BGP – A route too far

Michael Silvin
Fredrik Söderquist

Contents

- Background
- Mechanics
- ASN
- iBGP
- Route Reflection
- Confederation
- RR vs Confederation
- Potato Routing
- GigaSunet
- Security
- Interconnecting
- Politics
- Policies
- Filtering
- Learning more

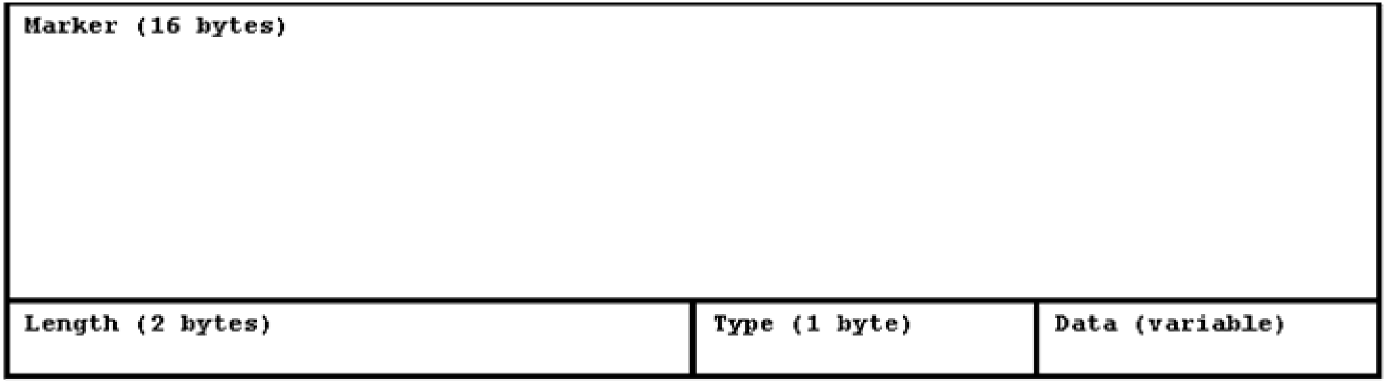
Background

- BGP first became an Internet standard in 1989
- Originally defined in RFC 1105
- The current version, BGP-4, was adopted in 1995 and is defined in RFC 1771
- BGP-4 supports Classless Inter Domain Routing (CIDR)
- Is the routing protocol that people use in today to route between autonomous systems.

Quick look at the mechanics

- Uses TCP to establish a reliable connection between two BGP speakers on port 179.
- Path vector protocol, stores routing information as a combination of a destination and attributes of the path to that destination.
- BGP runs in two modes: eBGP and iBGP
- Five message types are used:

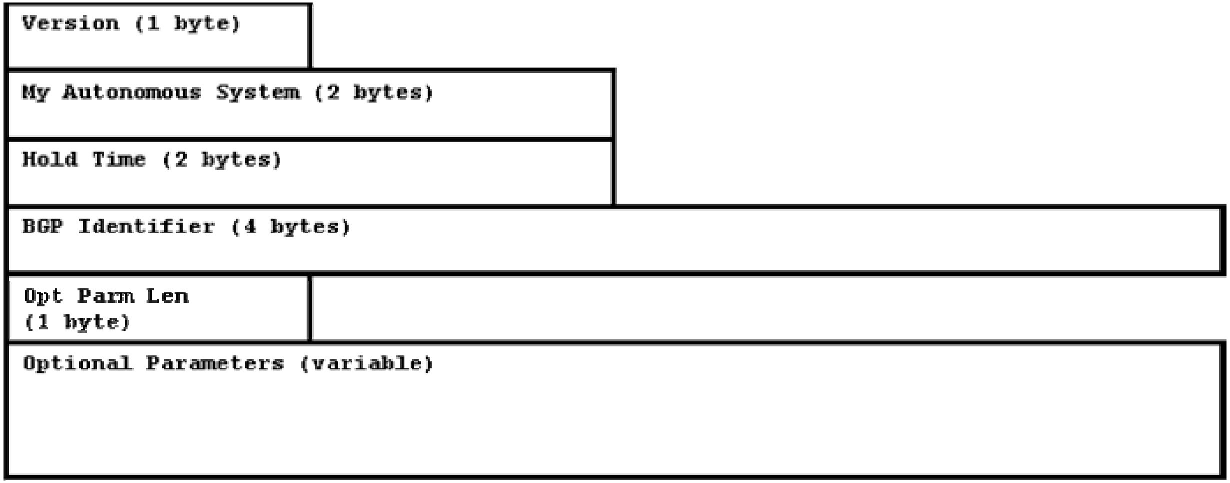
BGP Message Header



Marker (16 bytes)		
Length (2 bytes)	Type (1 byte)	Data (variable)

