BGP

Introduction and Basic Procedures

 2005/03/11

Border Gateway Protocol (BGP)

- BGP-3
  - Was classful
  - Central AS needed (didn't scale well)
  - Not further discussed here!
  - RFC 1267
- BGP-4
  - Classless
  - Meshed AS topologies possible
  - Used today – discussed in the following sections!!!
  - RFC 1771

BGP is a distance vector protocol. This means that it will announce to its neighbors those IP networks that it can reach itself. The receivers of that information will say "if that AS can reach those networks, then I can reach them via it".

If two different paths are available to reach one and the same IP subnet, then the shortest path is used. This requires a means of measuring the distance, a metric. All distance vector protocols have such means. BGP is doing this in a very sophisticated way by using attributes attached to the reachable IP subnet.

BGP sends routing updates to its neighbors by using a reliable transport. This means that the sender of the information always knows that the receiver has actually received it. So there is no need for periodical updates or routing information refreshments. Only information that has changed is transmitted.

The reliable information exchange, combined with the batching of routing updates also performed by BGP, allows BGP to scale to Internet-sized networks.

BGP-4 at a Glance

- **Carried within TCP**
  - **Manually configured neighbor-routers**
  - **Therefore reliable transport (port 179)**
- **Neighbor routers establish link-state**
  - **Hello protocol (60 sec interval)**
- **Incremental Updates upon topology changes**
  - **New routes are updated**
  - **Lost routes are withdrawn**
- **Each route is assigned a policy and an AS-Path leading to that network**
  - **Using attributes**

A router which has received reachability information from a BGP peer, must be sure that the peer router is still there. Otherwise traffic could be routed towards a next-hop router that is no longer available, causing the IP packets to be lost in a black hole.

TCP does not provide the service to signal that the TCP peer is lost, unless some application data is actually transmitted between the peers. In an idle state, where there is no need for BGP to update its peer, the peer could be gone without TCP detecting it.

Therefore, BGP takes care of detecting its neighbors presence by periodically sending small BGP keepalive packets to them. These packets are considered application data by TCP and must therefore be transmitted reliably. The peer router must also, according to the BGP specification, reply with a BGP keepalive packet.

Path Vector Protocol

- Metric: Number of AS-Hops
- All traversed ASs are carried in the AS-Path attribute
  - BGP is a "Path Vector protocol"
  - Better than Distance Vector because of inherent topology information
  - No loops or count to infinity possible

Each BGP update consists of one or more IP subnets and a set of attributes attached to them. The intrinsic metric is the number of AS hops. Note that this metric is given implicitly by a AS path attribute, which is a vector of all ASs traversed.

BGP Database

- BGP routers also maintain a BGP Database
  - Roadmap information through path vectors
  - Attributes
- Routing Table calculated from BGP Database
- CPU/Memory resources needed

The designers of the BGP protocol have succeeded in creating a highly scalable routing protocol, which can forward reachability information between Autonomous Systems, also known as Routing Domains. They had to consider an environment with an enormous amount of reachable networks and complex routing policies driven by commercial rather than technical considerations.

TCP, a well-known and widely proven protocol, was chosen as the transport mechanism. That decision kept the BGP protocol simple, but it put an extra load on the CPU or the routers running BGP. The point-to-point nature of TCP might also introduce a slight increase in network traffic, as any update that should be sent to many receivers has to be multiplied into several copies, which are then transmitted on individual TCP sessions to the receivers.

Whenever there was a design choice between fast convergence and scalability, scalability was the top priority. Batching of updates and the relative low frequency of keepalive packets are examples where convergence time has been second to scalability.

Some Interesting Numbers

- Today's Internet BGP Backbone Routers are burdened
  - About 100,000 routes (!)
  - About 10,000 Autonomous Systems
- Although excessive CIDR, NAT, and Default Routes
- Collapse expected
  - Looking for new solutions

Internet routers do a hard job. The number of networks is increasing exponentially since the early 1990s and the only way to overcome routing table exhaustion is to apply excessive supernetting (CIDR), NAT, and default routing. In 2001 about 100,000 routes have been counted in typical BGP Internet router. Moreover, 10,000 ASs have been registrated.

Although this techniques significantly reduce the table growths a collapse is expected to happen in the near future—unless other techniques will be explored.

Basic Idea of BGP is Easy !

1) BGP notifies other Autonomous Systems about reachabilities of networks

2) Each single route has attributes associated to it

3) Routers can apply policies for each route based on these attributes (e.g. filtering routes)

The text above summarizes the basic BGP-4 functionality. As it can be seen its not so complicated as many people think.

