] Jon Postel, "Transmission Control Protocol," RFC 793, September 1981.

[21] Fred Baker and Pekka Savola, "Ingress Filtering for Multihomed Networks," RFC 3704 [BCP 84], March 2004.

[22] David Meyer, "The Locator Identifier Separation Protocol (LISP)," *The Internet Protocol Journal*, Volume 11, No, 1, March 2008.

FRED BAKER, a Cisco Fellow, has been active in technology development and Internet standardization since the 1980s. He participated in early development of IEEE 802.1d switching and IP routing. In the IETF, he has written or edited RFCs on a variety of topics, and chaired both working groups and the IETF itself. At this time, he is the IETF's Voting Member on the U.S. NIST Smart Grid Interoperability Panel, a member of the SGIP's Architecture Committee, and co-chair of the IETF IPv6 Operations Working Group. At Cisco, his group supports research at universities; he is looked to for research advice and mentorship both within and outside the company. E-mail: **fred@cisco.com**

Reflecting on World IPv6 Day

by Phil Roberts, ISOC

On June 8, 2011, many websites around the world made their main webpage reachable over IPv6 for 24 hours, and many of those that did this left their sites IPv6-accessible afterward.

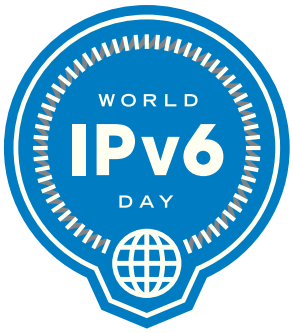
Major worldwide websites enabled IPv6 on their main page. Google enabled not only its main website but also YouTube and Blogger. Facebook and Yahoo! both enabled their main webpages as well. These websites are the five most visited websites in the world according to Alexa rankings. Other major worldwide websites that enabled IPv6 include Yahoo! Japan, Bing, Microsoft, BBC, CNN, and AOL.

Important local websites in countries around the world also joined in. In South Korea both Naver and Daum (the first and fourth most visited sites in South Korea according to Alexa) joined the event. In the Czech Republic four of the top 25 local websites joined. There were also major sites from Brazil, Portugal, and Indonesia.

Purposes

Enabling IPv6 in this way served numerous purposes:

- Network operators clearly saw that content is going to be available on IPv6. Although the major websites may not be quite there yet, it is clear that they are seriously moving in that direction.
- The industry worked to improve problems with IPv6 connectivity. Some immediate improvement resulted, and more fixes are underway to further improve IPv6 connectivity.
- Setting a public date created a deadline that accelerated deployment for many of the organizations that contacted us.
- It was important to be compared with Google, Facebook, and Yahoo!. Participants in this experiment wanted to be seen doing the same thing as the industry giants.
- This event was a clear example of how the Internet industry can work together to deploy technology that is for the good of the Internet, without intervention from outside entities. The multi-stakeholder model of Internet development continues to function well.



More than 1000 organizations contacted the Internet Society. Many of these organizations had already permanently enabled IPv6. Of the 430 or so websites the Internet Society monitored on the day, roughly two-thirds have continued to provide IPv6 access after the day.

In addition, major hosting companies enabled IPv6 for large numbers of domains, including Domain Factory, which, as a result of participating in World IPv6 Day, has made IPv6 "on by default" for all of its more than 800,000 domains. Another hosting company, Stratos, left IPv6 on after June 8 for its more than 4 million domains.

RIPE Labs did extensive measurements of IPv6 leading up to, on, and after the day, and it has published results indicating an increase in IPv6 traffic on the day—and an overall increase in IPv6 traffic also after the day.

References

[1] Phil Roberts, "World IPv6 Day," *The Internet Protocol Journal,* Volume 14, No. 1, March 2011.

[2] RIPE Labs, "Measuring World IPv6 Day—Long-Term Effects," **http://labs.ripe.net/Members/emileaben/measuring-world-ipv6-day-long-term-effects**

[3] RIPE Labs, "Measuring World IPv6 Day—Some Glitches And Lessons Learned," **http://labs.ripe.net/Members/emileaben/measuring-world-ipv6-day-glitches-and-lessons-learned**

[4] RIPE Labs, "Measuring World IPv6 Day—First Impressions," **http://labs.ripe.net/Members/mirjam/measuring-world-ipv6-day-first-impressions**

PHIL ROBERTS joined the Internet Society (ISOC) in 2008. Prior to that he spent several years with Motorola in research and product development, all in the area of mobile broadband systems. He has been active in the IETF for more than a decade. He can be reached at: **roberts@isoc.org**

Letters to the Editor

Hi Geoff,

Thanks you for your contribution to the March 2011 issue of *The Internet Protocol Journal*. Your description in "A Rough Guide to Address Exhaustion" and the article on "Transitional Myths" were very insightful into the whole issue of IPv4 to IPv6, and the issues concerning migration. Some of your thoughts on the migration hit home, as I am speaking to customers about the planning for the transition and I see a lot of "Got You" that I must now incorporate in my discussions with my customer.