- The BGP message header is used in all messages

OPEN Message (Type 1 – RFC 1771)



Version (1 byte)

My Autonomous System (2 bytes)

Hold Time (2 bytes)

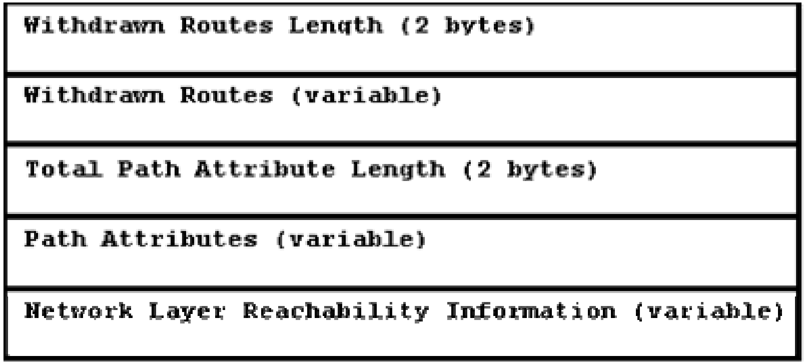
BGP Identifier (4 bytes)

Opt Parm Len (1 byte)

Optional Parameters (variable)

- The first BGP message that is sent after the TCP connection has been established is the OPEN message.
- It is used to exchange configuration information and to negotiate common parameters for the peering session.

UPDATE Message (Type 2)



Withdrawn Routes Length (2 bytes)

Withdrawn Routes (variable)

Total Path Attribute Length (2 bytes)

Path Attributes (variable)

Network Layer Reachability Information (variable)

- UPDATE messages are used to distribute the routing information in BGP
- Are only sent after the session is established.
- An UPDATE message can be used to withdraw existing routes, advertise new routes, or both.

Path Attributes (1/2)

- AS_PATH
  - lists the AS:es traversed by a prefix
  - last AS at the beginning
  - provides loop prevention
- NEXT_HOP
  - Next hop address to reach a prefix (BGP-wise)
  - Must be reachable before being considered by BGP
  - Third-party next hop: next hop received from network protocol peer

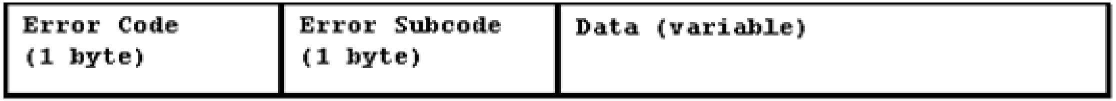
Path Attributes (2/2)

- COMMUNITY
  - A group of prefixes sharing a common property
  - Private (defined by administrator)
  - Well-known (predefined, RFC 1997)
    - NO_EXPORT (not advertised to eBGP)
    - NO_ADVERTISE (not advertised to any peer)
    - INTERNET (no restrictions)

KEEPALIVE Message (Type 3)

- KEEPALIVE messages are sent periodically at 1/3 the *Hold Time* to indicate that a peer is still operating normally to keep the BGP session alive
- Standard hold time in Cisco routers is 120s
- Suggested hold time in RFC1771 is 90s
- This message only contains the BGP header and no data.

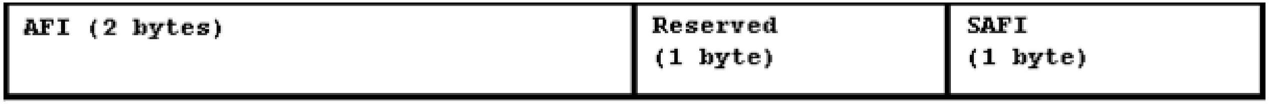
NOTIFICATION Message (Type 4)



Error Code (1 byte)	Error Subcode (1 byte)	Data (variable)

- The NOTIFICATION message is sent when BGP detects an error condition
- Peering session is terminated and the TCP is connection is closed.
- The cause of the error condition is sent to the peer for debugging and troubleshooting.