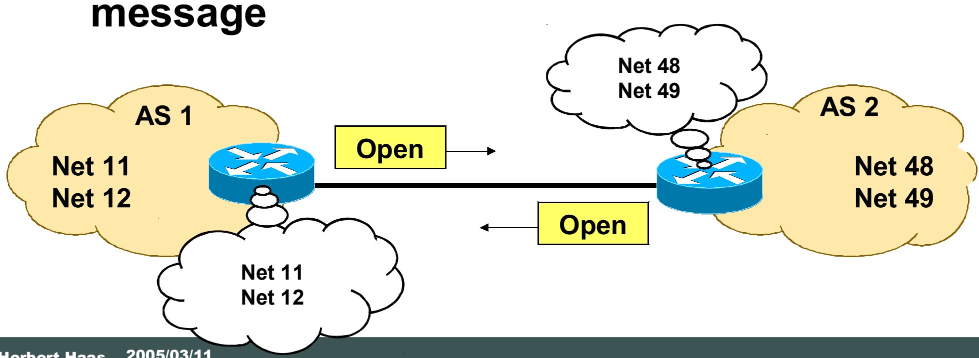
BGP Limitations

- Destination based routing
  - No policies for source address
- Hop-by-hop routing
  - Leads to hop-by-hop policies
  - Connectionless nature of IP
  - Mitigated through
    - Community attribute
    - Peer groups

There are still some limitations in BGP. It is impossible to implement source address-based policies with BGP (unless supported by vendor specific techniques). Furthermore BGP is still hop-by-hop routing, that is, the connectionless nature of IP makes it impossible to foresee what the next routers will do with the route.

Neighborship Establishment

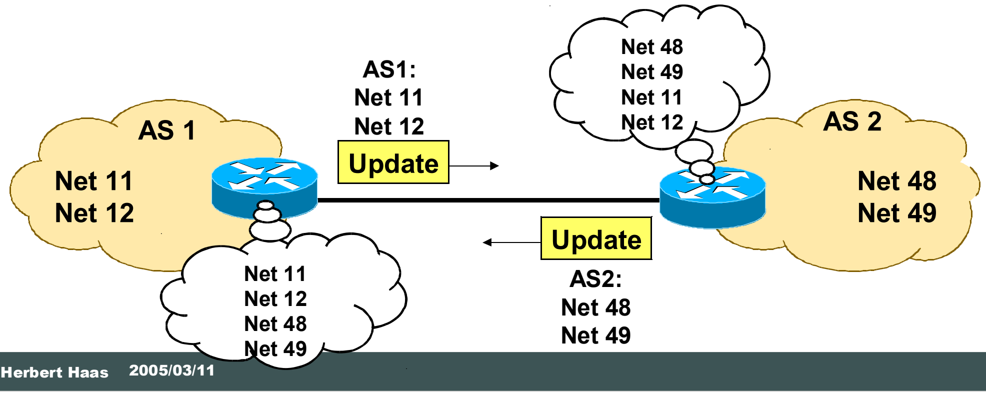
- Open Message
  - BGP Version (4)
  - AS number
  - BGP Router-ID (IP address)
  - Hold Time
- Problems are indicated with Notification message



The BGP protocol is carried in a TCP session, which must be opened from one router to the other. In order to do so, the router attempting to open the session must be configured to know to which IP address to direct its attempts.

NLRI Update

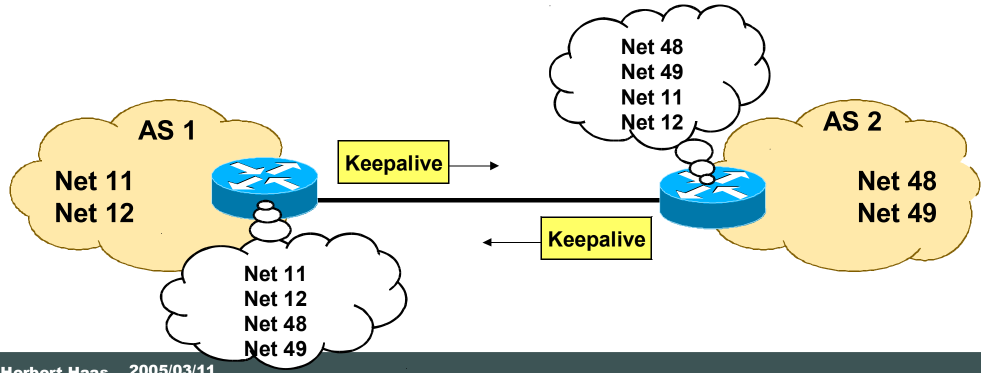
- After open message, all known routes are exchanged using **update** messages
- Contains network layer reachability information (**NLRI**)
  - List of prefix and length



Once the BGP session is established, routing updates start to arrive. Each BGP routing update consists of one or more entries (routes). Each route is described by the IP address and subnet mask along with any number of attributes. The next-hop, AS-path and origin attributes must always be present. Other BGP attributes are optionally present.

