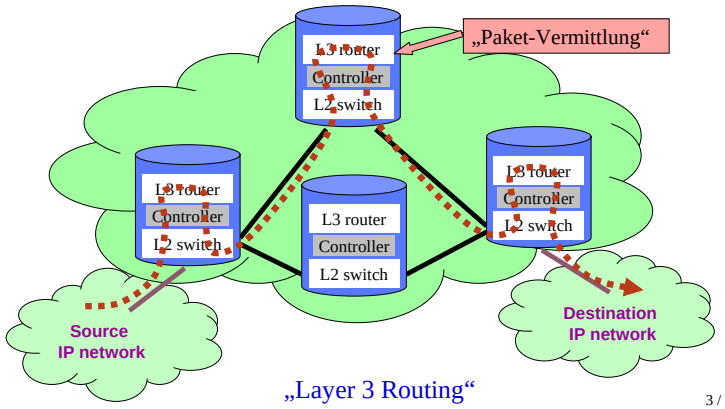
MPLS / BGP
neue Wege bei der QoS-Signalisierung

ITG-FG 5.2.3 - Next Generation Networks
17. Sitzung am 28. 3. 2007 – Arcor

Thomas Martin Knoll
TU Chemnitz - Professur Daten- und Kommunikationstechnik
Telefon **0371 531 33246**
Email **knoll@etit.tu-chemnitz.de**

IP Routing



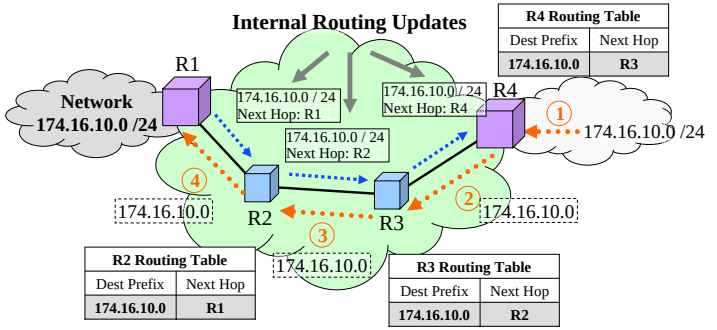
„Layer 3 Routing“

Gliederung

1. IP Routing
2. BGP
3. MPLS
4. MPLS-VPNs
5. Inter-AS-QoS

IP Routing Beispiel

- IGP Routing Updates + Packetweiterleitung auf IP-Schicht

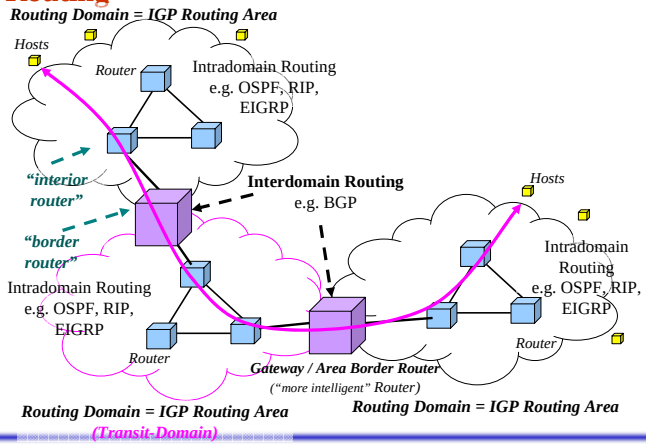


R4 Routing Table	
Dest Prefix	Next Hop
174.16.10.0	**R3**

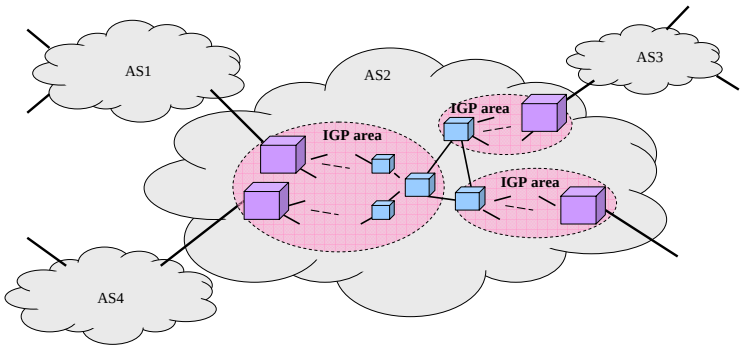
R2 Routing Table	
Dest Prefix	Next Hop
174.16.10.0	**R1**

