Telematik

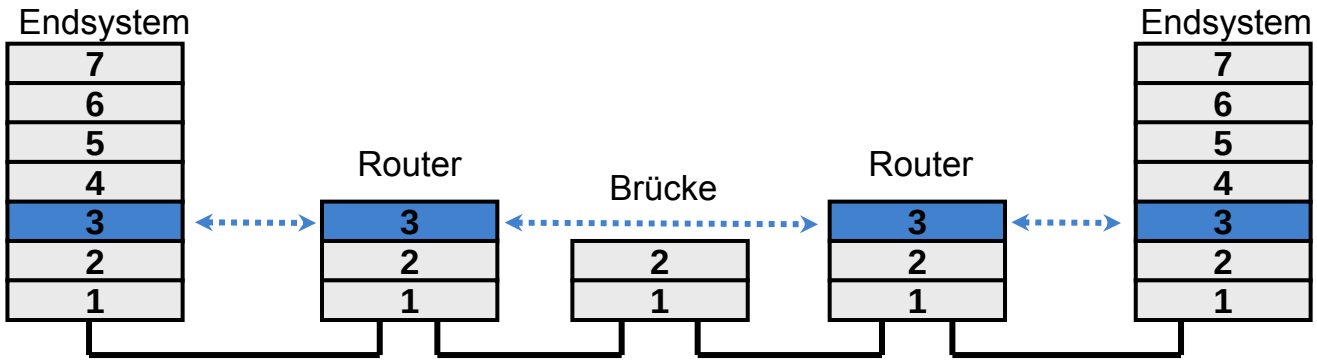
3. Routingprotokolle und -architekturen

Prof. Dr. Martina Zitterbart
Dipl.-Inform. Thomas Gamer
Dipl.-Inform. Martin Röhricht
[zit | gamer | roehricht]@tm.uka.de

3. Routingprotokolle und -architekturen

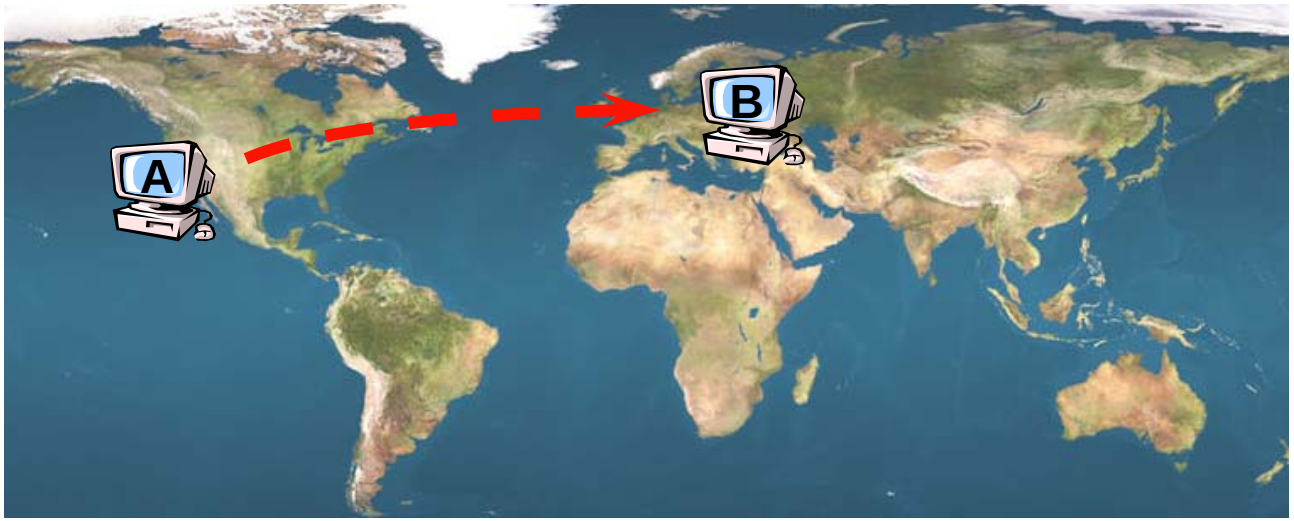
Einordnung in das ISO/OSI-Basisreferenzmodell

- Vermittlung im Schichtenmodell
  - Schicht 3: Vermittlungsschicht (*Network Layer*)
    - Verknüpft einzelne Übertragungsabschnitte zu einer Ende-zu-Ende-Übertragung
    - Wegewahl im Kommunikationssystem
      - Finden geeigneter Routen für die Ende-zu-Ende-Datenübertragung
    - Adressierung der Systeme
    - Multiplexen



Grundlegende Fragestellung

- Ausgangssituation
  - Endsystem A möchte Daten an Endsystem B senden



- Grundlegende Fragestellungen
  - Wie findet das Internet einen Weg zum Zielsystem?
    - Falls möglich einen „guten“ Weg / eine „gute“ Route
    - → Wegewahl / Routing
  - Wie werden die Systeme eindeutig identifiziert?

3.1 Routing – Allgemeines

- Grundlegende Anforderungen
  - Korrektheit und Einfachheit
  - Robustheit
    - Alternative Wege in Situationen lokaler Fehler bzw. Überlastungen
    - Idealerweise keine Verluste von Dateneinheiten bzw. keine Brüche virtueller Verbindungen
  - Stabilität
    - Keine Änderungen wenn nicht erforderlich
    - Vermeiden von Oszillationen
  - Fairness vs. Optimalität
  - Overhead durch Übertragung und Verarbeitung
    - Sollte geringer als der „Benefit“ sein
    - Je mehr Informationen über die Topologie und die aktuelle Verkehrslast verfügbar sind und je häufiger diese ausgetauscht werden, desto besser ist die Entscheidung zur Wegewahl
      - Aber: Hierdurch steigt auch die Verkehrslast im Netz durch Signalisierungsverkehr, d.h. Übertragungen, die nur der Verwaltung/dem Betrieb dienen und keine Nutzdaten transportieren