If you do have a means of updating the technical community with activities in the area of IPv6 and how to move customers to this protocol platform, can you please point me in that direction? I like your approach and so would like to stay close to what you are doing in this area. Again, thank you for your contribution!

Ole, thanks for getting this type of information out to the technical community. Great work.

—Joel Smith, Verizon Business, Toronto, Ontario, Canada
joel.smith@one.verison.com

The author responds:

Hi Joel,

Thank you for your comments.

Running IPv6 in a dual-stack configuration certainly presents some issues, some of which are unique to particular networks and configurations, some of which appear to be common to particular roles (such as content delivery platform, Internet Service Provider, Enterprise Provider, and end user), and some of which are common across most, if not all, circumstances.

In assisting to set up some dual-stack services a year ago, I wrote down some of the issues that I found helpful in an article: "Two Simple Hints for Dual Stack Servers" (**http://www.potaroo.net/ispcol/2010-05/v6hints.html**). You may find those hints to be of some value to your work. Some other sites that have a good collection of information are: **http://www.ipv6actnow.org/** and the community site **http://www.getipv6.info/index.php/Main_Page**, which also contains a wealth of information of a technical nature.

The basic guideline is to approach adding IPv6 to a network like any other engineering project: exercise care and attention to detail, and you will find it to be very straightforward!

Kind regards,

—Geoff Huston, APNIC
gih@apnic.net

Geoff and Ole,

Many thanks for your excellent papers in the March 2011 issue of IPJ. You have brought all the issues together in one place. They are clearly explained. Now I'll do my small part by suggesting to one and all that they read it. My IPv6 service comes from a manually configured tunnel from Hurricane Electric.

—Dan Cotts
dcotts@lisco.com

The author responds:

Thanks, Dan, for this feedback. It's certainly the right time for both users and content providers to act now to ensure that we continue to enjoy an Internet that still operates with a coherent end-to-end architecture into the future. The only way we can ensure that this happens is to act now and insist on IPv6—everywhere!

—Geoff Huston, APNIC
gih@apnic.net

Hello,

I enjoyed the recent IPv6 issue (Volume 14, No. 1, March 2011), but was dismayed by the lack of any frank discussion of the IPv6 "any-to-any" mantra versus the benefits of IPv4 *Network Address Translation* (NAT).

Internet purists don't hide their desire to rid the world of NAT and return to an any-to-any Internet where they could use FTP to/from any host. But for the past 15 years, NAT, RFC 1918, and perimeter security have been great for the Internet and for home and enterprise networking. When dealing with billions of endpoints, the implicit security of NAT far outweighs any alternative. Just think back to the pre-broadband/NAT days when hosts were attacked within seconds of dialing into an ISP.

Of the ~1.7 billion publicly addressed Internet devices, the vast majority would be perfectly happy behind *Carrier-Grade NAT* (CGN). In fact, as ISPs begin introducing NAT offerings, millions will stampede to them for their lower cost. Mobile phone networks are the lowest-hanging fruit, followed by residential broadband. ISPs will still offer public IP products, of course, just at a higher price point.

The IETF needs to stop pussy-footing around the issue. CGN is not just an IPv6 transitional technology; it could very well become the de facto operating standard for the next decade.

The IETF desperately needs to:

- Amend RFC 5382 ("NAT Behavioral Requirements for TCP") to allow endpoint-independent mapping. This will improve CGN scalability by several orders of magnitude. For example, rather than 2000 hosts per public IP mentioned in Mr. Huston's "Rough Guide" on address sharing, CGN could support 200,000 or more hosts per public IP.
- Develop an IETF standard for P2P connection establishment. It took 8+ years for the IETF to take an interest in P2P mechanics (RFC 5128). Now it's time to show leadership. If a CGN-compatible P2P establishment standard were drafted, it would be adopted by P2P libraries overnight. While they're at it, look at standards for tying *Universal Plug and Play* (uPnP) into CGN.
- Help coordinate a discussion of operational issues with ISP administration, law enforcement, DMCA enforcement, geolocation services, black/white lists, etc. Perhaps it's time to extol the benefits of millisecond-accurate IPFIX logs with NAT extensions, or develop a new TCP option to embed NAT details?
- Legitimize common ISP self-preservation tactics, such as restricting SMTP, metering connections/sec, and so on.