R3 Routing Table	
Dest Prefix	Next Hop
174.16.10.0	**R2**

IP Routing



IP Routing –> Inter-AS Routing



AS … autonomous system

Autonomous Routing Domains

= Flickenteppich von über IP verbundener physikalischer Netze, die ein einheitliches Regelwerk zur Wegewahl verwenden.

Autonomous Systems (ASes)

Ein "Autonomous System" ist eine "Autonomous Routing Domain", der für die administrative Verwaltung eine AS-Nummer zugewiesen wurde.

… the administration of an AS appears to other ASes to have a single coherent interior routing plan and presents a consistent picture of what networks are reachable through it.

RFC 1930: Guidelines for creation, selection, and registration of an Autonomous System

ASNs

- 16 Bit
- Letzte 1024 Nummern für private Nutzung (64512 … 65535)

BGP

- BGP = Border Gateway Protocol
- Policy-Based routing protocol (Distance Vector)
- De facto EGP des heutigen Internet
- Einfaches Protokoll, das jedoch durch Sichtbarkeitsregeln, Wichtungen und frei definierbaren Attributen schnell komplex wird.
- Optimiert für Stabilität → Trägheit bewußt in Kauf genommen
- Nachteil: neuralgischer Punkt der Interconnection
  - kleine Änderung → teils globale Auswirkung
  - Fehler sind sofort für andere sichtbar (redistribution overwrite)
  - starke vertragliche Reklementierung (confidential peering agrements)
  - Hidden filtering, multihoming, load balancing, AS spoofing, …

BGP Ablauf

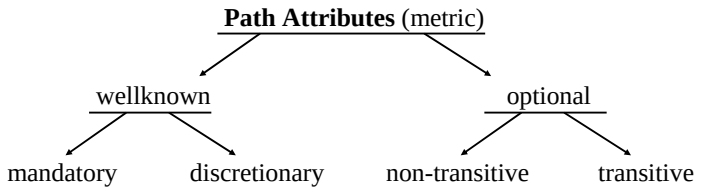
- Konfiguriere (manuell) Nachbarn mit IP-Adresse und ASN
- TCP-Verbindung auf Port 179
- Austausch von **Routing Prefixes & Attributen** !
- Inkrementale Updates (ca. 30 Sek. kumuliert + "gedämpft")

BGP Nachrichtentypen & Attribute

- Open : Aufbau einer "peering session"
- Keep Alive : regelmäßiges Handshake (falls kein Update)
- Notification : Sitzungsabbau
- Update : "Routing Announcement" = Prefixes + Attributes

http://www.iana.org/assignments/bgp-parameters

BGP Attribute



- Origin (i,e,?)
- AS-Path
- Next Hop

- Local Preference
- Atomic Aggregate

- MED (multi exit discriminator)

"partial bit"
- Aggregator
- Community

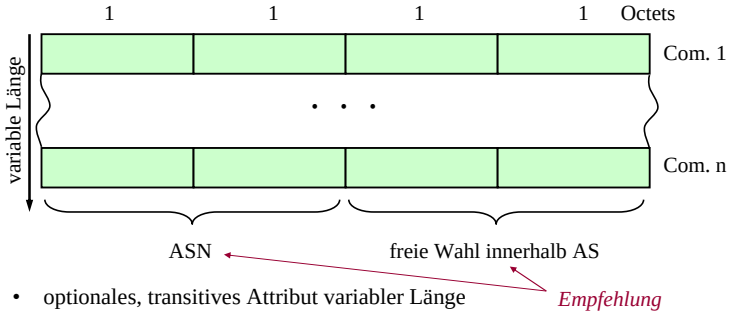
BGP Communities Attribute - RFC 1997

- BGP-Erweiterung für zusätzliche Informationen an Nachbar- und entfernte „BGP-Peers"
- zusätzliche (neben Prefix und AS-Path) Möglichkeit zur automatisierten Verwaltung / Filterung der Verteilung von Wegeinformation
- Routen mit gleicher „Community"-Information bilden eine Gruppe und können gebündelt verwaltet werden

Community
A community is a group of destinations which share some common property.