ROUTE-REFRESH Message (Type 5)



AFI (2 bytes)	Reserved (1 byte)	SAFI (1 byte)

- Not defined in RFC 1771, but as a BGP capability in RFC 2918.
- Is used to request a complete retransmission of a peer’s routing information without tearing down and reestablishing the BGP session.

ASN

- AS Number
- 16-bit number uniquely identifying an AS
- Reserved: 64512 to 65535
  - Should not be advertised on the Internet

iBGP

- Internal BGP (within AS)
- Differs from eBGP/”Normal BGP”
  - Does not add own ASN to AS_PATH
    - Routing information loops might form!
  - Disallow advertisment of prefixes learned via iBGP
    - Requires full mesh connectivity to work
- NOT an IGP!
  - IGP needed to provide infrastructure reachability

iBGP Scalability Issues

- Full mesh requirement → Large AS requires a lot of sessions
- A lot of sessions lead to high resource consumption

iBGP Scalability Issues - Solutions

- Route Reflection
- Confederation

Route Reflection (1/5)

- RFC 2796
- Route Reflectors
  - Relaxed iBGP loop-prevention rules
  - Allowed to readvertise in certain cases
- Speaker classification
  - Route Reflector (RR)
  - Route Reflector client (client)
  - Regular iBGP speaker (non-client)

Route Reflection (2/5)

- A RR reflects routes
  - from non-client to client (and vice versa)
  - from client to client
- Full mesh required between RRs and non-clients
- Ex: 5 routers
  - Full mesh: 10 sessions
  - Route Reflection with 1 RR, 2 clients and 2 non-clients: 5 sessions

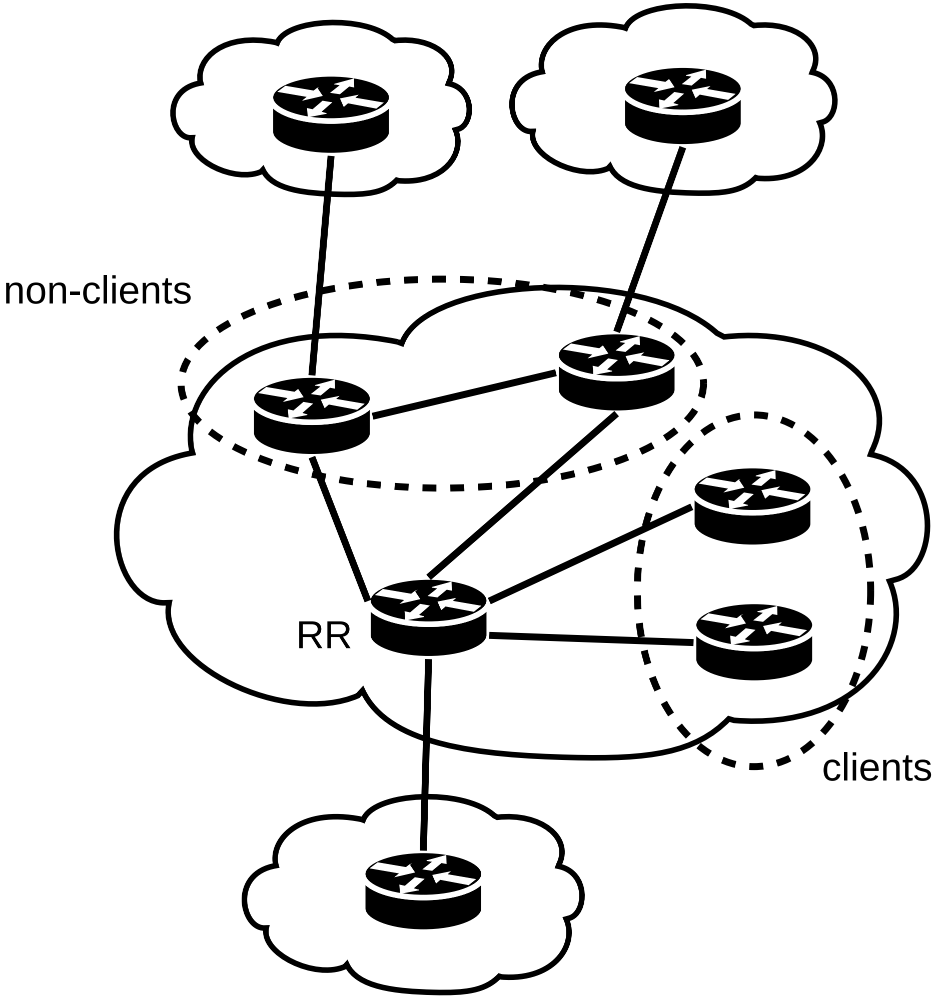
Route Reflection (3/5)

