John W. Stewart, III / Juniper
Enke Chen / Cisco
March 1998

Route Aggregation Tutorial
<draft-ietf-idr-aggregation-tutorial-01.txt>

Status of this Memo

Abstract

Route aggregation is critical to the long-term viability of the Internet. This document presents practical information that network managers can use to both understand the concepts of aggregation as well as put those concepts into use in order to do their part to make the Internet stable and allow its continued growth. The intended audience for this document is anyone responsible for configuring a network which has its own Autonomous System Number (ASN) and exchanges routing information with its Internet Service Provider(s) (ISP(s)) using the Border Gateway Protocol (BGP). This document does not cover multi-homing, though multi-homed sites can still benefit from understanding this material.

1. Introduction

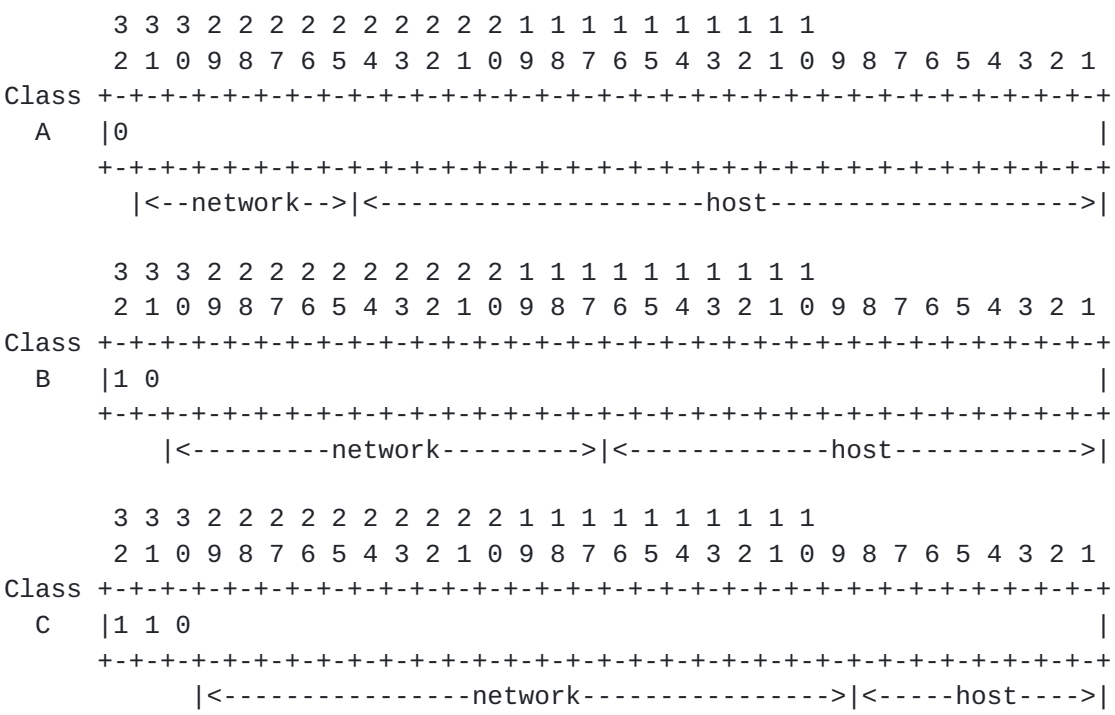
The long-term viability of the Internet depends on its ability to support the continued growth in demand. A large part of its ability to grow is dependent on the successful scaling of the routing system. Because the complexity of the Internet's routing system is a function of the number of reachable destinations, great care must be taken that, as the network grows, the demands on the routing system don't outpace advances in hardware and software.

In the early 1990s, the paradigm for large scale Internet routing changed from a "Classful" system to a "Classless" system. The Classless system applies techniques of hierarchy to achieve large scaling. In order for Classless routing to achieve its goal of allowing the routing system to scale very well, networks in all areas of the Internet must be vigilant about "route aggregation." This document provides educational information, both conceptual and practical, in an effort to encourage efficient aggregation throughout the Internet.

This document assumes only a very casual understanding of Internet addresses. Once readers clearly understand this document, they may wish to read "A Framework for Inter-Domain Route Aggregation" [9] to understand the big picture of large-scale aggregation in the Internet.