Well-known Communities:
NO_EXPORT (0xFFFFFF01)
NO_ADVERTISE (0xFFFFFF02)
NO_EXPORT_SUBCONFED (0xFFFFFF03)

BGP Communities Attribute - RFC 1997 cont.



- optionales, transitives Attribut variabler Länge
- Reservierte Werte (erste und letzte 64K):
  0x0000000 .. 0x0000FFFF und 0xFFFF0000 .. 0xFFFFFFFF
- kann ausgewertet, ignoriert, hinzugefügt, entfernt und geändert werden

BGP Entwicklung

- 1989 : BGP-1 [RFC 1105]
  - Replacement for EGP (1984, RFC 904)
- 1990 : BGP-2 [RFC 1163]
- 1991 : BGP-3 [RFC 1267]
- 1995 : BGP-4 [RFC 1771]
  - Unterstützung für Classless Interdomain Routing (CIDR)

BGP relevante „Basis“-RFCs

1930 Guidelines for creation, selection, and registration of an Autonomous System
1774 BGP-4 Protocol Analysis
1773 Experience with the BGP-4 protocol
1772 Application of the BGP in the Internet
1771 A Border Gateway Protocol 4 (BGP-4)
1745 BGP4/IDRP for IP---OSPF interaction
1675 BGP MIB

BGP relevante „Add-On“-RFCs

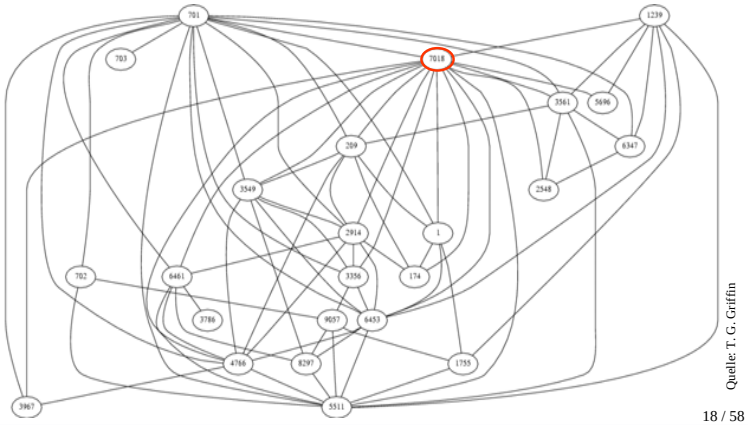
1965 Autonomous System Confederations for BGP
1997 BGP Communities Attribute
1998 An Application of the BGP Community Attribute in Multi-home Routing
2385 Protection of BGP sessions via the TCP MD5 Signature Option
2439 BGP Route Flap Damping
2796 BGP Route Reflection - An alternative to full mesh IBGP

BGP relevante „Add-On“-RFCs

2842 Capabilities Advertisement with BGP-4
2858 Multiprotocol extensions for BGP-4
2918 Route Refresh Capability for BGP-4

Internet-Routen / BGP

> 100 Nachbarn



IP Route Selection → Administrative Distances

Eigene Schnittstelle	0
Statische Route	1
eBGP	20
EIGRP	90
IGRP	100
OSPF	110
RIP	120
iBGP	200
Unbekannt	255

BGP Route Selection

- Entferne Routen mit unerreichbarem “Next-Hop”
- Bevorzuge “highest weight” (Router-lokal) → Cisco-spez. !
- Bevorzuge “highest local-preference” (global innerhalb AS)
- Bevorzuge Routen, die vom eigenen Router ausgehen
- Bevorzuge kürzere AS-Pfade (kürzeste Liste)
- Bevorzuge “lowest origin code” (“IGP < EGP < Unknown”)
- Bevorzuge “lowest MED”
- Bevorzuge “eBGP over iBGP”
- Für iBGP: bevorzuge Pfade zum nähesten IGP-Nachbarn
- Für eBGP: bevorzuge älteste (stabile) Pfade
- Bevorzuge Pfade des Routers mit der kleineren “BGP router ID”