- Rules for prefix advertisement
  - A RR reflects/advertises only its best path
  - A RR always advertises to eBGP peers
  - A client follows the regular iBGP rules
  - When advertising to iBGP peers rules depend on where the prefix was learned.

Route Reflection (4/5)

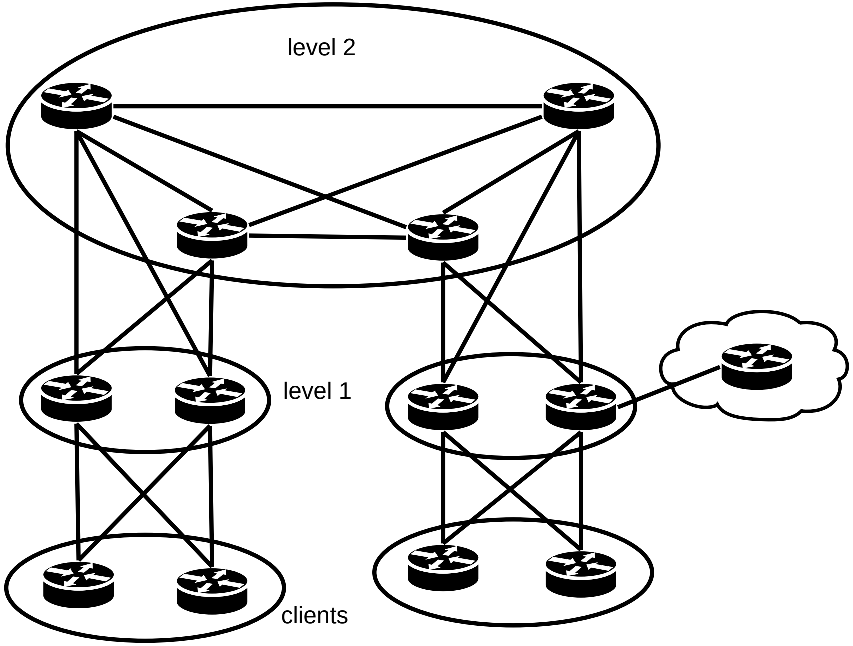
- RR learns prefix from
  - eBGP peer: Advertise to all clients and non-clients
  - non-client: Reflect to all clients
  - client: Reflect to all other clients and to non-clients

Route Reflection (5/5)



Hierarchical Route Reflection

- Several levels of RRs
- Lower level RRs act as clients to higher level RRs
- No limit on the number of levels



Confederation (1/4)

- RFC 3065
- Splits an AS into a number of smaller AS:es
  - Member AS:es/Sub AS:es
- eBGP used among sub AS:es (intraconfederation eBGP sessions)
- Full mesh within sub AS
  - Route Reflection can be used inside a sub AS