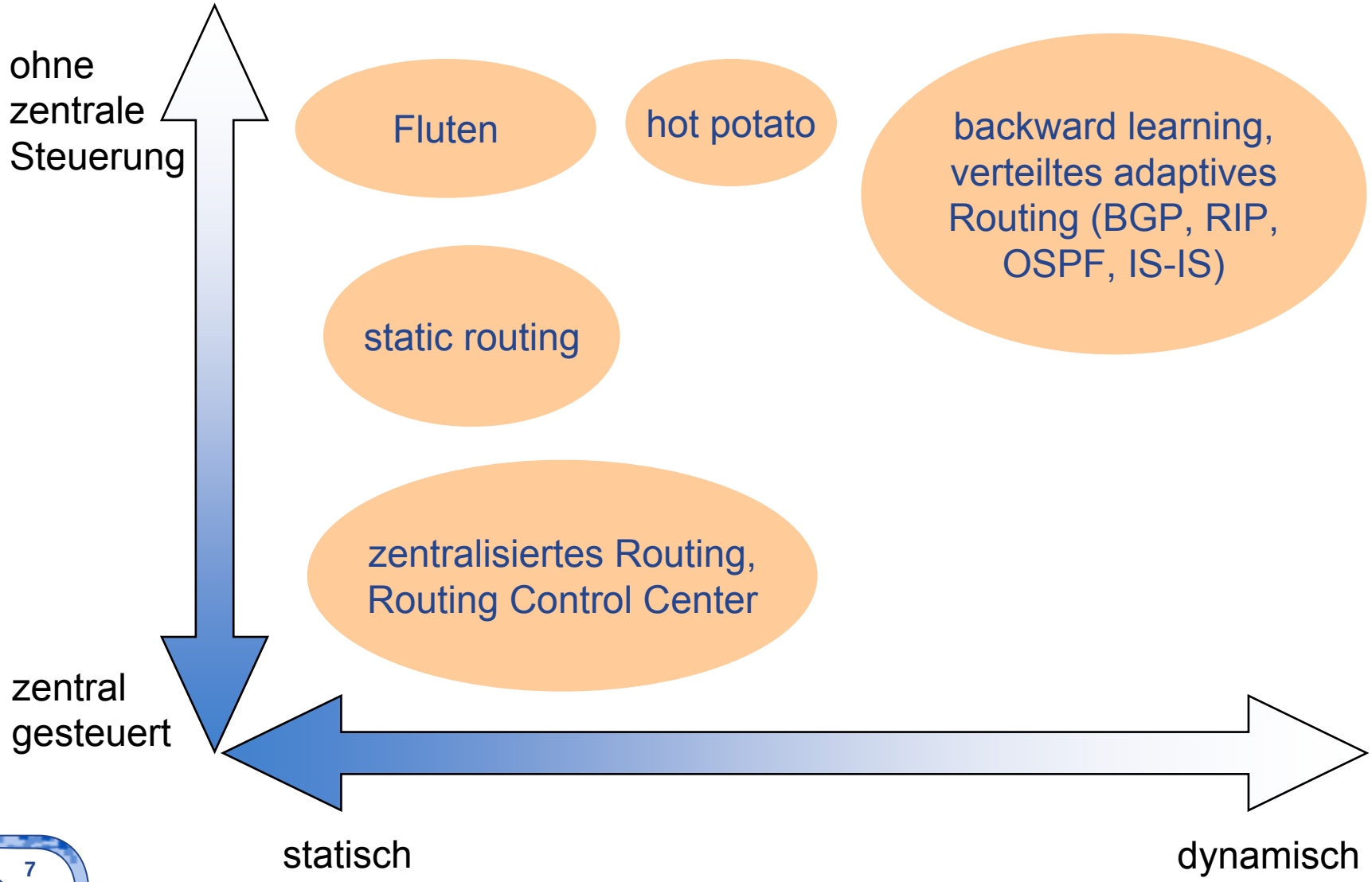
Routing – Begriffe

- Router
  - Auf Vermittlungsschicht operierendes Zwischensystem
    - Übt Wegewahl-Funktionen für eingehende IP-Dateneinheiten aus
  - Üblicherweise mit Schnittstellen zu mehreren anderen Systemen ausgestattet
  - Tauscht über Routingprotokolle Informationen mit anderen Routern aus
- Route
  - Hier: Weg einer Dateneinheit zum Ziel
    - Daher auch „Wege“wahl für die Auswahl der besten Route
  - In der Literatur unter wechselnden Bezeichnungen und Bedeutungen
    - Pfad, Weg, …

Routing – Begriffe

- Routing-Metrik
  - Bewertungskriterium einzelner Übertragungsabschnitte
    - Beeinflusst deren Bevorzugung/Vermeidung
    - Üblicherweise Ganzzahl, die unidirektional den Übertragungs-abschnitten zu weiteren Zwischensystemen zugeordnet wird
      - Auch als Kosten oder Gewicht bezeichnet
        - Bedeutung beachten (Bevorzugung vs. Vermeidung)
    - In Netzwerk-Graphen häufig als Zahl an einer Kante abgebildet (dann häufig bidirektional gültig)
- Routing-Policy
  - Betreibervorgaben zur Routing-Strategie
    - Üblicherweise für ganze Netzwerke
      - Beispielsweise Bevorzugung bestimmter Nachbar-Netzwerke bei der Wegewahl

Routing-Verfahren im Überblick



Adaptives Routing

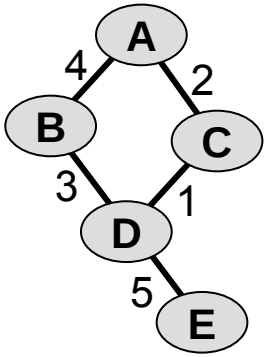
- Das Routing wird beeinflusst durch bzw. reagiert auf
  - Stau: Mindestens ein Übertragungsabschnitt ist mit Verkehr überlastet
  - Linkfehler: Ein Übertragungsabschnitt zwischen zwei Routern fällt aus
- Voraussetzungen für adaptives Routing
  - Beobachtung
    - Information über den Netzzustand wird gesammelt
  - Protokoll
    - Austausch von Routing-Information mittels Routing-Nachrichten
  - Berechnung
    - Berechnet die neuen besten Routen zu den entsprechenden Knoten

Adaptives Routing: Vor- und Nachteile

- Vorteile
  - Verbessert Robustheit des Netzes
    - Im Fehlerfall werden alternative Übertragungsabschnitte verwendet
  - Konfiguriert sich überwiegend eigenständig
- Nachteile gegenüber statischem Routing
  - Routing-Entscheidung ist komplexer
    - CPU-Last auf Routern steigt
  - Routing-Entscheidung basiert auf Informationen des Netzes
    - Können veraltet sein
    - Austausch belastet das Netz zusätzlich
  - Problem Reaktionszeit
    - Evtl. zu schnelle Reaktion des adaptiven Routing-Protokolls, woraus Stau-Situationen und Oszillation entstehen können

3.2 Shortest Path Algorithmen

- Modellierung
  - Netz wird als Graph modelliert. Router und Endsysteme sind die Knoten, die Verbindungen die Kanten des Graphs. Den Kanten werden jeweils Kosten zugewiesen.



- Shortest Path Algorithmen
  - Bellman-Ford
    - Findet kürzesten Pfad von einem Quellknoten zu allen anderen Knoten bzw. den kürzesten Pfad von allen Knoten zu einem Zielknoten
    - Iteration über Kanten
    - Distanz-Vektor-Algorithmus
  - Dijkstra
    - Findet kürzesten Pfad von einem Quellknoten zu allen anderen Knoten bzw. den kürzesten Pfad von allen Knoten zu einem Zielknoten
    - Iteration über Pfadlänge
    - Link-State-Algorithmus

Notation und Variablen

- N: Anzahl Knoten im Netzwerk
- s: Anzahl Iterationsschritte
- h: Pfadlänge
- $D_i^{(h)}$: Kürzeste Pfadlänge von Knoten 1 zu Knoten i
- d_{ij}: Kosten (Distanz) von Knoten i zu Knoten j
- P: Menge schon behandelter Knoten

3.2.1 Bellmann-Ford Algorithmus


[BeGa91]

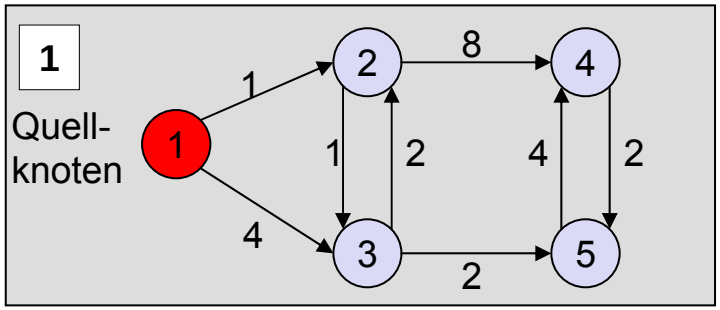
- Funktionsweise
  - Initial: Finde zuerst den kürzesten Pfad mit Kantenlänge 1
  - Dann: Schleife für Kantenlänge 2, 3, 4, …
    - Finde alle kürzesten Pfade, die höchstens entsprechend viele Kanten lang sind
- Komplexität: O(N^3) mit N = Anzahl der Knoten
  - Muss im schlimmsten Fall (N – 1) Mal für (N – 1) Knoten und (N – 1) Alternativen durchgeführt werden
- Algorithmus
  - Sei $D_i^{(h)}$ die kürzeste (≤ h) Pfadlänge von Knoten 1 zu Knoten i Nach Konvention gilt: $D_1^{(h)} = 0$ für alle h
  - Initial:

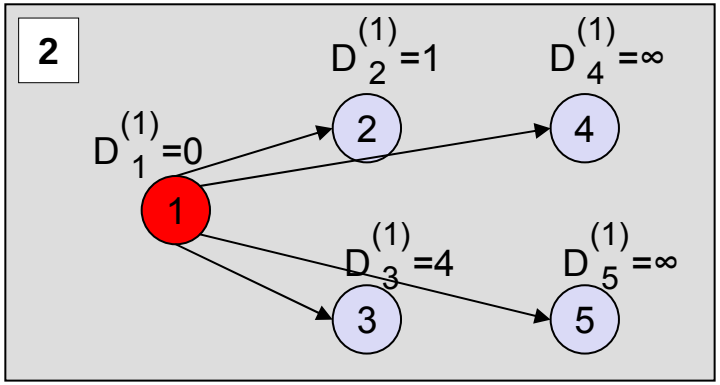
$$D_i^{(0)} = \infty, \forall i \neq 1$$

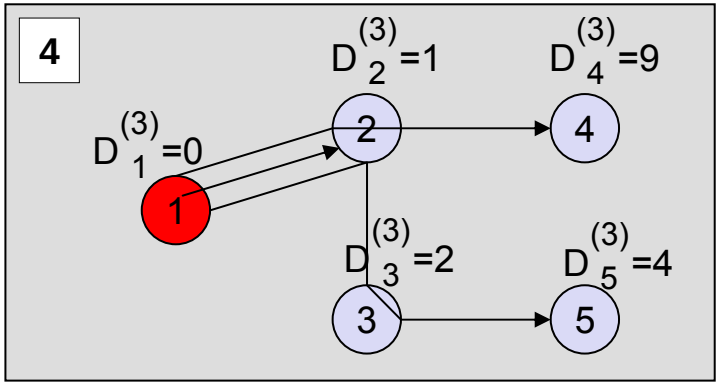
  - Für jede Wiederholung $h \geq 0$:

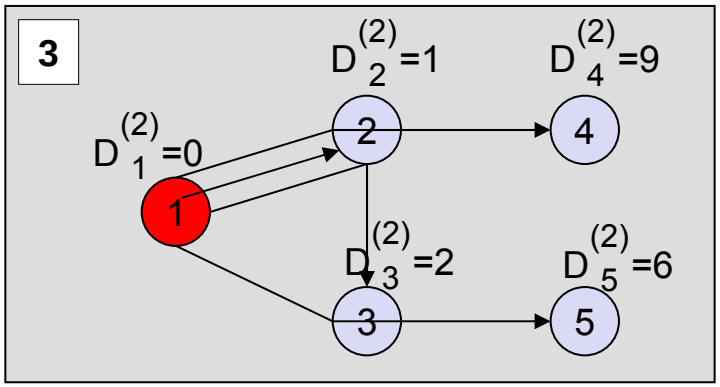
$$D_i^{(h+1)} = \min_j [D_j^{(h)} + d_{ij}], \forall i \neq 1$$

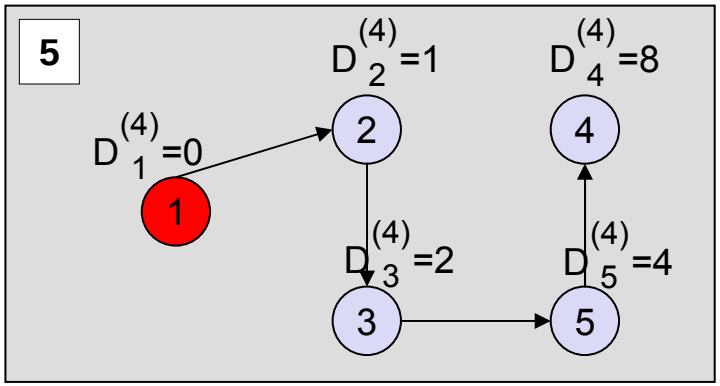
Beispiel: Bellmann-Ford











3.2.2 Dijkstra Algorithmus

- Funktionsweise
  - Finde bei jedem Iterationsschritt den kürzesten Pfad
  - Finde den kürzesten Pfad mit Kantenlänge 1 vom Startknoten
  - Der nächste Pfad ist entweder der nächstkürzeste Pfad der Kantenlänge 1 oder der kürzeste Pfad der Kantenlänge 2
  - usw.
  - Hinweis: Es sind keine negativen Kosten (Distanzen) für Pfade erlaubt

[BeGa91]

- Komplexität: O(s^2) mit s = Anzahl der Iterationen
- Algorithmus:
  - Initial:

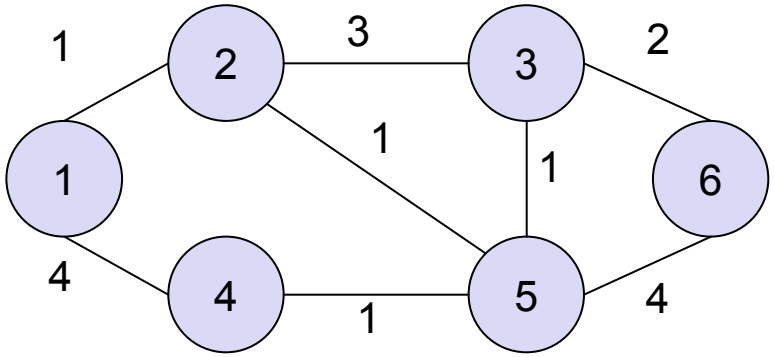
    Sei $P = \{1\}$, $D_1 = 0$ und $D_j = d_{1j}$ für $j \neq 1$
  - 1. Schritt:

    Finde den nächst-nahen Nachbarn, so dass gilt:

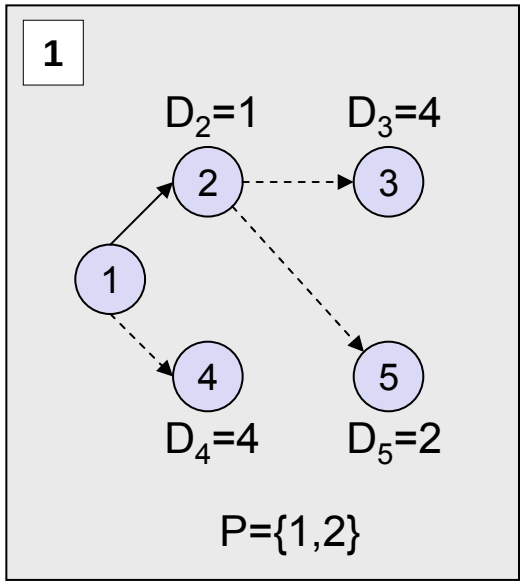
    $$i \notin P; \quad D_i = \min_{j \notin P} D_j; \quad P := P \cup \{i\}$$
  - 2. Schritt:

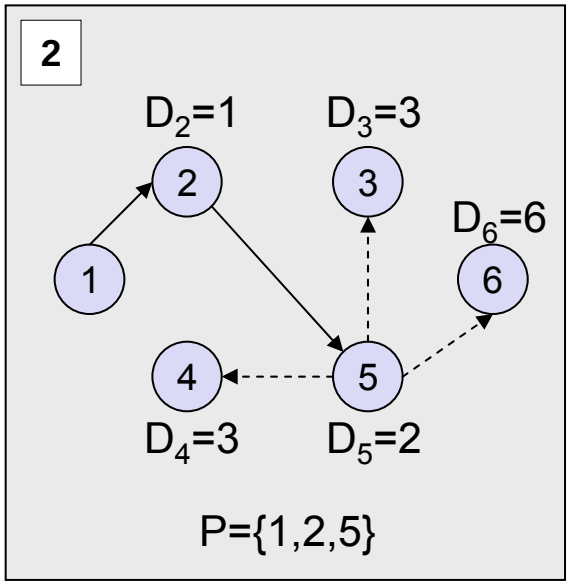
    Berechne $D_j = \min[D_j, D_i + d_{ij}]; \quad \forall j \notin P$
  - Gehe zu Schritt 1