BGP-“Tricks“

- **AS „spoofing“**
- **Source „spoofing“**
- **Filterung nach regulären Ausdrücken** auf (nahezu) alle Attribute
- **Wichtung der Update-Quellen** (Cisco metric)
- **AS confederations**
- **TTL-Security Check**
- **Split Horizon / Route Reflector** → gute Beispiele für skalierbare Implementierungen
- „**Penalty System**“ für Ausfälle
- „Politik der ruhigen Hand“ (**Update-Zeit, Dampening**) sorgt für Stabilität
- **Community-Technik** für direkte Absprache mit Nachbarn

Multi-Protocol Label Switching

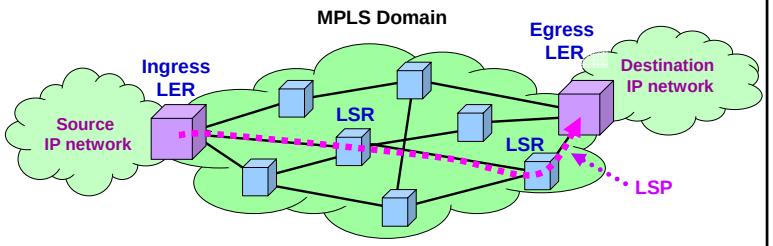
Motivation / Entwicklungsziele:

- **Weiterentwicklung der Routing-Architektur** von IP Netzen
- **Leistungssteigerung** (oder besseres Preis/Leistungsverhältnis) in der Vermittlungsfunktion (Router)
- Stetig steigende Komplexität der **Abbildung von IP**-Datenströmen **auf ATM**-Netze
- **Skalierbarkeit** der IP-Vermittlung
- Einführung **neuer Vermittlungsfunktionen** (Steuerbarkeit)

Einheitlicher (universeller) **Vermittlungsalgorithmus** +
freie Auswahl an Steuerprotokollen

MPLS wird derzeit in der “**IETF MPLS working group” spezifiziert.**

MPLS-Netz-Architektur



LER (ingress / egress) … Label Edge Router (Eingang / Ausgang)
LSR … Label Switch Router
LSP … Label Switched Path

Forwarding Equivalence Class - FEC

- Einteilung aller möglichen Pakete, die ein Router weiterleiten kann, in eine endliche Zahl **getrennter Klassen** => **“Forwarding Equivalence Classes”**
- Router **behandeln alle Pakete einer Klasse gleich** – ungeachtet der Kopfinformationen der Netzwerkschicht

“Forwarding”-Granularität:

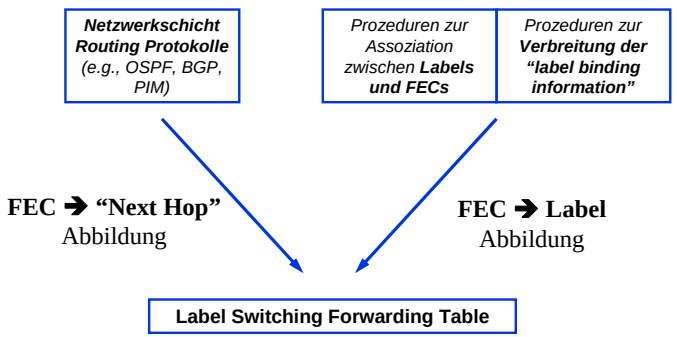
Fein (z.B. gleiche IP-Adresse + Ports)
(-) komplex
(+) applikationsspezifisches Weiterleiten

Grob (z.B. gleicher IP-Prefix)
(+) gut skalierbar
(-) starr / unflexibel

Granularität pro LSP: freie Auswahl + dynamische Steuerung im Router

Conventional routing: packet assigned to a FEC at each hop (i.e. L3 look-up)
MPLS : packet assigned to a FEC only done at network ingress