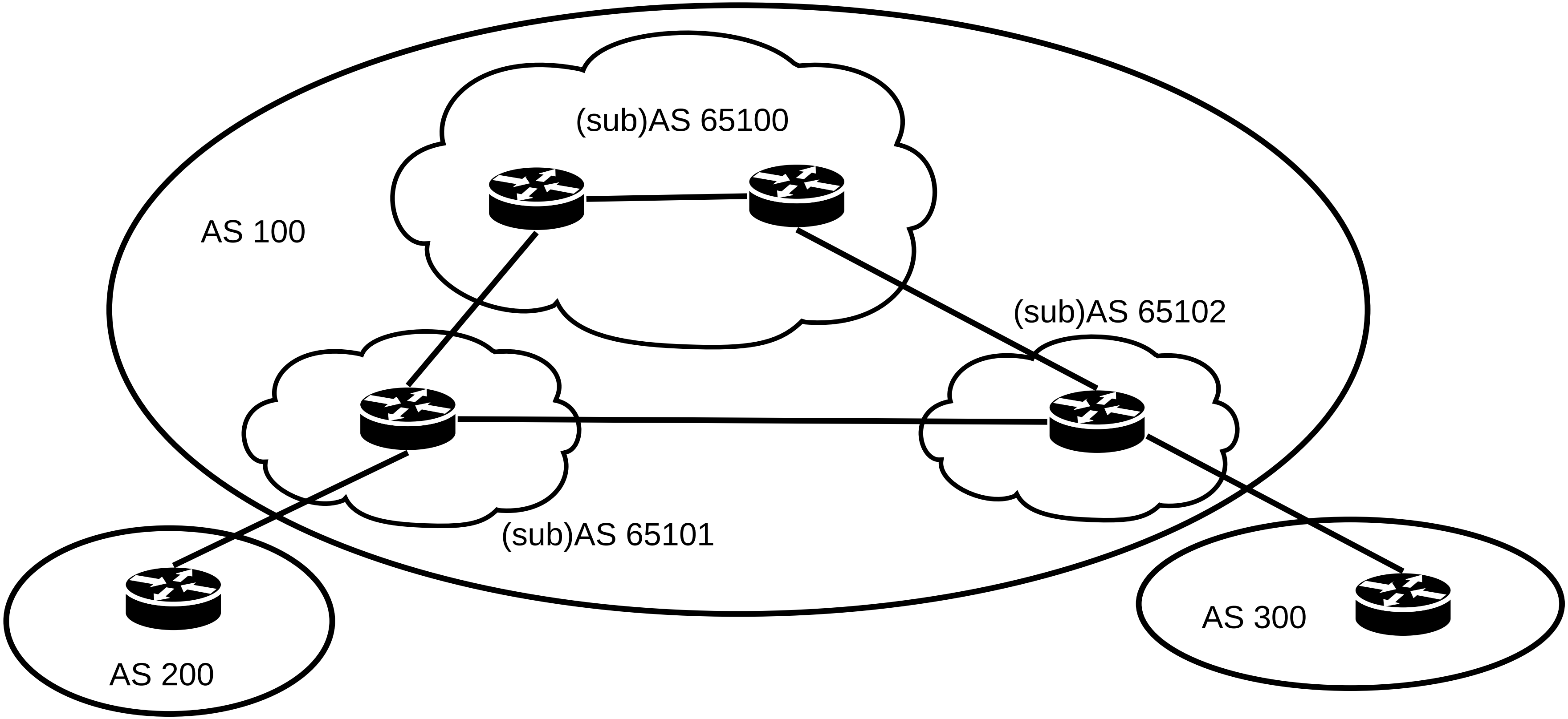
Confederation (2/4)

- Intraconfederation eBGP sessions follow iBGP rules in some cases and eBGP rules in some cases
  - AS_PATH is updated when sending updates
- Three different types of peering
  - External (from confederation to external)
  - Confederation external (between sub AS:es)
  - Internal (within sub AS)

Confederation (3/4)

- The following applies to the different session types (for AS_PATH)
  - External: Sub ASN removed, Confed. ASN prepended
  - Confederation external: Sub ASN prepended
  - Internal: Not modified
- Any range of ASNs can be used in a confederation since these ASNs are not exported.

Confederation (4/4)