2. Network Classes and the Bit-Level Detail

As originally specified in the early 1980s, the Internet Protocol (IP) included the idea of network "Classes." [1] In IP, a certain number of bits in the 32-bit addresses refer to the network and the remainder of the bits refer to a host on that network. (In the mid 1980s IP was extended such that part of the host bits can refer to a subnet and the remainder would refer to a host on that subnet. [2]) The point of the different Classes was to have addresses with different numbers of network/host bits. The Class of an address could be determined by the high-order bits. A Class A address had "0" as the high-order bit, and then 7 bits of network and 24 bits of host; a Class B address had "10" as the high-order two bits, and then 14 bits of network and 16 bits of host; and a Class C address had "110" as the high-order three bits and then 21 bits of network and 8 bits of host. Looking at an address in "dotted quad notation" (e.g., 166.45.3.46), Class A networks have a first number of 0-127, Class B networks have a first number of 128-191 and Class C networks have a first number of 192-223. A Class A network could number 1.7 million hosts, a Class B 65,000 and a Class C 256. Diagramatically:



Although the original intent of having Classes was to allow for flexible addressing, experience showed that the hard boundary of the three Classes actually made the addressing less flexible. For example, if a site connecting to the Internet needed to address 300 hosts, then a Class C network wouldn't be adequate and a Class B would need to be assigned. This resulted in poor utilization of the assigned address space and caused a faster-than-necessary rate of consumption of the available IP address space.

Another problem with the scalability of Internet routing under the Classful system had to do with the address allocation policies used. At that time, when a site connected to the Internet, it would go to a central registry to get a unique IP network and then it would go to an ISP to procure connectivity. What this means is that if an ISP had 1000 customers, each of whom had been assigned a Classful network of some type, then that ISP would have to announce each of those 1000 networks to other providers in the Internet. In other words, the Internet's routing system was not taking advantage of the inherent provider/subscriber hierarchy and instead was being "flat-routed."

3. The Introduction of CIDR

In the early 1990s, a number of ISPs began to have operational problems related to the size of a full Internet routing table because of the limited amount of memory available in commercial routers. (A "full routing table" means a routing table which does not contain a default route and

instead contains an entry for every active network in the Internet.) Because of these problems, Classless Inter-Domain Routing (CIDR) was created. [3]

CIDR removed the idea of Classes from IP. Instead of having networks with an implied number of bits referring to network/host, there are "prefixes" with an associated mask explicitly identifying which bits refer to network/host. For example, the prefix "38.245.76.0" with a mask of "255.255.255.0" has 24 bits of network and 8 bits of host (i.e., it can address the same number of hosts as a Class C network even though the prefix is in the Class A range). The CIDR paradigm prefers the term "prefix" over "network" because it's more clear that no Class is being implied. Another way to write this example prefix is "38.245.76.0/24", meaning that the mask contains 24 1s in the high-order portion of the mask.

The strength of CIDR is that the masks can be on arbitrary bit boundaries and don't have to be on byte boundaries. So for example, going back to the case of the site which needs to address 300 hosts, the site could be allocated a "/23" (i.e., a prefix which has 23 bits for network and 9 bits for host, thus allowing 512 hosts to be addressed with the single prefix).

To complete the picture, in order for CIDR to actually help achieve better scaling of Internet routing, a specific address allocation architecture must be used. [4] Rather than the pre-CIDR style where sites would go to a centralized registry to get an address which does not take into account where that site connects to the Internet, CIDR-style address allocation involves registries allocating address space to ISPs who, in turn, sub-allocate it to their customers. So for example, a registry might allocate the prefix 204.71.0.0/16 (called a "CIDR block") to ISP1, and then ISP1 could sub-allocate 204.71.1.0/24 to SmallCustomer1, 204.71.2.0/24 to SmallCustomer2, 204.71.128.0/22 to MediumCustomer and 204.71.136.0/20 to LargeCustomer. The benefit, then, is that when ISP1 exchanges routing information with other ISPs, it only needs to announce the single prefix 204.71.0.0/16 and not each of the individual prefixes used by its customers. The ability to merge multiple prefixes which have some number of leading bits in common is called "aggregation."

In 1993, the deployment of a routing protocol which supported CIDR (specifically BGP Version 4 [5]) had an immediate and measurably positive effect on route scaling. Immediately after its deployment a full routing table went down in size in absolute numbers (this was possible only because address allocation had already been done for some time in the CIDR style even though the routing hadn't yet taken advantage of it) and, more importantly, the rate of growth was slowed.

4. A Note on Renumbering

The crux of CIDR is that the Internet's generally hierarchical topology is being reflected in the addressing. As a result, if a site started out as a customer of ISP1 and is thus numbered out of one of ISP1's CIDR blocks, but then that site terminates the relationship with ISP1 and "rehomes" to ISP2, then the site would need to renumber its nodes to be part of one of ISP2's CIDR blocks. The major reason for this is to retain efficiency in the routing system. [6]

Renumbering is an unfortunate necessity in the current IPv4 Internet. This is the reason for the recent advance of renumbering technology in IPv4 (e.g., DHCP [7]) as well as the focus of easy renumbering in IPv6. [8] Sites should keep this "unfortunate necessity" in mind when deploying equipment to make sure that their infrastructure can be renumbered easily if that becomes necessary.