Label Switching Forwarding Table



Steuerung des LSP-Setups

MPLS path =
Assoziierung von "***FEC next-hops***" und
"**incoming + outgoing labels**"

Independent LSP Control

- **jeder LSR** -> **unabhängige Entscheidung über Label-Generierung und Verteilung**
- versende Label-FEC-Bindung an "Peers" sobald "next-hop" ermittelt
- LSP ergibt sich dann, wenn alle Strecken-abschnitte Label vergeben haben

(+) **schneller Pfad-Aufbau**
(-) unabhängige Wahl der FEC
(-) Homogener Aufbau erfordert Konfiguration aller LSRs
(-) Gefahr der Schleifenbildung

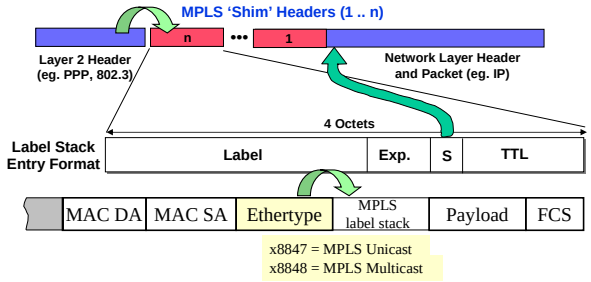
Ordered LSP Control

- nur der **Kopf(ingress) oder** das **Ende(egress)** eines LSP darf **LSP-Setup initiieren**
- LSP-Setup 'fliest' von Ende zu Ende

(-) langsamer Aufbau ("binding" durchläuft gesamte LSR-Region bevor LSP aufgebaut ist
(+) **FEC-Auswahl** erfolgt beim initiierenden LSR => alle LSR nutzen die gleiche FEC !
(+) **einfache Steuerung, welche Pakete** mittels MPLS weitergeleitet werden
(+) **Schleifen können vermieden werden**

Beide Methoden sind im Standard unterstützt und sind interoperabel.

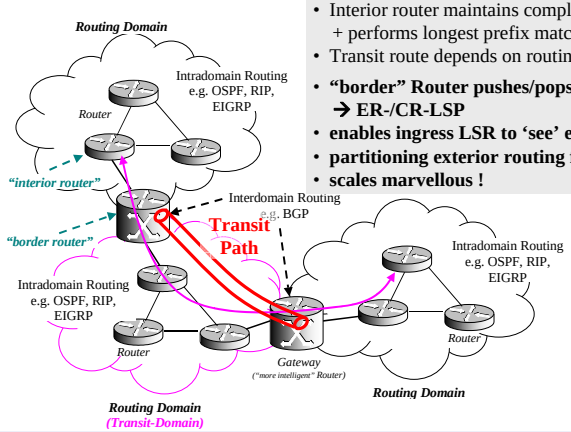
Label-Hierarchie / Label-Stack



- **Labels are pushed/popped** as packets **enter/leave MPLS domain**
- **TTL value copied** as packets **enter/leave MPLS domain** ***(loop mitigation)***
- Top-Label used for forwarding
- Stack allows for **Hierachie of MPLS domains !!**

Traffic Engineering – „Tunnelling" / „Label Stack"

- Interior router maintains complete routing table + performs longest prefix match
- Transit route depends on routing protocol
- **"border" Router pushes/pops label ➔ ER-/CR-LSP**
- **enables ingress LSR to 'see' egress LSR**
- **partitioning exterior routing from interior**
- **scales marvellous !**



Ausfallsicherheit („Restoration“)

Umgang mit Verbindungsausfällen

“Path protection”
(Ausfallschutz für gesamten Pfad)

“Link / Node protection”
(Ausfallschutz für Knoten bzw. Teilstrecken)

Herkömmliche Herangehensweise:

- **Fehlererkennung** (Meldung durch Schicht 1 bzw. 2 / fehlende Updates etc.)
- Ermittlung eines geeigneten Ersatzweges (**Routing-Update**)
- Erneute **Ressoucenreservierung** auf Ersatzweg
- MPLS: zusätzlich Aufbau des neuen **LSP**s