Confederation vs. Route Reflection

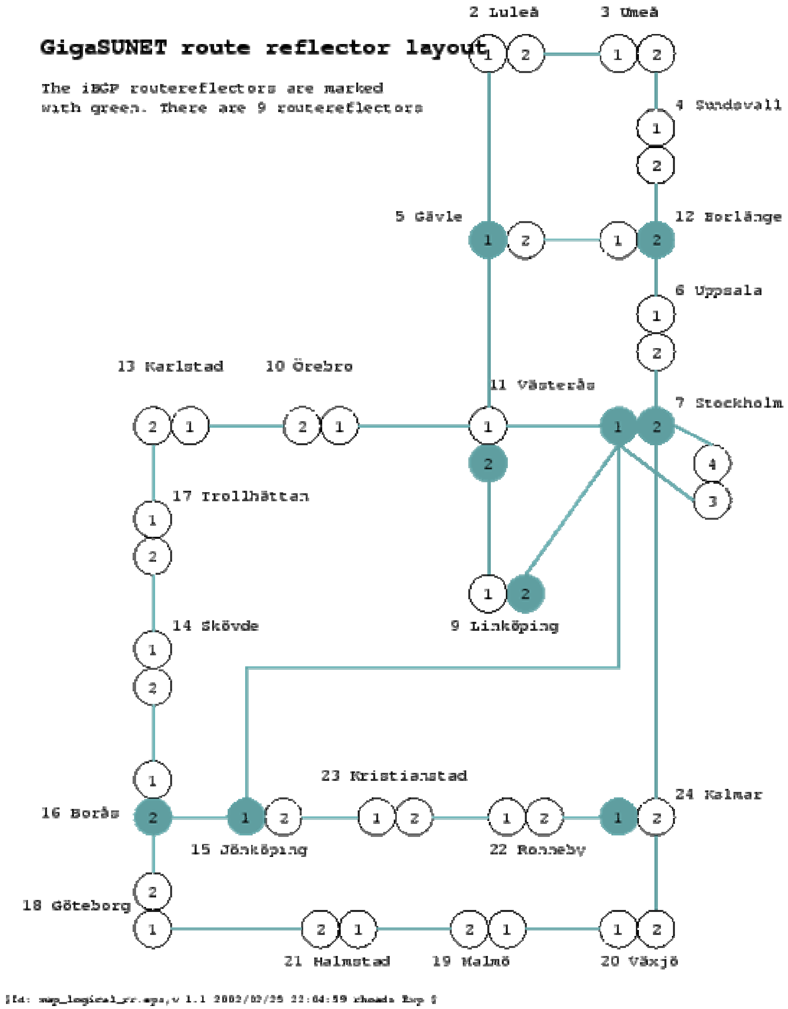
- Hierarchies allowed for both (using Route Reflection sub AS:es in Confederation case)
- Route Reflection requires minor changes when implementing – Confederation requires major changes in configuration and architecture
- Route Reflection requires router support – Confederation requires router support for AS_PATH elements
- Single IGP inside AS for Route Reflection – Single and separate IGP possible in

Hot/Cold Potato Routing

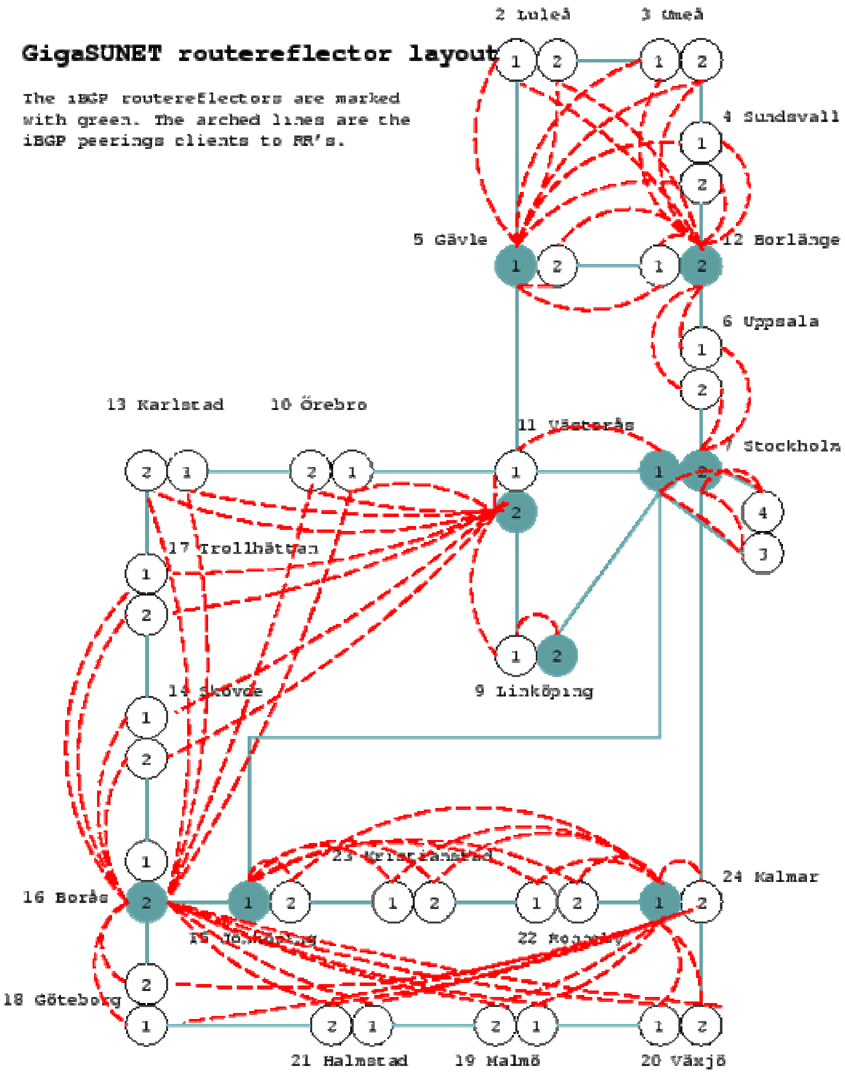
- Hot
  - Let the traffic take the shortest path out of the network (get rid of the ”hot potato”)
  - Cheaper
- Cold
  - Keep the traffic as long as possible
  - Good for QoS

