Routing im Internet - Eine Einführung in BGP

Thorsten Dahm
12.10.2006
t.dahm@resolution.de

Routing?

Weiterleiten von Paketen aufgrund von

- Ziel-Adresse
- Quell-Adresse
- Interface
- weiteren Kriterien

- PI = Provider Independend
- PA = Provider Assigned

CIDR
(Classless Inter-Domain Routing, RFC 1518/1519)

- Suffix = Anzahl „1“-Bits in der Netzwerkmaske
  - z. B.: 192.168.100.0 255.255.255.0 wird zu 192.168.100.0/24 (24 Bits auf "1" gesetzt)

Subnetting:

Aufteilen eines großen Netzes in mehrere kleine

- Supernetting:

Zusammenfassen mehrerer kleiner Netze zu einem großen

- Es gibt keine IP-Klassen mehr!

Summarization / route aggregation

- reduziert Länge der Subnetzmaske bis sie alle zu summierenden Netze umfasst:
  - z.B. 192.168.100.0/24 und 192.168.101.0/24 wird zu 192.168.100.0/23
- Die ultimative Summary-Route: 0.0.0.0/0 (Default-Route, „matched“ alle IP-Adressen)

Administrative Distanz

- „Glaubwürdigkeit“ einer Route
- Entscheidet bei mehreren gleichen Routen aus unterschiedlichen Quellen (Protokollen) welche benutzt wird, z.B.:
  - connected = 0,
  - static = 1,
  - eBGP = 20,
  - OSPF = 110,
  - iBGP = 200

Routing-Protokolle

- Dienen zum Austausch von Routen / Netzinformationen
- Dienen zum Aufbau der Routing-Tabellen
- Anhand Routing-Tabellen wird die Forwarding-Tabelle aufgebaut

Grobe Unterscheidung der Routing-Protokolle

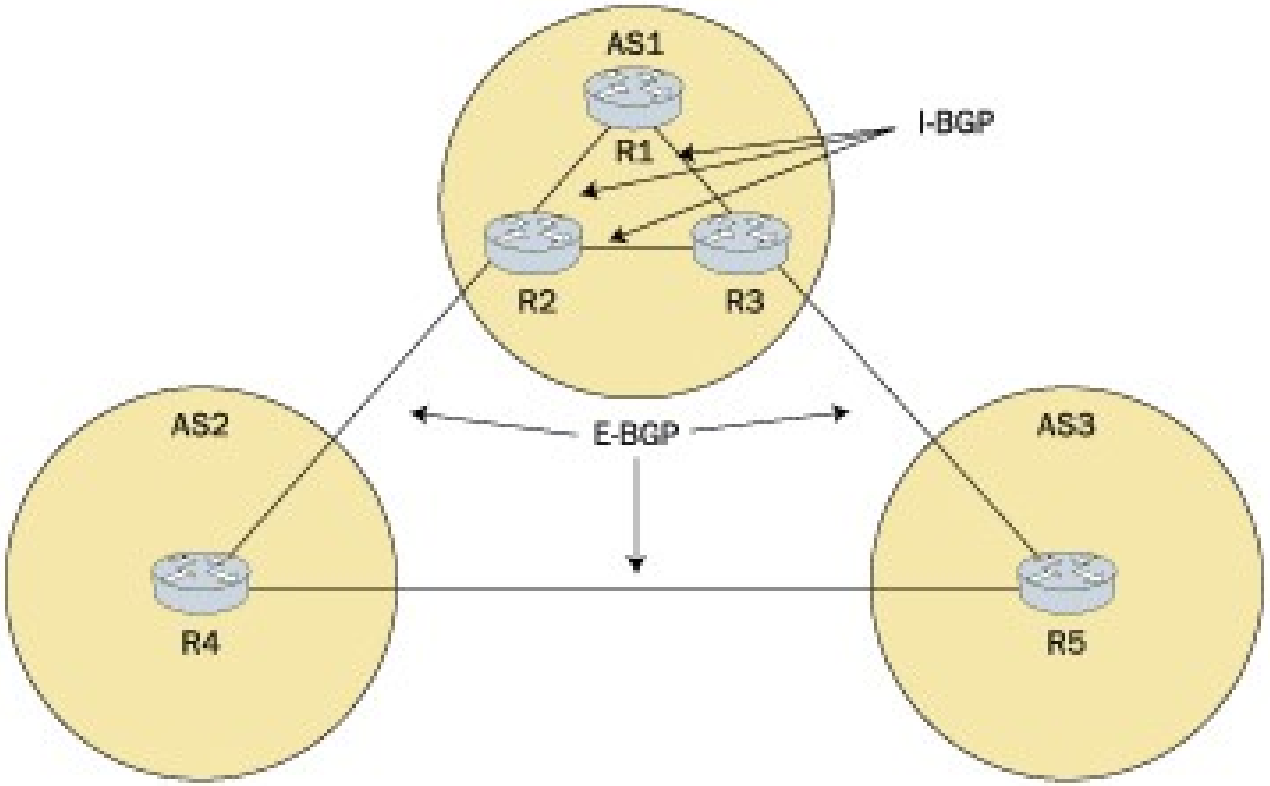
- EGP (Exterior Gateway Protocol):
  BGP 4 (RFC 1771, 1995)
- IGP (Interior Gateway Protocol):
  IS-IS, OSPF, EIGRP, IGRP und RIP