Most importantly, IPv6 proponents should stop taking CGN as a personal affront. There is no malice; it's simply the path of least resistance for the IPv4 conundrum.

—*Craig Weinhold, Madison, Wisconsin*
craig.weinhold@cdw.com

The author responds:

Thank you for your note, Craig.

The discussion of how far the Internet could scale with integration of NATs into the interior of the network as well as the current pattern of NATs at the edge is not a new discussion. The *Realm Specific IP* (RSIP) Working Group was active over a decade ago in the IETF, looking at how a network would operate that consisted of a union of distinct realms, each of which was, in address terms, a discretely addressed IP network. With the benefit of hindsight, the outcomes of that effort in supporting a case for infrastructure NATs as a long-term architectural direction for the Internet were not overly encouraging.

From the perspective of the technology community, it reinforced the conclusion that IPv6 represented the best possible response to the recognized problem of IPv4 address exhaustion. NATs were a poor compromise in so far as, at the most basic level, NATs add state into the interior of the network. This imposition of state into the network infrastructure imposes a cost in terms of service fragility and network robustness that cannot be avoided.

There was an assumption some years ago that the industry would grapple with the transition to IPv6 well before the exhaustion of IPv4 addresses, and we would never have to deal with a dual-stack transition where one-half of the dual stack, the IPv4 part, would need to operate in a mode that included infrastructure NATs. We now appear to be beyond choice here—for the Internet to continue to grow by a further 300 million new services per year at present, and grow by yet more in the coming years, there is no choice but to operate the IPv4 part of the dual-stack environment with infrastructure NATs.

But this is a short-term hack, as distinct from a tenable longer-term position. The address pool of IPv4 is not getting any larger, and as more and more new services are added into a dual-stack network, the growth in the IPv4 part of the network can be absorbed only by progressive reduction of the number of available ports to each client of the infrastructure NAT. Services become more fragile and the network becomes less resilient. The inevitable next step in progressive scarcity of IPv4 addresses in the face of such inexorable growth is to drop the entire notion of end-to-end service and introduce application-level proxies into the IPv4 network. At this point we lose any ability to further sustain an open IPv4 Internet. The only applications that could be supported are those that are supported by the application-level proxies, and all other applications simply fail. The segregation of one Internet into a number of effectively disconnected "walled gardens" of networking is a rapid outcome in such a scenario.

One of the strengths of the Internet is its openness and neutrality. The open architectural model allows novel services to be added into the network by simply equipping clients and services with the service, leaving the interior of the network untouched. The interior of the network is entirely neutral to such innovations, as it is unaware of the content or intent of the packets that are passed through its switching infrastructure.

So the long-term path of greatest common benefit to all in the Internet is a network that, as far as possible, simply vanishes! It is an Internet where content and services can rendezvous with users without having to negotiate with any network elements. It is a network that is free of toll gates. And the network has now grown to such an extent that the only path from here that can sustain that architectural simplicity and sustain yet more growth is one that shifts determinedly and rapidly to IPv6. With the limited time and resources available, attempting to improve upon NATs is, in my opinion, not the best use of the resources we can apply to this problem.

Regards,

—Geoff Huston, APNIC
gih@apnic.net

Call for Papers

The Internet Protocol Journal (IPJ) is published quarterly by Cisco Systems. The journal is not intended to promote any specific products or services, but rather is intended to serve as an informational and educational resource for engineering professionals involved in the design, development, and operation of public and private internets and intranets. The journal carries tutorial articles ("What is...?"), as well as implementation/operation articles ("How to..."). It provides readers with technology and standardization updates for all levels of the protocol stack and serves as a forum for discussion of all aspects of internetworking.

Topics include, but are not limited to:

- Access and infrastructure technologies such as: ISDN, Gigabit Ethernet, SONET, ATM, xDSL, cable, fiber optics, satellite, wireless, and dial systems
- Transport and interconnection functions such as: switching, routing, tunneling, protocol transition, multicast, and performance
- Network management, administration, and security issues, including: authentication, privacy, encryption, monitoring, firewalls, troubleshooting, and mapping
- Value-added systems and services such as: Virtual Private Networks, resource location, caching, client/server systems, distributed systems, network computing, and Quality of Service
- Application and end-user issues such as: e-mail, Web authoring, server technologies and systems, electronic commerce, and application management
- Legal, policy, and regulatory topics such as: copyright, content control, content liability, settlement charges, "modem tax," and trademark disputes in the context of internetworking

In addition to feature-length articles, IPJ contains standardization updates, overviews of leading and bleeding-edge technologies, book reviews, announcements, opinion columns, and letters to the Editor.

