Inter domain Routing: (addenda to routing 101)

The true value of the Internet is in its connectedness. Inter-domain routing protocols are the glue which make this to happen while allowing policy flexibility.

AS concept:
==========

Autonomous system (AS) = set of networks and routers which are internally connected and under a single administrative control. The notion of an AS is critical to EGPs. Note that AS refers to a set of routers, not IP prefixes. Since admin borders are crossed, there is not necessarily an optimization goal on routing, but it should be able to support policy constraints.

Key roles of an AS:

1. Routing/connectivity :

- ***Internally connected,*** i.e. should not use path through another AS for internal connectivity
    - ***Designates at least 1 router that participates in AS-level routing***
      AS is assigned a ***16-bit AS number*** for AS-level routing
    - Due to paucity of AS numbers, a roughly hierarchical AS structure is expected, where customer/stub AS connected to a single ISP ***shares the AS number*** of its ISP. This stub AS is not directly visible to other ISPs.
    - The "border router"(s) must ***advertise the reachability*** of networks within the AS to the external world/core…
        - Note: "***reachability***" is different from "path" or "distance" or "link state" => no quantitative metric specified => suboptimal…
        - Note ***advertising networks*** is different ***advertising AS-numbers*** as part of a AS-level routing. But an AS-number is used always with respect to a set of networks (designated by prefixes)

2. Autonomy/administrative :

Many networks and routers, under ***single administrative control.***

This autonomy allows the AS administrator flexibility to

- ***a) "Hide" certain networks*** from the external world. This typically leads to private intranets using IP which may use a block of private IP addresses.
- b) Limit transit traffic destined to other ASs (AS-level ***route-filtering***). Eg: enterprise AS may not want to act as transit.
- c) Specify that its traffic not use a particular AS (typically a neighbor) or other such policies for routing: ***policy-based routing.***

3. Addressing/aggregation:

AS is also expected ***aggregate addresses*** as much as possible and deploy address ***renumbering*** systems (DHCP/NAT) to allow provider-based address allocation, and the core routing systems to scale. These issues are examined in more detail below:

==========

Route/Address aggregation:

Central issue: Route/address aggregation is good as it:

- reduces the size, and slows the growth, of the Internet routing table. Thus, the amount of resources (e.g., CPU and memory) required to process routing information is reduced and route calculation is sped up.
- Another benefit of route aggregation is that route flaps are limited in number, frequency and scope, which saves resources and makes the global Internet routing system more stable.

- The AS boundary is a natural place to implement address/route aggregation, i.e., the set of networks could be represented by a single aggregated (more-specific) prefix.

Problem: Classful addressing.

Solution: ***CIDR***. Similar to subnetting (actually called ***supernetting"),*** but only no equivalent of subnet masks. See the CIDR bullet in routing 101 for more info on CIDR.

- A contiguous set of addresses is assigned to an AS. ***Contiguous => summarizable*** through an address prefix.

Such prefixes are represented by the ***notation: 192.16.0.0/16***

- Notice that the classful addressing imposed an ***implicit mask*** (8-, 16- , or 24-bit), whereas CIDR imposes an ***explicit mask*** of the form /16, /21 etc

- The address allocation system becomes decentralized: IANA -> tier 1 ISPs -> tier 2 ISPs -> tier 3 ISPs -> customers etc ***[Provider-based address allocation]***

- Because of historical reasons, i.e. networks ***already had classful address assignments and have not deployed address renumbering systems (eg: NAT or DHCP)*** to change over to the new address allocation scheme.

- This implies that a less specific prefix may be assigned to a new customer even though a chunk of that address space is occupied by a different network/AS.
- Therefore, when a packet comes to a core/border router, its destination address may match multiple prefixes in the routing table.
- ***Rule: choose longest-prefix match, i.e., more specific prefix.***
- This constraint also influences the ***design of forwarding tables, and super-fast route-lookup algorithms.***
- Since AS may have non-aggregatable prefixes, these need to be advertised outside the AS.