- Intra-AS: wenige Routen, schnelle Konvergenz
- Inter-AS: viele Routen, Stabilität

Border Gateway Protocol - Ein paar Fakten

- AS: Autonomous System
- IP-Netz unter einer einheitlichen administrativen Hoheit mit einer einheitlichen und klar definierten Routing-Policy
- iBGP: internal BGP (innerhalb eines AS)
- eBGP: external BGP (zwischen mehreren AS)
- Path Vector Protocol
- Vergabe von AS-Nummern vom IANA bzw. vom RIPE/ARIN/APNIC usw.
- Private (64512 bis 65535, RFC 1930) und Public AS-Nummern wie bei IP Adressen

Internal vs. External BGP



Aufbau einer BGP-Session

- TCP 179

6 States:

- Idle
- Connect
- Active
- Open Sent
- Open Confirm
- Established

BGP Message-Typen

- Open
- Update
- Keepalive
- Notification

Open

beide Seiten senden ein Open, welches enthält:

- BGP-Versionsnummer
- AS-Nummer
- Hold Time (max. Zeit bis Update oder Keepalive kommt), default: 180 s
- Identifier (höchste IP-Adresse eines Loopback-Devices, dann höchste IP)
- Optionaler Zusatzkrams (z. b. Authentication)

Update

- Network Layer Reachability Information (NLRI) - die Routen
- BGP Attribute
- withdrawn routes

  entfernte Routen & Netze, welche unreachable geworden sind
- neu hinzugekommene Routen

Keepalive & Notification

- Keepalive alle 60 Sekunden
- Notification, dass Verbindung beendet wird

BGP Attribute

BGP benutzt verschiedene Attribute, um den besten Pfad zu einer Destination zu bestimmen:

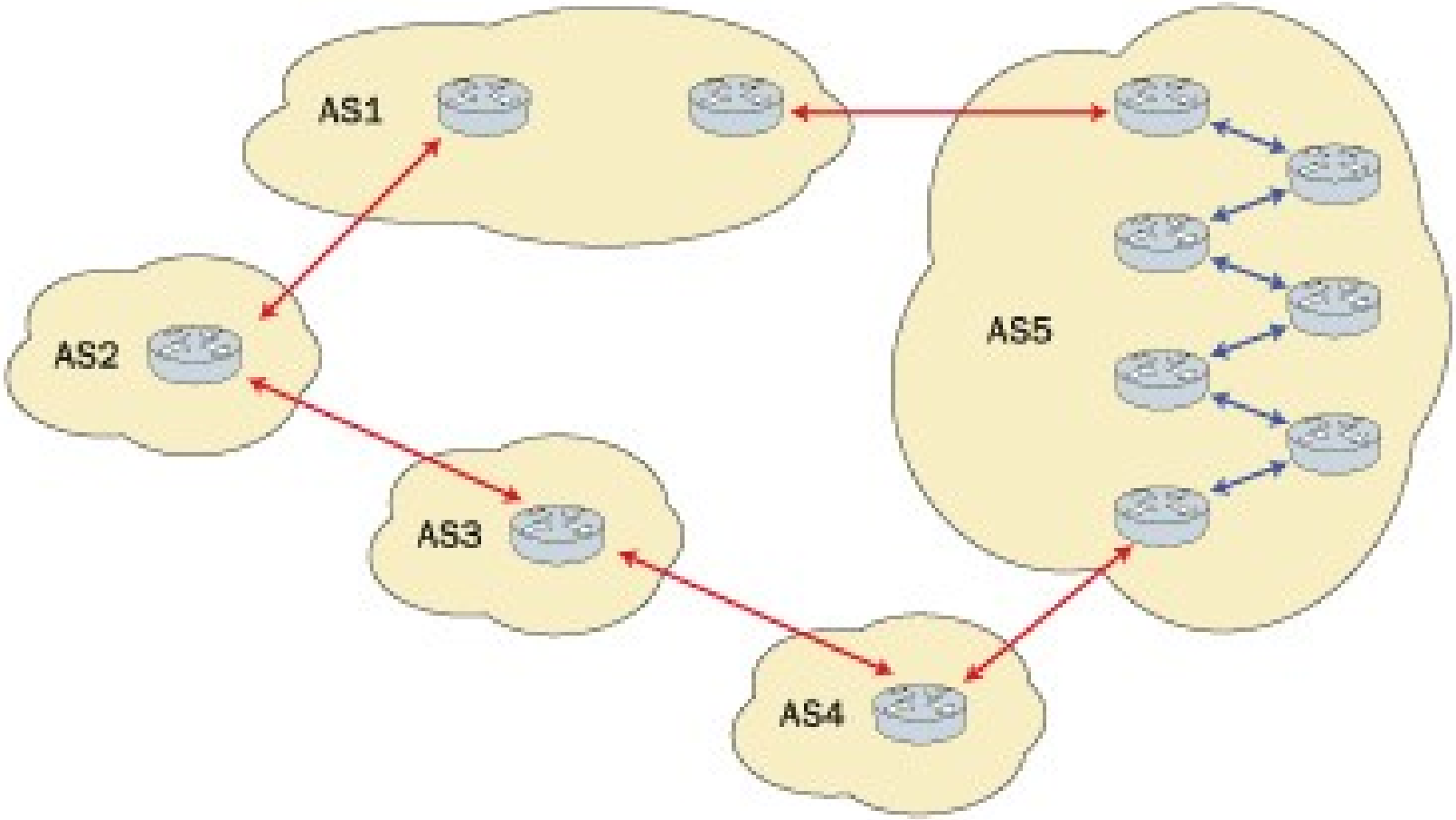
- MED (multi-exit discriminator)
- AS-Path
- local preference
- community
- ...

BGP Attribute (Fortsetzung)

- MED ist Vorschlag ans Neighbor-AS:

  welcher Pfad soll in Richtung des eigenen AS genommen werden?
- In jedem AS, welches eine Route passiert: ein Hop wird zum AS_Path hinzugefügt
- local_pref bestimmt den Weg den ausgehender Traffic nimmt (wird innerhalb des eigenen AS propagiert)

AS-PATH



Communities / route tagging

- eine Route kann mehr als 1 Community haben
- Policies können Communities benutzen
- Community 1000 = alle Routen,

  Community 2000 = alle Routen aus Deutschland usw.