➔ benötigt mehrere Sekunden

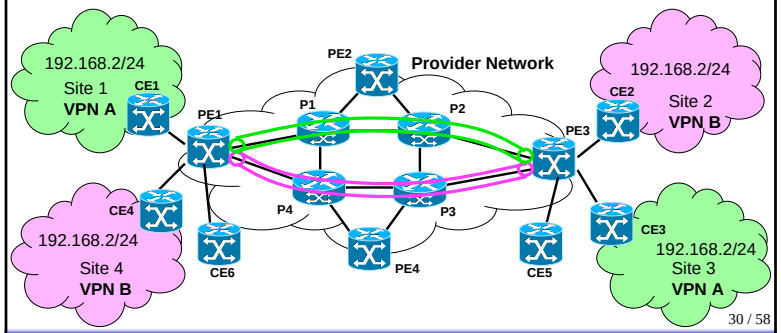
MPLS – “Fast Re-Route” :

- **zuvor Aufbau eines disjunkten Backup-LSPs (mit / ohne Reservierung)**
- **Fehlererkennung**
- **Kopf des Backup-LSP: PUSH-Operation -> Verkehr im Backup-Tunnel**
- **Backup-LSP-Ende: POP-Operation (gegebenenfalls mehrfach)**

➔ benötigt ca. 50 ms

„Virtual Private Network – VPN“

- Vermeidung einer vollen Vermaschung
- Problem: Verwendung privater IP-Adressen → **MP-BGP Adressen**

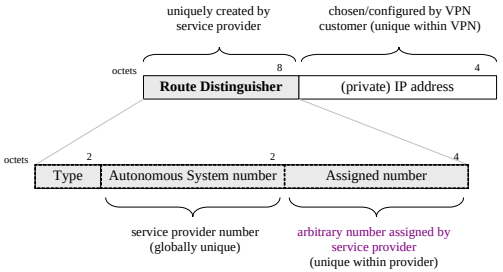


VPN – IP-Adressen – „MP-BGP“

- BGP nimmt an, daß global eindeutige IP-Adressen verwendet werden
- VPN Kunden benutzen jedoch **private IP-Adressen** in ihren privaten Netzen

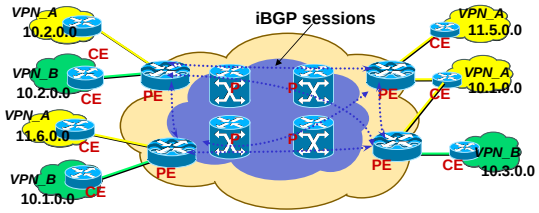
⇒ Generierung von “**globally unique addresses”** aus “VPN-Identifier” und privater IP-Adresse

⇒ **neuer BGP Adresstyp “VPN-IP address”**



VRF-Konzept / MPLS-VPN-Backbone

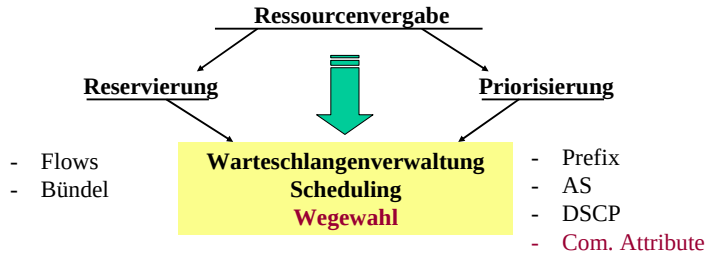
- VRF = VPN Routing and Forwarding
- Virtuelle Router-Instanz pro VPN (→ Subinterface assoziiert)
- VPN-spezifische Routing-Tabellen und (FIBs bei MPLS)
- VPN-Bildung & Dienstebeschränkung durch “Redistribution”
- Vollständiges LSP-Netz zwischen PE-Routern → loopback addresses!
- P-Router = LSP-Relay-Stationen ohne BGP-Funktion



[Cisco]