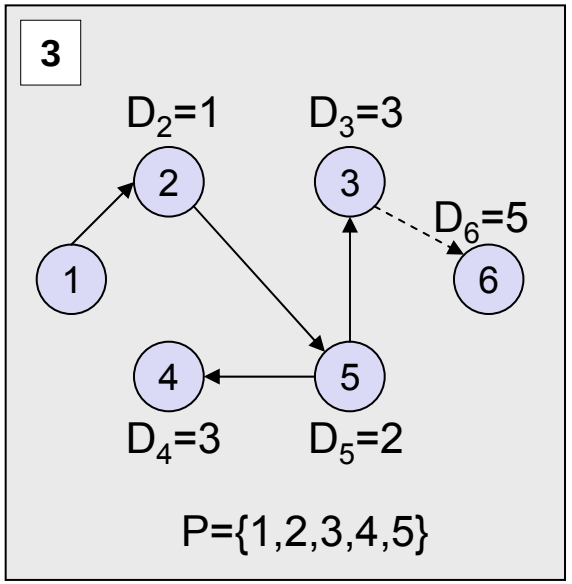
Beispiel: Dijkstra



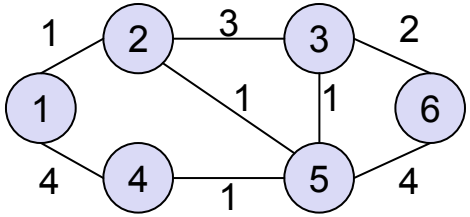
$d_{ij} = d_{ji}$ für alle (i,j)





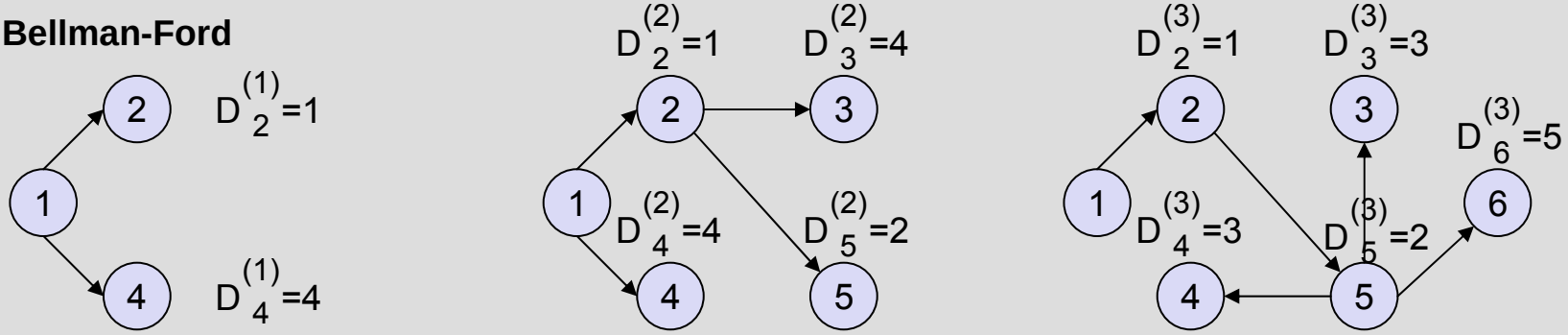


Beispiel: Vergleich Bellmann-Ford und Dijkstra

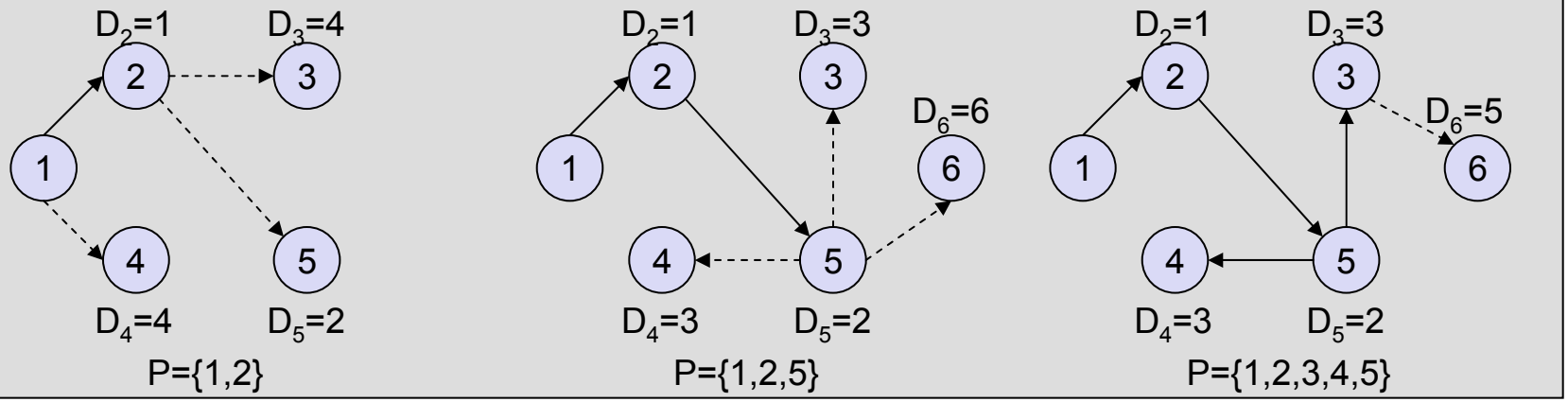


$d_{ij} = d_{ji}$ für alle (i,j)

Bellman-Ford

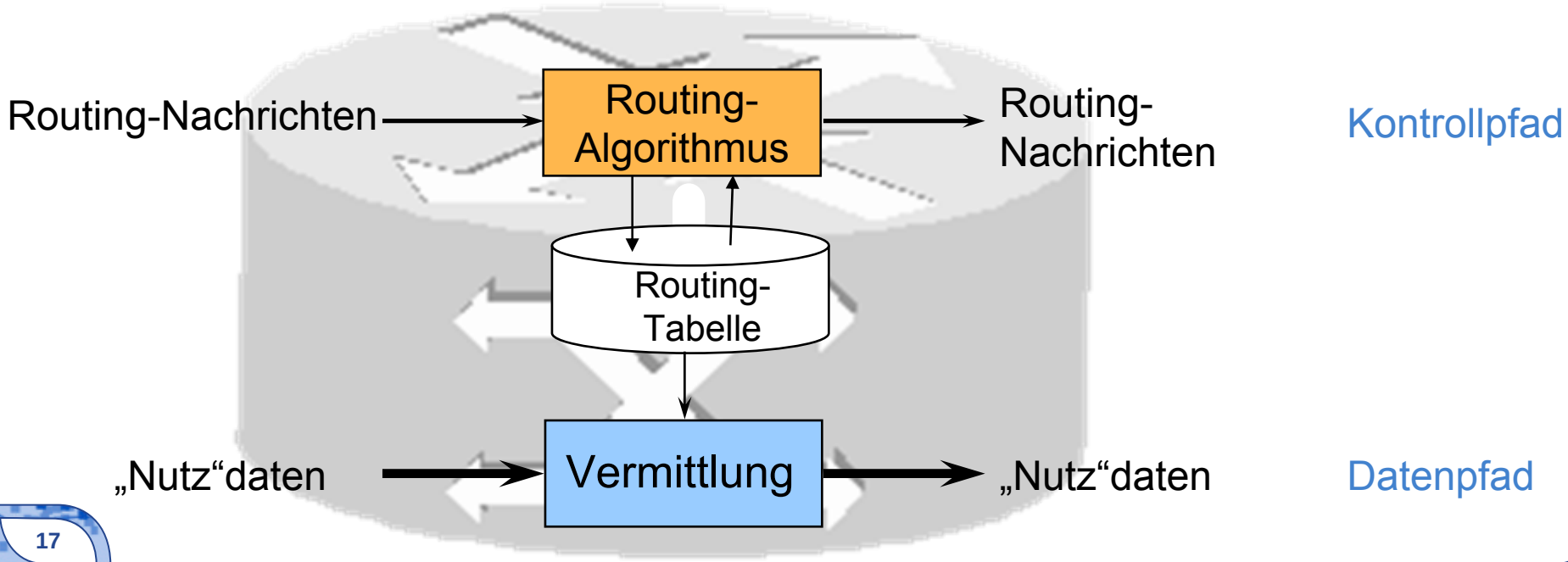


Dijkstra



3.3 Routing im Internet

- Aufbau und Funktionsweise eines Routers
  - Verbindet unterschiedliche Netze auf Schicht 3
  - Unterscheidung zwischen Kontrollpfad und Datenpfad
  - Weiterleitung von Dateneinheiten nach Informationen aus der Routing- (bzw. Forwarding)-Tabelle
    - Routing-Algorithmen bestimmen Inhalt der Routing-Tabelle
    - Implementierungsabhängig: Forwarding-Tabelle mit nur den „besten“ Pfaden der Routing-Tabelle