Well-known Communities

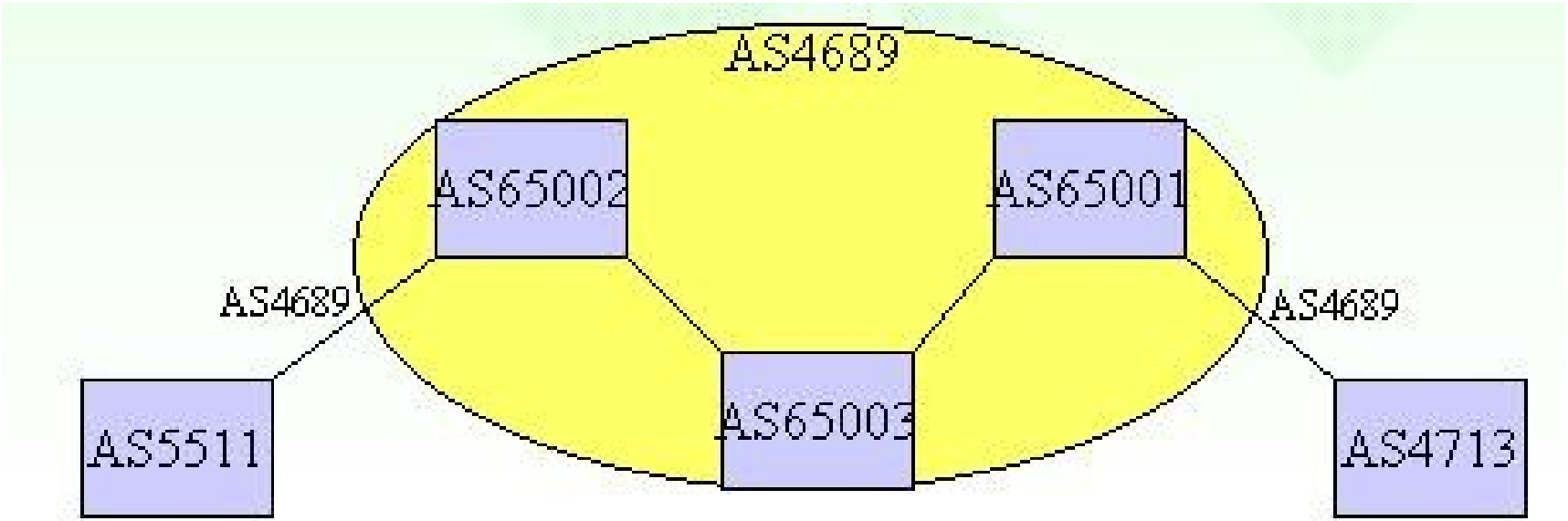
- NO_EXPORT:

  darf nicht an andere AS „advertised“ werden
- NO_ADVERTISE:

  darf nicht an andere BGP-Peers weitergegeben werden
- NO_EXPORT_SUBSET/LOCAL_AS:

  darf nicht an andere AS (auch nicht innerhalb einer Confederation) weitergegeben werden

Confederations (RFC 3065)



Peerings

- private oder public Peering
- Sonderformen: paid Peering
- viele public peering points
  z.B. DECIX in Frankfurt
- Austausch von (Kunden-) Routen zwischen AS
- „jeder trägt seine eigenen Kosten“,
  evtl. schriftliches Peering Agreement

Aufbau einer Forwarding-Tabelle

```
rou-e19-1#sh ip route
Codes: C - connected, S - static, I - IGRP, R - RIP, M - mobile, B - BGP
       D - EIGRP, EX - EIGRP external, O - OSPF, IA - OSPF inter area
       N1 - OSPF NSSA external type 1, N2 - OSPF NSSA external type 2
       E1 - OSPF external type 1, E2 - OSPF external type 2, E - EGP
        i - IS-IS, su - IS-IS summary, L1 - IS-IS level-1, L2 - IS-IS level-2
        ia - IS-IS inter area, * - candidate default, U - per-user static route
       o - ODR, P - periodic downloaded static route

Gateway of last resort is 10.146.52.5 to network 0.0.0.0

B    152.163.0.0/16 [20/0] via 10.149.243.145, 1w1d
   207.200.70.0/27 is subnetted, 1 subnets
B       207.200.70.192 [20/0] via 10.149.243.145, 1w1d
O E2 198.81.11.0/24 [110/20] via 10.146.52.5, 4w5d, FastEthernet0/0
                    [110/20] via 10.146.52.6, 4w5d, FastEthernet0/0
```

show ip bgp summary

rou-e16-1>sh ip bgp sum
BGP router identifier 64.236.165.135, local AS number 25723
BGP table version is 2067, main routing table version 2067
197561 network entries using 20348783 bytes of memory
215369 path entries using 10337712 bytes of memory
38059 BGP path attribute entries using 2284140 bytes of memory
32400 BGP AS-PATH entries using 844780 bytes of memory
406 BGP community entries using 29480 bytes of memory
0 BGP route-map cache entries using 0 bytes of memory
0 BGP filter-list cache entries using 0 bytes of memory
BGP using 33844895 total bytes of memory
BGP activity 327031/129470 prefixes, 1449490/1234121 paths, scan interval 60 sec