Steady State

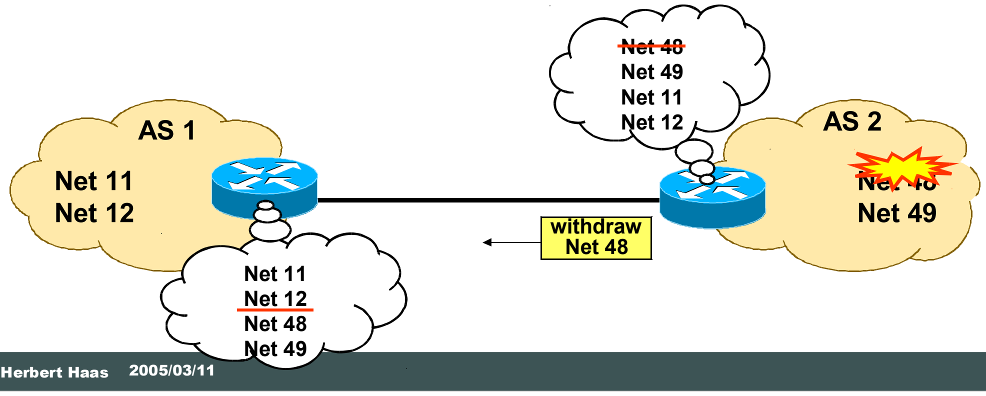
- After Open/Update procedure, BGP is nearly quiet – *No periodic updates* !
- Only keepalive messages are sent
  - 19 Bytes
  - Per default every 60s



After finishing the update process, no periodic updates are sent, just keepalives by default every 60 seconds

Topology Change:

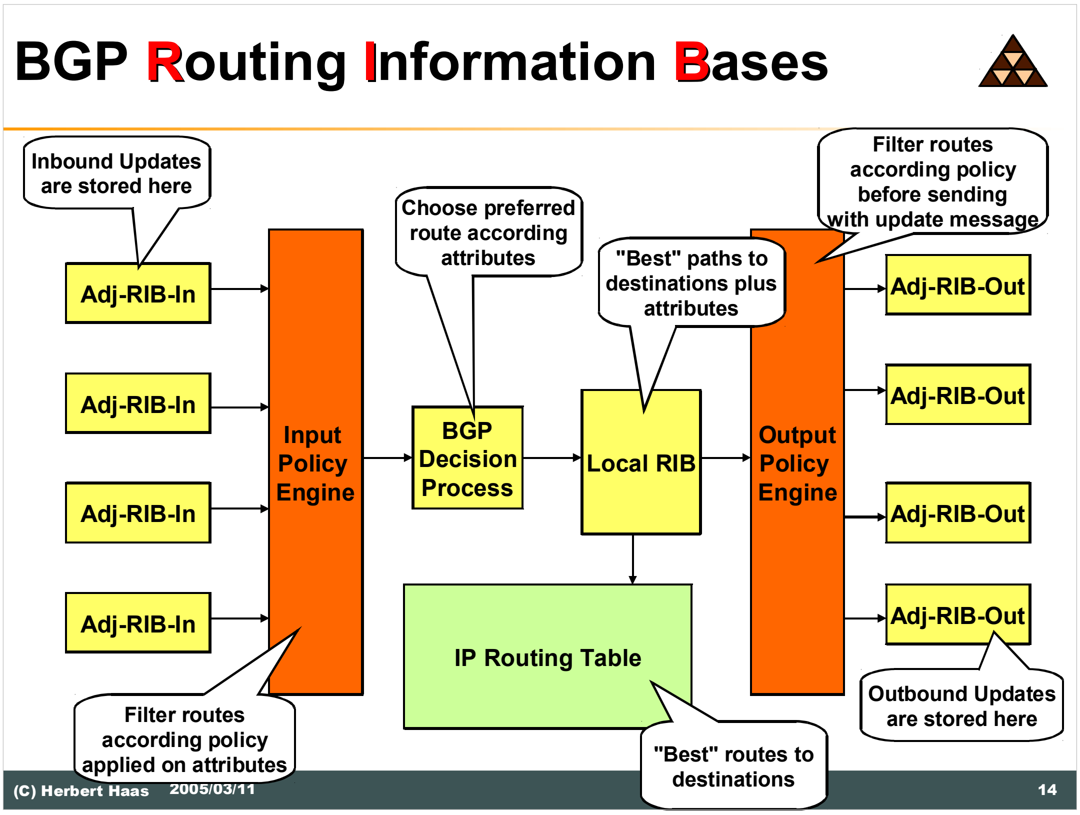
- Incremental Updates upon topology or attribute changes
- Withdraw message upon loss of network



If there is a topology change, only information about the changes is transmited.

RIB

- BGP routing information is stored in RIBs
- RIBs might be combined (vendor specific)
- Only best paths are forwarded to the neighboring ASs
- Alternative paths remain in the BGP table
  - "Feasible routes" in Adj-RIB-In
  - Are used if the original path is withdrawn



The Adj-RIB-In maintains also feasible routes, whereas only the best route is kept in the Local RIB. In case of a withdrawn message for this single best route, the best feasible route becomes active.

Quiz

- How many routes are maintained by BGP today?
- How many AS-numbers have been defined already?
- How long is the typical BGP convergence time?