5. Practical Aggregation

As stated earlier, aggregation refers to the combining of multiple contiguous prefixes into a single prefix. For example, assume the prefixes 209.123.10.0/24 and 209.123.11.0/24. The binary representation for 209.123.10.0/24 is:

```
 3 3 3 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1
 2 1 0 9 8 7 6 5 4 3 2 1 0 9 8 7 6 5 4 3 2 1 0 9 8 7 6 5 4 3 2 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1 1 0 1 0 0 0 1 0 1 1 1 1 0 1 1 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|<---- 209 ---->|<---- 123 ---->|<----  10 ---->|<----  0  ---->|
```

And the binary representation for 209.123.9.0/24 is

```
 3 3 3 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1
 2 1 0 9 8 7 6 5 4 3 2 1 0 9 8 7 6 5 4 3 2 1 0 9 8 7 6 5 4 3 2 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|1 1 0 1 0 0 0 1 0 1 1 1 1 0 1 1 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0|
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|<---- 209 ---->|<---- 123 ---->|<----  11 ---->|<----  0  ---->|
```

The important thing to note here is that the two networks can be aggregated into the single prefix 209.123.10.0/23 because they have the leading 23 bits in common.

The example above is very simple. A real example of a very large degree of aggregation is the prefix 208.0.0.0/12, which covers the 4096 24-bit prefixes 208.0.0.0/24 through 208.15.255.0/24. It's obvious from this example how profound an impact aggregation can have on the size of a routing table and the resources required for the associated storage and computation.

It is important to aggregate as much as possible, even in the simple example presented earlier, because small non-optimalities can add up and result in a poorly aggregated global routing system. If you exchange routes with your provider using BGP, then it is your responsibility to do the aggregation configuration. (Note that aggregation can only be done with BGP4, so if you are running an earlier version of BGP, you should upgrade your software; most major router manufacturers have implemented BGP4.) Assuming that your AS number is 5555, your provider's AS number is 2222 and the IP address of your provider's BGP speaker is 1.2.3.4, the Cisco syntax for configuring the aggregation would be:

```
interface Ethernet0
 ...
 ip address 209.123.10.1 255.255.255.0
...
interface Ethernet1
 ...
 ip address 209.123.11.1 255.255.255.0
...
router bgp 5555
 network 209.123.10.0 mask 255.255.254.0
 neighbor 1.2.3.4 remote-as 2222
...
ip route 209.123.10.0 255.255.254.0 Null0 254
```

The "network" line in the BGP section tells the router to announce that network if it has a route to it. The "ip route" statement for 209.123.10.0/23 is a static route that creates a "pull-up" route for the aggregate; this gives the router a route to the prefix so that the "network" line takes effect and the prefix is announced. The static route for the aggregate is only needed in order for the "network" line to take effect; that static route will never be used for packet forwarding because the static routes for the individual /24 prefixes are more specific and therefore take precedence. This configuration information is required only on the router which speaks BGP with your provider's router.

In this example, it is assumed that the router which speaks BGP to the

provider has local interfaces numbered out of the address space being aggregated. This is assumed for simplicity of the example; the "network" line and pull-up route would be used the same ways to do the aggregation even if the routing for the address space were done statically, based on an IGP, etc.

6. References

[1] Postel, J., "Internet Protocol", RFC 791, September 1991.

[2] Postel, J., Mogul, J.C., "Internet Standard Subnetting Procedure", RFC 950, August 1985.

[3] Fuller, V., Li, T., Yu, J., and K. Varadhan, "Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy", RFC 1519, September 1993.

[4] Rekhter, Y., Li, T., "An Architecture for IP Address Allocation with CIDR", RFC 1518, September 1993.

[5] Rekhter, Y., and Li, T., "A Border Gateway Protocol 4 (BGP-4)", RFC1771, March 1995.

[6] Ferguson, P., Berkowitz, H., "Network Renumbering Overview: Why would I want it and what is it anyway?", RFC 2071, January 1997.

[7] Droms, R., "Dynamic Host Configuration Protocol", RFC 2131, March 1997.

[8] Thomson, S., Narten, T., "IPv6 Stateless Address Autoconfiguration", RFC 1971, August 1996.

[9] Chen, E., Stewart III, John W., "A Framework for Inter-Domain Route Aggregation", draft-ietf-idr-aggregation-framework-01.txt, July 1997.
TBD -- RFC NUMBER

7. Authors' Addresses

John W. Stewart, III
Juniper Networks, Inc.
385 Ravendale Drive
Mountain View, CA 94043
phone: +1 650 526 8000
email: jstewart@juniper.net

Enke Chen
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134-1706
Phone: +1 408 527 4652
email: enkechen@cisco.com