Neighbor	V	AS	MsgRcvd	MsgSent	TblVer	InQ	OutQ	Up/Down	State/PfxRcd
64.236.12.5	4	1668	21535	21532	2067	0	0	2w0d	182228
64.236.165.136	4	25723	21535	21532	2067	0	0	2w0d	9320

show ip bgp nei x.x.x.x

```
rou-e19-1#sh ip bgp nei 10.149.243.145
BGP neighbor is 10.149.243.145,  remote AS 65471, external link
 BGP version 4, remote router ID 10.149.244.9
 BGP state = Established, up for 1w1d
 Last read 00:00:12, hold time is 90, keepalive interval is 30 seconds
 Neighbor capabilities:
   Route refresh: advertised and received(old & new)
   Address family IPv4 Unicast: advertised and received

 Inbound soft reconfiguration allowed
 Incoming update prefix filter list is protect_our_net
 Outgoing update prefix filter list is only_our_net

 Connections established 11; dropped 10
 Last reset 1w1d, due to BGP Notification received, cease
Connection state is ESTAB, I/O status: 1, unread input bytes: 0
Local host: 10.149.243.146, Local port: 179
Foreign host: 10.149.243.145, Foreign port: 3102
```

Route Selection

- Basiert auf AS-Path Hop Count (interne Routen werden wegen weniger AS-Hops bevorzugt)
- max. Path per Default nur 1
  - kann für eBGP auf max. 6 geändert werden
  - iBGP kann immer nur 1 Link benutzen
- Sieht ein Router sein eigenes AS im Path verwirft er die mittels eBGP gelernte Route

Route Selection, Fortsetzung

- weight (Cisco proprietär)
- Höchste local_pref
- shortest AS-Path
- lowest MED
- shortest path to next hop
- lowest BGP-ID
- ...

show ip bgp

```
rou-e19-1#sh ip bgp
BGP table version is 7182, local router ID is 10.146.52.135
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal
Origin codes: i - IGP, e - EGP, ? - incomplete

  Network          Next Hop            Metric  LocPrf  Weight  Path
*> 10.0.0.0/16     10.149.243.145                           0 65471 65470 65003 i
* i                 10.146.52.136            100             0 65471 65470 65003 i
*> 10.1.0.0/16     10.149.243.145                           0 65471 65470 65003 i
* i                 10.146.52.136            100             0 65471 65470 65003 i
```

Filter

- RFC 2267 (informational)
- Ausgehend:

  nur eigene Netze und Kundennetze erlauben
- Eingehend:

  private und martian Adressen filtern
- RIPE-Tool: automatisches Generieren von Filtern

route dampening

- valide Routen werden invalid und dann wieder valid (route flapping)
  - Penalty für jeden Flap
  - Ab einem bestimmten Penalty-Wert wird die Route nicht mehr advertised (route suppression)
  - BGP schickt „dampening announcement“ für diese Route
  - Pro definiertem Zeitabschnitt wird Penalty-Wert halbiert
  - Bei Unterschreiten eines bestimmten Penalty-Werts wird die Route wieder advertised

BGP Syncronisation

- „fully mesh“ ist notwendig:
  - um „routing loops“ zu verhindern
  - sicherzustellen, dass jeder Router im Pfad die notwendigen Informationen hat, um ein Paket weiterzuleiten
- Eine iBGP-Route wird erst benutzt,

  wenn es im IGP eine Route zum next-hop gibt

Kann Linux / Unix das auch?

- Zebra (http://www.zebra.org/)
  unterstützt BGP, OSPF, RIP
- Quagga (http://quagga.net)
  Weiterentwicklung v. Zebra
  - modular, für jedes Protokoll 1 Prozess
  - Konfiguration Cisco CLI (IOS) nachempfunden
  - Zentrale Login-Shell für alle Daemons

Zebra / Quagga

- Open Source
- unterstützt zur Zeit Linux, FreeBSD, NetBSD, OpenBSD, Sun Solaris
- Zebra-Daemon für Kommunikation zwischen Kernel + Routing-Protokoll-Daemon
- Erstellt die Kernel-Routing-Tabelle
- Tauscht Routing-Informationen zwischen den einzelnen Routing-Prozessen aus

C-BGP & Cisco Simulator

- BGP-Simulator für Linux/Unix
- Simulation mehrerer AS am PC
- Beherrscht fast alle BGP-Features
- URL: http://cbgp.info.ucl.ac.be/

- http://www.ipflow.utc.fr/index.php/Cisco_7200_Simulator
- Simuliert Cisco 7200 und 3600 Series auf einem normalen PC

Ende

Fragen?