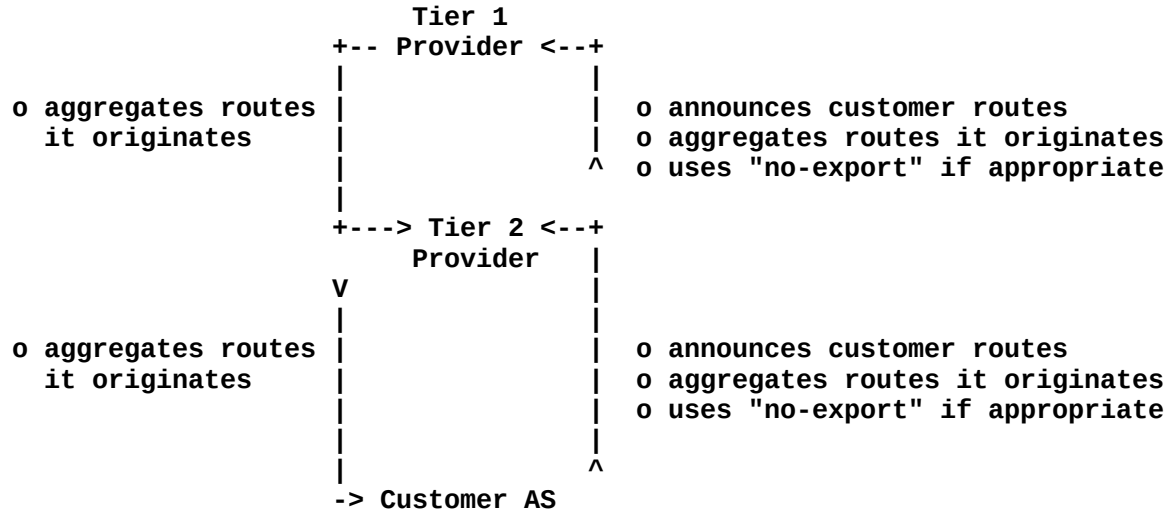
- ***Framework*** of aggregation: "***aggregation by the source AS***", both within routing domains as well as towards upstream providers.

- Route aggregation is done in a ***distributed fashion***, with emphasis on aggregation by the party or parties injecting the aggregatable routing information into the global mesh.
- The ***flexibility*** of a routing domain to be able to inject more granular routing information to an adjacent domain to control the resulting traffic patterns, without having an impact on the global routing system.
  - Such flexibility/granular information can be used for load-balancing, multi-homing etc.
  - "Routes originated by an AS" refers to routes which have that AS first in the AS path attribute; including statically configured routes.

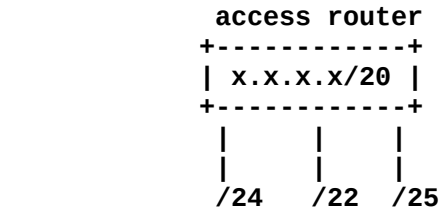
- ***Implementation of aggregation:*** use 'no-export' BGP community to balance the provider-subscriber need for more granular routing information with the Internet's need for scalable inter-domain routing.

- That is, in its route announcements toward its upstream provider, an AS tags the BGP community "no-export" to routes it originates that ***do not need to be propagated beyond its upstream provider*** (e.g., prefixes allocated by the upstream provider).

Figure 1 : aggregation implementation

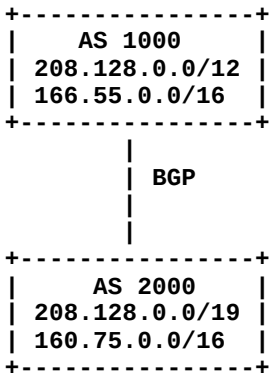


- The implementation of hierarchical address allocation is done primarily through static configuration. Every BGP router can be configured with an address block. Note that /24 and /22 blocks are more specifics of /20 (see diagram below)



Example of Route/Address Aggregation: (From RFC 2519)

> Consider the example shown in Figure 3 where AS 1000 is a "Tier 1" provider with two large aggregates 208.128.0.0/12 and 166.55.0.0/16, and AS 2000 is a customer of AS 1000 with a "portable address" 160.75.0.0/16 and an address 208.128.0.0/19 allocated from AS 1000. Assume that 208.128.0.0/19 does not need to be propagated beyond AS 1000.