GigaSunet RR Layout

- Original layout of GigaSunet (2002)
- 2 RRs per ring
- Cisco did not think their equipment could handle full mesh…



GigaSunet RR Layout (peerings)



GigaSUNET routereflector layout

The iBGP routereflectors are marked with green. The arched lines are the iBGP peerings clients to RR's.

GigaSunet Layout

- RR to Full mesh in the summer 2003
- Allows for ”hot potato routing”

BGP Security

- Infrastructure attacks
  - Resetting of sessions
- Malicious advertisements
  - Graded route flap dampening
  - Peer/route filtering
  - Public peering
- DDoS countermeasures
  - Dynamic Black Hole Routing

Resetting sessions

- Possible to reset BGP session by guessing TCP session parameters
- Use MD5 signatures (a TCP option) to make this more difficult

Route Flapping

- Routing change that causes a change in the BGP tables (e.g. link goes up/down)
- Problem reduced by using Route Flap Dampening

Route Flap Dampening

- Maintain history for routes/prefixes
- Several parameters control the dampening
  - State (damp, history)
  - Penalty
  - Suppress limit (and maximum suppress limit)
  - Half life
  - Reuse limit

Graded Route Flap Dampening

- All prefixes are equal... not...
- More hosts in /8 than in /24, so shorter suppression time for /8:s
- For essential services such as DNS no graded dampening should be performed.

Public Peering

- Pointing default
  - Point default route into ISP via NAP router
  - Full BGP routes should not be carried by NAP router
- Third-party Next-Hop
  - Redirect peering traffic elsewhere
  - Full BGP routes should not be carried by NAP router

Dynamic Black Hole Routing

- Advertise BGP prefix with next-hop to a null route
- Victim of DDoS will have its prefix advertised with next-hop set to null route
- Prefix advertised to edge of network
- Traffic can also be redirected for analysis (sink router)
- What if we black hole a customers entire prefix?

Interconnecting to other networks (1/2)

- Transit
  - customer allowed to transit the network to reach its destination
- Peering
  - reachability between ISPs (and their direct customers)
  - public peering (Network Access Points (NAPs), Internet eXchange Points (IXPs) and Metropolitan Area Exchanges (MAEs))
  - private peering (ISP to ISP)

Interconnecting to other networks (2/2)

- ISP Tier
  - Level 1, peering only
  - Level 2, peering and transit
  - Level 3, mostly transit, may have peering

Zen of the Day

”Once a customer, never a peer”

Dual-homing / Multi-homing

- Primary reason to use BGP
- Is done by announcing reachability information for your network to two ISPs
- Multihomed networks will need a real AS number which can be obtained from the RIRs.