3.3.1 Historie zum Internet-Routing

- Von Beginn an: verteiltes adaptives Routing
  - Knoten im Netz tauschen Information aus, die dann für die Routing-Entscheidung mit herangezogen wird. Die Routing-Entscheidung wird jeweils lokal in den einzelnen Knoten gefällt
  - ARPANET
    - „Grand-daddy of packet networks“
    - Zu Beginn (70er Jahre) 56 kbit/s Mietleitungen
    - Hat Routing-Architektur des heutigen Internet stark geprägt

[McFR78]

Ziel der Betrachtung der Historie

- Vorstellung der historischen Routing-Ansätze zur Darstellung der
  - gewonnenen Erkenntnisse
  - evolutionären Verbesserungen
  - Gründe, warum welche Lösungen zum Einsatz kamen

3.3.1.1 Erste „Routing-Generation“

- Original ARPANET Routing-Algorithmus
  - Entwickelt 1969
  - Verteilter adaptiver Algorithmus (Shortest Path) basierend auf Bellmann-Ford Algorithmus
    - Minimal geschätzte Verzögerung als Metrik
    - Intention: Wegewahl, auf denen Dateneinheiten möglichst schnell ihr Ziel erreichen
      - Implementierung: Aktuelle Warteschlangenlänge (Anzahl Dateneinheiten in Warteschlange) plus festes Inkrement
      - Problematisch, da immer eine Verzögerung hinzugefügt wird
      - Problematisch bei unterschiedlichen Übertragungskapazitäten (bauen Warteschlange unterschiedlich schnell ab)

[McFR78]

Erste „Routing-Generation“

- Jeder Knoten verwaltet drei Tabellen
  - „Network-Delay“-Tabelle
    - Geschätzte Verzögerung, die eine Dateneinheit erfährt für alle Schnittstellen und alle Ziele
  - „Minimum-Delay“-Tabelle
    - Minimale geschätzte Verzögerung zu allen möglichen Zielen
    - Wird alle 2/3 Sekunden bestimmt. Tabelle wird alle 2/3 Sekunden an alle Nachbarn versendet.
    - Zusätzlich wird noch der „Hop-Count“ versendet
      - Gibt Entfernung in Anzahl von Zwischensystemen an
  - Routing-Tabelle
    - Enthält Einträge für mögliche Ziele mit Schnittstellen, auf denen Dateneinheiten zum nächsten Zwischensystem in Richtung dieser Ziele gesendet werden
- … bis Mai 1979 im ARPANET verwendet

Erste Verbesserungen

- Anpassung von Routen

[McFR78]

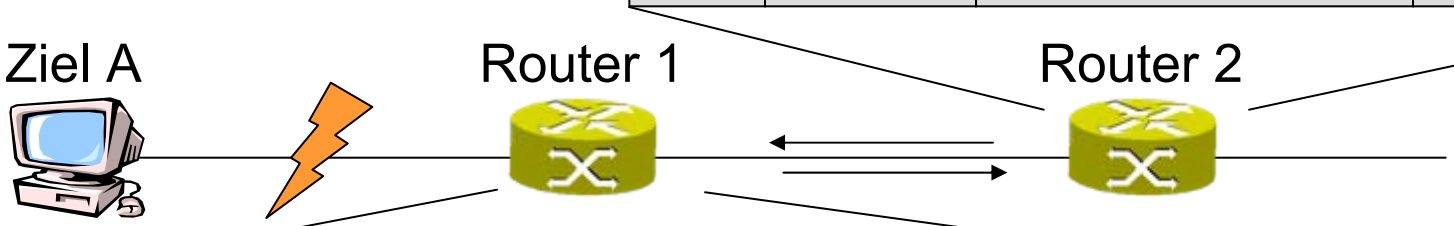
  - *„the reachability algorithm reacts very quickly to „good news“, and very slowly to „bad news“*
    - Grundlegendes Problem bei Systemen mit wiederholter verteilter Minimierung oder Maximierung
  - Beobachtung
    - Hop-Count zu einem Zielknoten unterscheidet sich bei benachbarten Knoten höchstens um Eins
    - Erniedrigen des Hop-Counts um Eins: Alle Knoten verwenden diesen Hop-Count und den entsprechenden Weg
    - Erhöhen des Hop-Counts: Solange Nachbarn niedrigeren Wert haben, wird der erhöhte Wert ignoriert – es gibt alternative Wege

    → Count-to-Infinity

Erste Verbesserungen

- Count-to-Infinity
  - Beobachtung: Schleifen entstehen bis Distanzwert „Unendlich“ erreicht wird
  - Grundlegendes Problem:
    - Router 1 ändert Weg zu einem Ziel, d.h. nutzt neuen nächsten Router (2), ohne zu wissen, dass dessen kürzester Weg zum Ziel über Router 1 selbst verläuft
  - Split Horizon
    - Routing-Informationen zu Zielen werden nicht an Nachbarn gesendet, über die auch die Wege zu diesen Zielen laufen

Ziel	Distanz	Nächstes System	Schnittstelle
A	2	Router 1	1
...	...	...	...



Ziel	Distanz	Nächstes System	Schnittstelle
A	3	Router 2	1
...	...	...	...

Erste Verbesserungen

- Poisoned Reverse: Vermeiden der Schleifenbildung
  - Sendet Knoten K neue Routing-Informationen zu einem Nachbarn, so wird die Routing-Metrik solcher Routen, die über diesen Nachbarn laufen mit „Unendlich“ markiert

→ Schleifen, die mehr als zwei Knoten involvieren können nicht vermieden werden

  - Überlegen Sie sich ein Beispiel

- Hold-down
  - Routing-Algorithmus nutzt beste Route für einen gewissen Zeitraum weiterhin, auch wenn sich diese verschlechtert hat

  → kann manche Schleifen vermeiden

  → Verlangsamt Anpassung an Topologieänderungen
    - Wieso ist dies der Fall?

Erste Verbesserungen

- Erhöhte Verfügbarkeit (Anteil der Zeit, zu der das Netzwerk erfolgreich seinen Übertragungsdienst erbringt)
  - Extrem wichtig, denn ohne Routing kein funktionierendes Netz
  - Lokaler Fehler kann globale Auswirkungen haben
    - Beispiel: ein Knoten signalisiert, dass er auf dem besten Weg zu allen anderen Knoten liegt
  - Aufgetretene Probleme im ARPANET u.a.:
    - Speicherprobleme im Router. „Minimum-Delay“-Tabelle enthält nur Nullen. Router hat also eine Verzögerung von Null zu allen Knoten, stellt also besten Weg zu allen Knoten bereit.
    - Speicherproblem verursacht inkorrekte Anweisung. Falscher Zeiger wurde berechnet und damit die Wegewahlentscheidung auf der Basis zufälliger Daten im Speicher getroffen. Stark oszillierende Routen.
  - Abhilfe, u.a. durch
    - Nutzung von Prüfsummen
      - Spezielle Prüfsummen zur Absicherung der Routing-Information
      - Programme zur Berechnung von Routen mit Prüfsummen versehen

Versenden der „Minimum-Delay“-Tabelle

- Probleme

[McRR80]