Cisco will pay a stipend of US$1000 for published, feature-length articles. Author guidelines are available from Ole Jacobsen, the Editor and Publisher of IPJ, reachable via e-mail at `ole@cisco.com`

Fragments

RFC Series Editor Search Announcement

The *Internet Engineering Task Force* (IETF) is seeking an *RFC Series Editor* (RSE). The RSE has overall responsibility for the quality, continuity, and evolution of the *Request for Comments* (RFC)[3] Series, the Internet's seminal technical standards and publications series. The position has operational and policy development responsibilities. The overall leadership and supervision of RFC Editor function is the responsibility of the RFC Series Editor. The RSE is a senior professional who must be skilled in leading, managing and enhancing a critical, multi-vendor, global information service. The following qualifications are desired:

- Leadership and management experience. In particular, demonstrated experience in strategic planning and the management of entire operations. Experience that can be applied to fulfill the tasks and responsibilities described in "RFC Editor Model (version 2)"[1].
- Excellent written and verbal communication skills in English and technical terminology related to the Internet a must; additional languages a plus.
- Experience with editorial processes.
- Familiar with a wide range of Internet technologies.
- An ability to develop a solid understanding of the IETF, its culture and RFC process.
- Ability to work independently, via e-mail and teleconf, with strong time management skills.
- Willingness and ability to travel as required.
- Capable of effectively functioning in a multi-actor and matrixed environment with divided authority and responsibility; ability to work with clarity and flexibility with different constituencies.
- Experience as an RFC author desired.

More information about the position can be found on the RFC Editor Webpage[2]. The RSE reports to the *RFC Series Oversight Committee* (RSOC). Expressions of interest in the position, Curriculum Vitae (including employment history), compensation requirements, and references should be sent to the RSOC search committee at **`rse-search@iab.org`**. Questions are to be addressed to the same e-mail address. Applications will be kept confidential. The RSOC will interview interested parties at the IETF meeting in Quebec City that begins July 24, 2011, but the application period is open until the position is filled.

—Fred Baker, Chair, RFC Series Oversight Committee

References

[1] **http://www.ietf.org/id/draft-iab-rfc-editor-model-v2-02.txt**

[2] **http://www.rfc-editor.org/rse/RSE-position.html**

[3] Leslie Daigle, "RFC Editor in Transition: Past, Present, and Future," *The Internet Protocol Journal,* Volume 13, No. 1, March 2010.

Global IPv6 Deployment Monitoring Survey 2011

The *Global IPv6 Deployment Monitoring Survey 2011* is now online at: **http://www.surveymonkey.com/s/GlobalIPv6survey2011**

This survey has been designed by GNKS Consult in collaboration with TNO and the RIPE NCC to further understand where the community stands on IPv6 and what needs be done to ensure that the Internet community is ready for the widespread adoption of IPv6.

Anyone can participate in this survey and we hope that the results will establish a comprehensive view of current IPv6 penetration and future plans for IPv6 deployment. The survey comprises 23 questions and can be completed in about 15 minutes. For those without IPv6 allocations or assignments or who have not yet deployed IPv6, there will be fewer questions.

The survey closes July 31, 2011. We thank you for your time and interest in completing this survey. If you have any questions concerning the survey, please e-mail: **info@gnksconsult.com**

For more information about the survey and links to previous year's survey results, please see:

https://www.ripe.net/internet-coordination/news/industry-developments/global-ipv6-deployment-monitoring-survey-2011

RFC 6127 Published

The topic of IPv4 depletion and IPv6 deployment is covered in the recently published RFC 6127 entitled "IPv4 Run-Out and IPv4-IPv6 Co-Existence." From the introduction: "When IPv6 was designed, it was expected that the transition from IPv4 to IPv6 would occur more smoothly and expeditiously than experience has revealed. The growth of the IPv4 Internet and predicted depletion of the free pool of IPv4 address blocks on a foreseeable horizon has highlighted an urgent need to revisit IPv6 deployment models. This document provides an overview of deployment scenarios with the goal of helping to understand what types of additional tools the industry needs to assist in IPv4 and IPv6 co-existence and transition." RFCs can be obtained from the RFC Editor web page, see:

http://www.rfc-editor.org/rfc.html

The Internet Protocol Journal, Cisco Systems
170 West Tasman Drive
San Jose, CA 95134-1706
USA

ADDRESS SERVICE REQUESTED

PRSRT STD
U.S. Postage
PAID
PERMIT No. 5187
SAN JOSE, CA

The Internet Protocol Journal is published quarterly by the Chief Technology Office, Cisco Systems, Inc.
www.cisco.com
Tel: +1 408 526-4000
E-mail: ipj@cisco.com