Quality of Service - QoS

"To NGN or not to NGN - That's the QoStion" (Miguel Lopez)

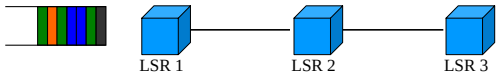


„QoS through lack of knowledge ?" → Wege & Attribute

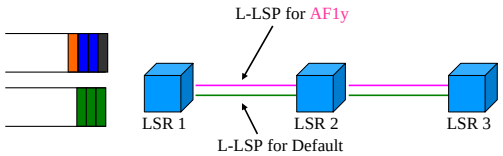
MPLS-Support für DiffServ

- Markierung → Label: "**L-LSP**" (LSP pro DSCP) / "**E-LSP**" *(DSCP → EXP-Bits)*
- **Ressourcenvergabe pro Klasse** – nicht pro Verkehrsfluß

E-LSP (max. 8 Klassen; Unterscheidung anhand **EXP-Bits**)



L-LSP (z.B. **2 LSPs** => PHB + „Dropping Level" innerhalb AF1 LSP anhand EXP-Bits)



Inter-AS QoS-Unterstützung

Ausschließliche IP-DiffServ-Unterstützung

- Einigung auf einheitliche DSCP-Werte notwendig
- ⊕ Sehr einfache Realisierung möglich
- ⊖ Mißbrauch nicht verhinderbar
- ⊖ Üblicherweise keine Auswirkung bei der Wegewahl

DiffServ+MPLS - Unterstützung

- Einigung auf einheitliche DSCP-Werte notwendig
- ⊕ Sehr einfache Realisierung möglich mittels E-LSPs
- ⊕ Verkehrstunnel innerhalb AS → Mißbrauch eingeschränkt
- ⊕ Auswirkung bei der Wegewahl mittels L-LSPs
- ⊖ Inter-AS-Tunnel nicht üblich, da Tunnel-Setup auf den jeweiligen (nicht öffentlichen) IGP-Routen basiert

Inter-AS QoS-Unterstützung

MP-BGP+DiffServ+MPLS - Unterstützung

- Einigung auf einheitliche DSCP-Werte notwendig
- ⊕ Sehr einfache Realisierung möglich mittels E-LSPs
- ⊕ Verkehrstunnel AS-übergreifend → Mißbrauch stark eingeschränkt
- ⊕ Auswirkung bei der Wegewahl mittels L-LSPs & E-LSPs ! (Lösung: Mehrfach-Prefix durch MP-BGP-Adressierung möglich)
- ⊕ Inter-AS-Tunnel möglich, wenn Tunnel-Setup auf den öffentlichen MP-BGP-Routen basiert
- ⊕ Nutzung der „Community"-Attribute zur SLA-basierten QoS-Signalisierung (z.B. Ticketvergabe, Mehrfachklassifizierung bei unterschiedlichen Klassensätzen entlang des Pfades)
- ⊕ „**QoS through lack of knowledge**" durch Attribut-Filterung bei Routing-Updates
- ⊕ Einsatz der BGP-TTL-Security zum vertrauensvollen Austausch der Routen und QoS-Informationen

Zusammenfassung

- **MPLS- (und / oder Carrier grade Ethernet !) Tunnel** sind zentraler Bestandteil der Internet-Struktur
- **VRF** ist ein praktisches Werkzeug zum Kapseln von Routing-Information
- Zunehmendes **Interesse / Notwendigkeit an** Ressourcenreservierung und besonders **Priorisierung des IP-Verkehrs**
- Schwierigkeit der **einheitlichen Klasseneinteilung und Signalisierung**
- **Inter-AS-Tunnel** bedürfen **AS-übergreifender Routing-Informationen**
- **BGP als Vermittler** (sowohl iBGP als auch eBGP !)
- **Idee: Nutzung der BGP-Attribute und MP-BGP-Adressen zum Aufbau QoS-orientierter Inter-AS-Tunnel**
- **Idee: Globale QoS-orientierte Wegewahl durch Filterung der BGP-Updates anhand der QoS-bezogenen Attribute**

Vielen Dank für Ihre Aufmerksamkeit.