  - Routing-Nachrichten beeinflussen den Datenverkehr
    - Alle Knoten verteilen periodisch Routing-Nachrichten
    - Routing-Nachricht enthält jeweils die komplette „Minimum-Delay“-Tabelle des sendenden Knotens. Skaliert nicht, denn
      - Größe der Routing-Nachricht wächst mit Anzahl der Knoten
      - Anzahl der sendenden Knoten wächst
  - Berechnung der besten Routen erfolgt verteilt, basierend auf den Informationen der anderen Knoten und den eigenen Messungen
    - schwer zu kontrollieren, dass alle Knoten die gleichen besten Routen verwenden. Unterschiedliche Knoten haben verschiedene Sichten auf das Netz.
  - Frequenz, mit der Informationen gesendet werden, beeinflusst Reaktionszeit des Netzes. Dilemma:
    - Gering: Zu langsam, um auf Stausituationen und wichtige Topologieänderungen zu reagieren
    - Hoch: Zu schnelle Reaktion auf weniger wichtige Änderungen

Verzögerungsmessungen

- Periodische Überprüfung der Anzahl Dateneinheiten in der Warteschlange in jedem Knoten
  - Anzahl der Dateneinheiten entspricht Länge der Warteschlange
  - Feste Konstante dazu addiert
    - Durchlaufen eines Routers benötigt auch ohne Warteschlangen Zeit
- Probleme
  - Verzögerung hängt auch von der Datenrate des Netzanschlusses und von der Länge der Dateneinheiten ab
  - Ressourcen des Knotens haben ebenfalls Einfluss auf Verzögerung
    - Z.B. Prozess ist durch anderen Prozess blockiert

→ Instantane Messung nicht geeignet zur Vorhersage von mittlerer Verzögerung. Evtl. sehr starke Schwankungen der Warteschlangenlänge und damit häufige Wechsel der Routen

3.3.1.2 Zweite „Routing-Generation“

- Neuer Routing-Algorithmus, lief ab 1979 im ARPANET. Shortest Path First Algorithmus basierend auf Dijkstra.
  - Jeder Knoten kennt Netztopologie und die Verzögerungen zwischen den Knoten
    - Alle Knoten führen Berechnungen auf gleicher Datenbasis durch
    - Keine permanenten Schleifen. Temporäre Schleifen können existieren.
  - Periodische Messung der Verzögerungen auf den lokalen Netzanschlüssen
    - Weiterleitung dieser Information nicht nur an die Nachbarn sondern an alle Knoten im Netz
      - Fluten an alle Knoten
      - Schnelle Ausbreitung, da die Information unverändert weitergeleitet wird
    - Routing-Nachrichten relativ klein und Größe unabhängig von der Anzahl der Knoten im Netz
    - Schnelle Ausbreitung der Information, konsistentes Routing aller Knoten

[MCRR80, McRR80a]

Mittlere Verzögerung

- Periodische Messung der Verzögerung
  - Verzögerungen werden gemessen. Aufenthaltszeit, Sendezeit und Ausbreitungsverzögerung werden berücksichtigt
    - Dateneinheit erhält Zeitstempel mit Eingangszeit T_{Ein} im Knoten
    - Ausgangszeit T_{Aus} zeigt Aussenden des ersten Bits der Dateneinheit an.
    - Bei Empfang der Quittung:
      - Differenz zwischen Eingangszeit und Ausgangszeit der Dateneinheit („Aufenthaltszeit“).
      - Dazu addiert: Ausbreitungsverzögerung t_a und Sendezeit t_s.

      → Verzögerung t_v der Dateneinheit.

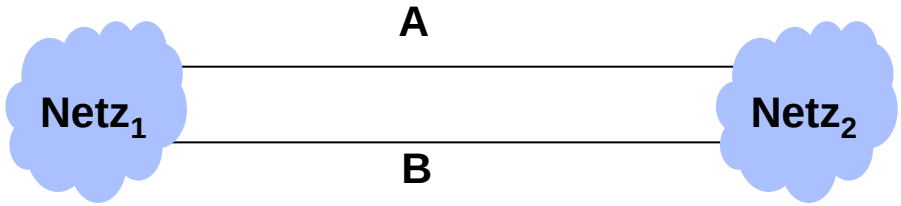
$$t_v = (T_{Aus} - T_{Ein}) + t_a + t_s$$

Mittlere Verzögerung

- Mittlere Verzögerung wird alle 10 Sekunden berechnet
  - Bei erheblichen Änderungen werden alle Knoten im Netz informiert
    - Schnelle Reaktion auf große Änderungen
  - Kleinere Änderungen werden nicht sofort im Netz propagiert
    - Dennoch müssen längerfristige, langsame Änderungen verbreitet werden
  - Variabler Schwellenwert zur Bestimmung „erheblicher“ Änderungen
    - Wird mit der Zeit reduziert
    - Initial auf 64 ms gesetzt. Bei jeder Berechnung (alle 10s) der mittleren Verzögerung um 12,8 ms erniedrigt
    - Nach 1 Minute wird die aktuell gemessene Verzögerung auf jeden Fall propagiert, da der Schwellenwert dann Null beträgt

Problem: Verzögerungs-Metrik

- Annahme
  - Gemessene Verzögerung ist ein guter Indikator der Verzögerung, die Dateneinheiten nach dem Re-routing erfahren
- Problem
  - Funktioniert gut bei niedriger bis moderater Verkehrslast
    - Verzögerung in Warteschlangen nicht dominant
  - Trifft bei hoher Verkehrslast so nicht zu und führt zu Routing-Oszillationen bzw. schlecht ausgelasteten Netzen
- Szenario



  - Von $Netz_1$ nach $Netz_2$ wird entweder A oder B verwendet. A und B seien identisch dimensioniert. Alle Knoten in $Netz_1$ bzw. $Netz_2$ ändern ihre Routen simultan.
  - Der Link A sei aktiv und ausgelastet → hohe Verzögerung in der Warteschlange → Routing-Änderung zu B → hohe Verzögerung …

[KhZi89]

3.3.1.3 Dritte „Routing-Generation“

- Änderung der Routing-Metrik 1987
  - Routing-Algorithmus selbst nicht geändert
  - Bestimmung des Pfades adaptiert
    - Grundlegende Idee: Bei hoher Auslastung wird nicht der optimale Pfad für alle Routen berechnet, sondern die Routen sollen einen „guten“ Pfad bereitgestellt bekommen
    - Auslastung wird neben Verzögerung mit betrachtet
      - Berücksichtigt damit auch unterschiedliche Datenraten auf verschiedenen Links
    - Damit: höhere Auslastung und stabilere Routen erzielbar

[KhZi89, Stal06]

Geänderte Routing-Metrik

- Berechnung:
  1) Gemessene durchschnittliche Verzögerung über die letzten 10 Sekunden wird in eine Schätzung der Linkauslastung umgewandelt. (M/M/1-Warteschlange)

$$p = \frac{2(T_B - T)}{T_B - 2T}$$

p – Link-Auslastung

T_B – Bedienzeit (gemittelt durch: 600 bit durchschnittliche Größe einer Dateneinheit / Datenrate des Links)