Then, based on the framework, AS 1000 would

- originate and advertise the BGP routes 208.128.0.0/12 and 166.55.0.0/16, and suppress more-specifics originated by itself/private-ASs/dedicated-ASs

- advertise the routes received from the customer AS 2000

and AS 2000 would

- originate BGP route 208.128.0.0/19 and 160.75.0.0/16

- advertise both 160.75.0.0/16 and 208.128.0.0/19 to its provider AS 1000 and suppress the more specifics originated by itself/private-AS/dedicated-AS, tagging the route 208.128.0.0/19 with "no-export"

- advertise both 160.75.0.0/16 and 208.128.0.0/19 to its BGP customers (if any) and suppress the more-specifics originated by itself/private-AS/dedicated-AS, plus any other routes the customers may desire to receive

Address Renumbering:

Central issue: The only known addressing scheme which produces scalable routing mechanisms depends on topologically aggregated addresses.

This scheme requires that sites ***renumber*** when their "***position in the global topology***" changes either voluntarily or involuntarily.

- "***Automatic***" renumbering is really important when prefixes are to be changed periodically to get the addressing topology and routing optimized.
- However, renumbering goes beyond just changing addreses ! Configuration dependencies in protocols and applications exists which pivot around IP addresses. Eg: FTP, TCP "socket", IPSec/IKE associations, DNS name-address mappings use/depend upon IP addresses ! Eg issue is whether existing TCP connections (using the old address(es)) should be maintained across renumbering.
- Renumbering remains operationally difficult and expensive. One of the "carrots" of Ipv6 is that it provides basic automatic renumbering mechanisms. because of the dependencies mentioned, IPv6
- renumbering cannot be truly automatic or instantaneous, but it has
- the potential to be much simpler operationally than IPv4 renumbering,

Addressing: Private Addresses, NAT, RSIP:

Central issue: Enterprises for various reasons prefer to use private IP address spaces for intranet communication. These private addresses must only be locally unique.

- Private addresses are NOT global/public addresses and routers in the core Internet will drop such packets with private IP addresses. Therefore an AS requires a block of public IP addresses, and an entity to map/translate them to private addresses. Such a block might be smaller or larger than the number of private addresses in use.
- The mapping may involve processing every packet on the data path and changing associations at higher layers like FTP etc. This is done by a **"network address translator" (NAT).** Typically NAT functionality is bundled with other AS edge data-plane components (firewalls, traffic management boxes, and rarely with border routers). NAT is completely transparent to hosts in the AS.
- **RSIP: "Realm-specific IP":** It is similar to NAT, in that it allows sharing a small number of external IPv4 addresses among a number of hosts in a local address domain (called a 'realm'). However, it differs from NAT in that the hosts know that different externally-visible IPv4 addresses

are being used to refer to them outside their local realm, and they know what their temporary external address is. The addresses and other information are obtained from an RSIP server, and the packets are tunneled across the first routing realm. Whereas NAT gateways preclude the use of IPsec across them, RSIP servers can allow it

Addressing: Identification vs Location (Mobility issues):

Central issue: In the original IPv4 network architecture hosts are ***globally, permanently and uniquely identified by an IPv4 address.***

- Such an IP address is used for identification of the node as well as for locating the node on the network. IPv4 in fact mingles the semantics of node identity with the mechanism used to deliver packets to the node.

- Debates:

Private vs Public addresses;
Transient vs Permanent;
Unique vs non-unique [Lets focus on this]

- The deployment of mechanisms that separate the network into multiple address spaces breaks the assumption that a host can be uniquely identified by a single IP address.
- Besides that, hosts may wish to move to a different location in the network but keep their identity the same. ***Mobility and roaming require a globally unique identifier.*** Mobile nodes must have a widely usable identifier for their location on the network, which is an issue if private IP addresses are used or the IP address is ambiguous.
- Several technologies at this moment use ***tunneling*** techniques to overcome the problem or cannot be deployed in the case of separate address spaces. ***Tunnels reduce the MTU size.*** The alternative of ***keeping state in the network*** is usually considered a bad thing: it ***reduces the flexibility and breaks the end to end model*** of the network.
- Separation of identity and location will not be available as a near-term solution, and will probably require changes to transport level protocols

References:

- Ferguson, P and H. Berkowitz, "Network Renumbering Overview: Why would I want it and what is it anyway?", RFC 2071, January 1997.
- M. Borella, J. Lo, D. Grabelsky, G. Montenegro "Realm Specific IP: Framework", Work in Progress. NAT working group, IETF. http://www.ietf.org/html.charters/nat-charter.html
- Chen et al, A Framework for Inter-Domain Route Aggregation, http://www.ietf.org/rfc/rfc2519.txt