Requesting IP address space and AS number

- The Internet Assigned Numbers Authority (IANA) is responsible for assigning the protocol numbers used on the Internet. This includes IP addresses and AS numbers.
- But IANA has delegated these activities to a few Regional Internet Registries (RIRs)

RIRs

- APNIC (Asia Pacific Network Information Centre) - Asia/Pacific Region
- ARIN (American Registry for Internet Numbers) - North America and Sub-Sahara Africa
- LACNIC (Regional Latin-American and Caribbean IP Address Registry) – Latin America and some Caribbean Islands (since nov 2002)
- RIPE NCC (Réseaux IP Européens) - Europe, the Middle East, Central Asia, and African countries located north of the equator
- Work to establish an African RIR (AfriNIC).

Two types of IP addresses

- Provider Aggregable (PA)
- Given to ISP which will give out parts of the block to its customers
- Provides Independent (PI)
- Are given directly to network customers. These are fairly rare and to be able to get a PI-block you need to be multihomed.

Different ways to announce your IPs

- Announcing a provider independent prefix
- Shooting holes in ISP PA block
- Request your own PA-block

Policy Routing - IRR

- Repository for routing policies
- Provides information for troubleshooting failures
- European users should use RIPE’s IRR
- Router configurations can be generated directly from the IRR data.

Routing Policy Specification Language (RPSL)

- Based on RIPE-181
- Allow you to specify you routing configuration so that you and others can check your policies and announcements for consistency.
- You can base your policies and router configuration on other peoples policies.
- Maintainer object, AS object, Route object, Set objects, Role object and more.
- RFC2622 RPSL
- RFC2650 Using RPSL In Practice.

RPSL – Maintainer object (1/2)

- Used to introduce some kind of authorization for registrations.
- Lists various contact persons and describes security mechanisms that will be applied when updating objects in the IRR.
- First step in creating policies for an AS.

RPSL – Maintainer object (2/2)

```
mntner:     MAINT-AS3701
descr:      Network for Research and Engineering
remark:     Internal Backbone
admin-c:    DMM65
tech-c:     DMM65
upd-to:     noc@nero.net
auth:       CRYPT-PW  949WK1mirBy6c
auth:       MAIL-FROM .*@nero.net
notify:     noc@nero.net
```

Autonomous System Object (1/2)

- Contains the peering policies of an AS
- Very simple or very elaborate
- whois –h whois.ripe.net AS1653

Autonomous System Object (2/2)

```
aut-num:   AS2
as-name:   CAT-NET
descr:     Catatonic State University
import:    from AS1 accept ANY
import:    from AS3 accept <^AS3+$>
export:    to AS3 announce ANY
export:    to AS1 announce AS2 AS3
admin-c:   AO36-RIPE
tech-c:    CO19-RIPE
mnt-by:    OPS4-RIPE
changed:   orange@ripe.net
source:    RIPE
```

Route Object

```
route:      130.236.0.0/16
descr:      LIUNET
origin:     AS2843
mnt-by:     AS2843-MNT
changed:    ripe-dbm@ripe.net 19941121
source:     RIPE
```

Set Objects

- Used for grouping other objects.
- AS-SET, ROUTE-SET, FILTER-SET

as-set: AS-SUNET

descr: ASes served by SUNET

members: AS1653, AS2831, AS2832, AS2833, AS2834, AS2835, AS2836, AS2837, AS2838

Route filtering

- Inbound / Outbound
- Security reasons
- Resource reasons

Route filtering - Inbound

- RFC1918 addresses, Intended for private networks, should never be advertised globally.
- System local addresses, 127.0.0.0/8 is reserved for use internal to a system.
- End node autoconfiguration block, 169.254.0.0/16, intended for automatic address assignment when a DHCP server is unavailable
- 0.0.0.0/8, sometimes used internally. Is not assigned and should not be used.
- Test network addressing, 192.0.2.0/24, is reserved for test networks. Intended for use in documentation and sample code.
- Class D and E space. Class D is 224.0.0.0/4 and reserved for multicast group and are not advertised by unicast routing protocols. Class E 240.0.0.0/4 is reserved and not in use.
- Unallocated
- whois –h whois.ripe.net FLTR-BOGONS

Route filtering - Outbound

- To protect you from misconfiguration

Learning more

- BGP, Iljitsch van Beijnum, O’Reilly
- http://www.bgp4.as
- http://www.ripe.net/ripe/meetings/archive/ripe-40/tutorials/bgp-tutorial/index.html
- http://www.irr.net

Questions?

Thank you for listening!

Have a nice afternoon