T – Gemessene Verzögerung

  2) Das Ergebnis wird geglättet, indem vorherige Schätzungen mit einbezogen werden:

$$U(n+1) = 0{,}5 * p(n+1) + 0{,}5 * U(n)$$

$U(n)$ – Durchschnittliche Auslastung zum Sample-Zeitpunkt n

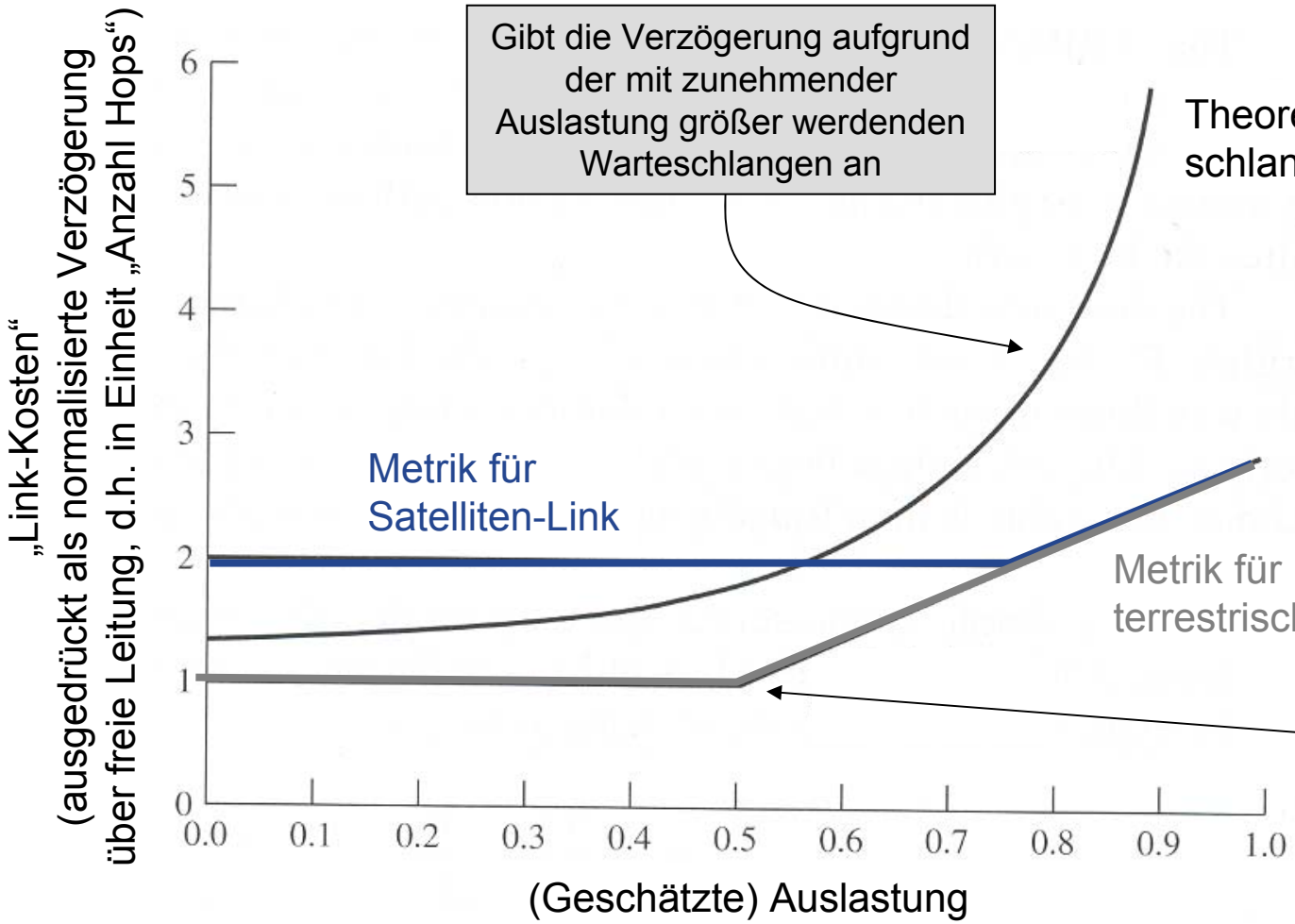
$p(n)$ – Link-Auslastung gemessen zum Sample-Zeitpunkt n

  3) Die Link-Kosten werden als eine Funktion der durchschnittlichen Auslastung derart dargestellt, dass eine Oszillation der Routen verhindert wird
  Hierzu werden die absolut möglichen Link-Kosten, sowie deren Änderungen zwischen zwei Routing-Updates begrenzt (siehe Abbildung übernächste Folie)

Geänderte Routing-Metrik

- Vorteil:
  - Routing reagiert bei schwach ausgelasteten Links besser auf Verzögerungen durch Ausbreitungsverzögerung, Warteschlangen und Übertragung
  - Routing wird unempfindlich bzgl. Ausbreitungs- und Warteschlangenverzögerung bei stark belasteten Links

Geänderte Routing-Metrik



3.3.1.4 Zusammenfassung

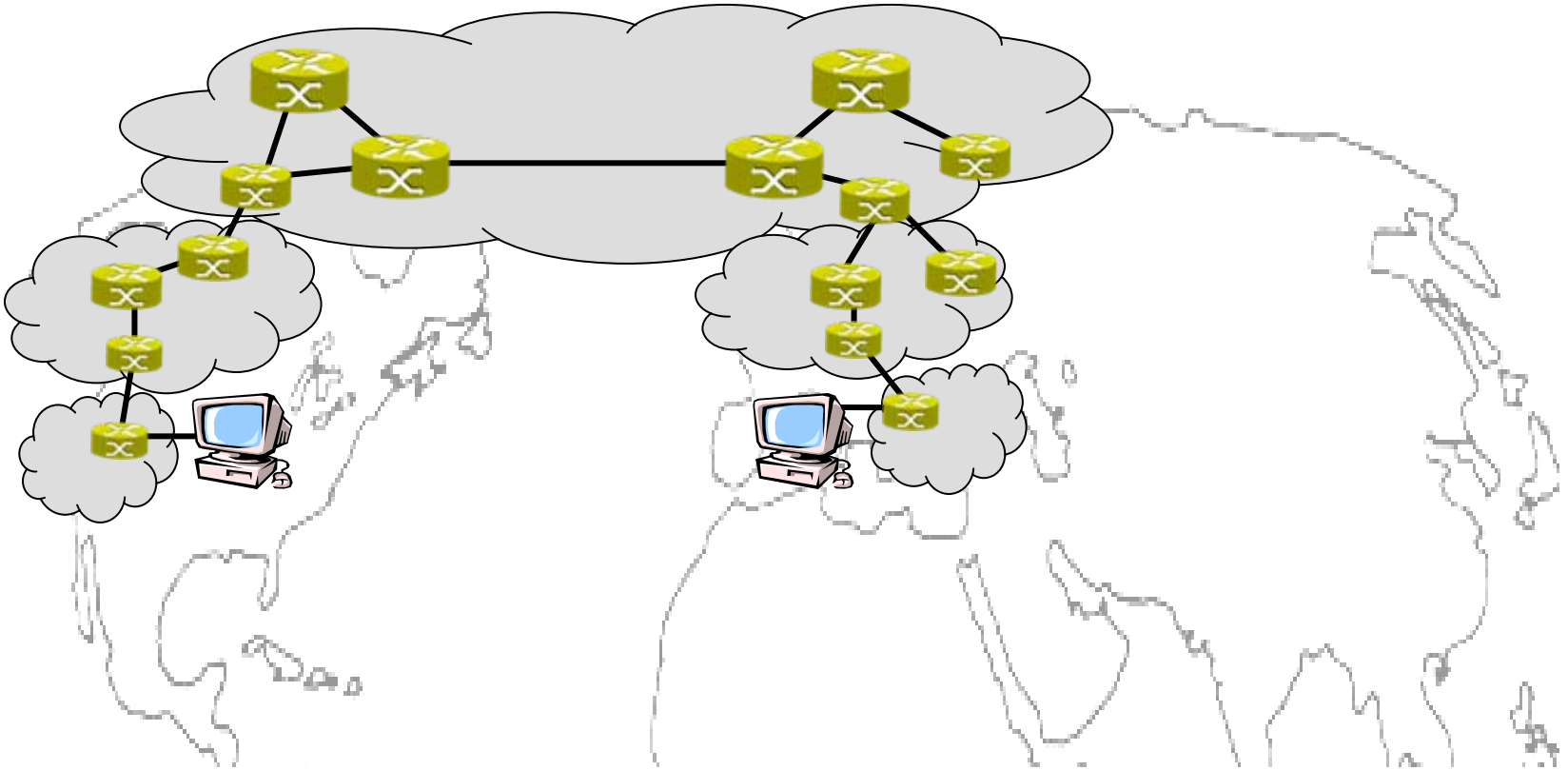
- Erste „Routing-Generation“
  - 1969: Die ersten Knoten von ARPANET sind funktionsfähig
  - Verteilter adaptiver Algorithmus, basierend auf Bellman-Ford
- Zweite „Routing-Generation“
  - 1979: ARPANET hat 200 Knoten
  - Verteilter adaptiver Algorithmus, basierend auf Dijkstra
    - Bessere Skalierbarkeit, keine langlebigen Schleifen, stabilere Routen
- Dritte „Routing-Generation“
  - 1987: ARPANET wächst weiter
  - Verbesserungen der Routing-Metrik
    - Verbessertes Verhalten bei hoher Last
    - Nicht mehr nur auf Verzögerung basierend. Berücksichtigt auch die Auslastung eines Links

3.3.2 Autonome Systeme

- Ursprüngliches Internet
  - ARPANET-Backbone wurde von BBN administriert und betrieben
- Das globale Internet besteht heute aus einer Menge von Autonomen Systemen (Autonomous Systems, AS)
  - Bessere Verwaltbarkeit des Internet
    - Ohne Autonome Systeme würde das Internet zum Beispiel aus einem großen Netz mit einem einheitlichen Routing-Protokoll bestehen
  - Jedes Autonome System hat eine eindeutige Nummer (derzeit 16 bit, Erweiterung auf 32 bit geplant) und besteht aus einer Ansammlung von Routern, die
    - eine gemeinsame Routing-Policy verfolgen
    - unter der gleichen technischen Administration liegen
    - ein gemeinsames Routing-Protokoll verwenden
    - nach außen wie eine Einheit erscheinen
  - Zwei grundsätzlich unterschiedliche Arten von Routing-Protokollen existieren:
    - Interior Gateway Protocol (IGP) innerhalb eines Autonomen Systems
    - Exterior Gateway Protocol (EGP) zwischen Autonomen Systemen

Aufteilung in Autonome Systeme

- Anschauliches Beispiel



Routing-Beispiel (traceroute)

Traceroute aus dem Netz der Uni Karlsruhe zu einem Host in den USA:

```
$ traceroute 66.35.250.151
traceroute to 66.35.250.151 (66.35.250.151), 30 hops max, 60 byte packets
 1  i72marbgate.tm.uni-karlsruhe.de (141.3.71.126)  0.429 ms  0.391 ms  0.366 ms
 2  172.16.4.1 (172.16.4.1)  0.796 ms  1.388 ms  1.691 ms
 3  192.168.1.190 (192.168.1.190)  2.406 ms  2.623 ms  4.596 ms
 4  172.21.3.9 (172.21.3.9)  4.811 ms  5.027 ms  5.109 ms
 5  Karlsruhe1.belwue.de (129.143.166.141)  5.325 ms  5.788 ms  5.974 ms
 6  Frankfurt1.belwue.de (129.143.1.178)  7.951 ms  5.024 ms  7.761 ms
 7  ffm-b2-link.telia.net (213.248.88.25)  7.949 ms  6.098 ms  6.060 ms
 8  ffm-bb2-link.telia.net (80.91.252.175)  6.222 ms ffm-bb1-link.telia.net
   (80.91.252.173)  6.688 ms ffm-bb1-link.telia.net (80.91.249.100)  6.408 ms
 9  prs-bb1-link.telia.net (80.91.247.36)  15.887 ms prs-bb2-pos7-0-0.telia.net
   (213.248.65.117)  16.094 ms prs-bb2-link.telia.net (80.91.252.233)  15.853 ms
10  ash-bb2-link.telia.net (80.91.251.102)  103.979 ms ash-bb2-link.telia.net
   (80.91.254.214)  103.668 ms ash-bb2-link.telia.net (80.91.251.102)  104.131 ms
11  208.173.52.121 (208.173.52.121)  99.603 ms  104.321 ms  99.292 ms
12  cr1-tengig0-7-2-0.washington.savvis.net (204.70.197.242)  104.268 ms  107.727 ms
   103.822 ms
13  cr1-pos0-0-0.sanfrancisco.savvis.net (204.70.192.114)  178.741 ms  178.725 ms
   182.397 ms
14  er1-7-0-0.SanJoseEquinix.savvis.net (204.70.200.197)  272.311 ms  209.762 ms
   209.953 ms
15  hr1-te-1-0-0.santaclarasc8.savvis.net (204.70.200.214)
   179.715 ms
16  204.70.203.142 (204.70.203.142)  184.144 ms !H * *
```

3 Autonome Systeme

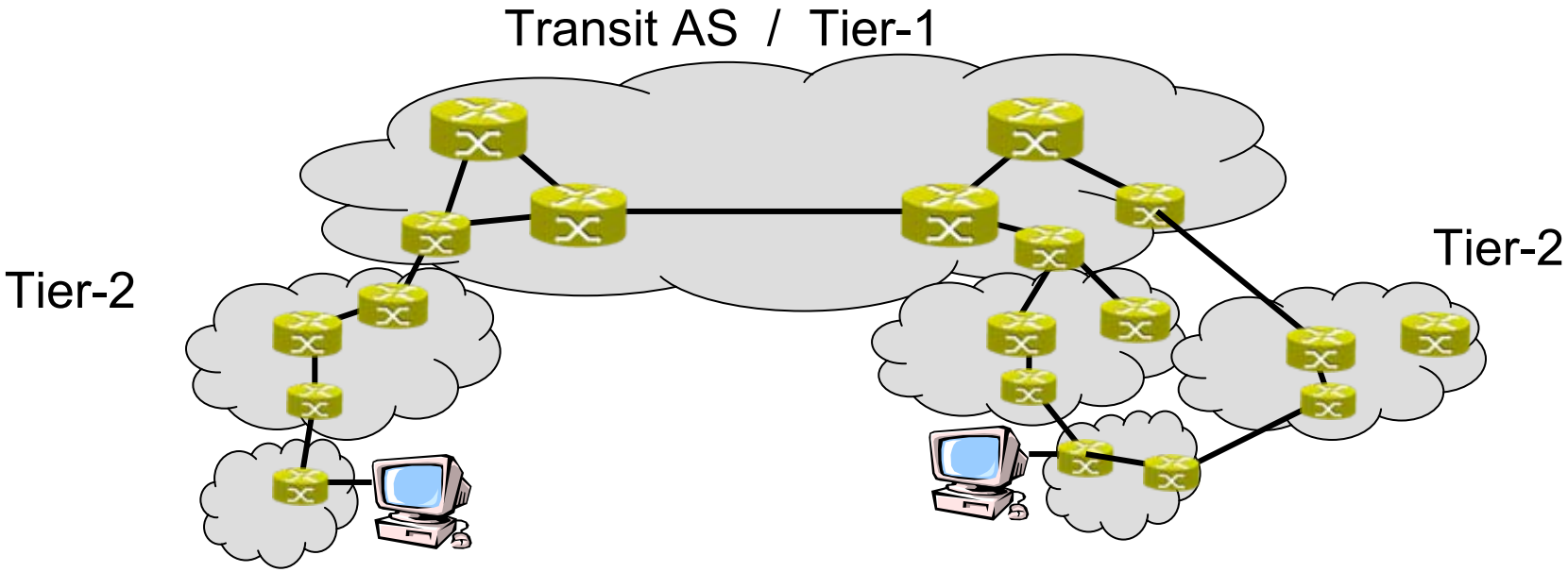
Ø zu bel. Ziel (!): 2-3 ASe

Einteilung Autonomer Systeme

- Verschiedene Typen von Autonomen Systemen lassen sich unterscheiden:
  - Stub-AS
    - Kleine Unternehmen
      - meist nur regional tätig
    - Anschluss an genau einen Provider
    - Kein Transitverkehr
  - Multihomed Stub-AS
    - Große Unternehmen
    - Anschluss an mehrere Provider (Ausfallsicherheit)
    - Kein Transitverkehr
  - Transit AS
    - Provider
      - üblicherweise globaler Ausdehnung

Einteilung Autonomer Systeme

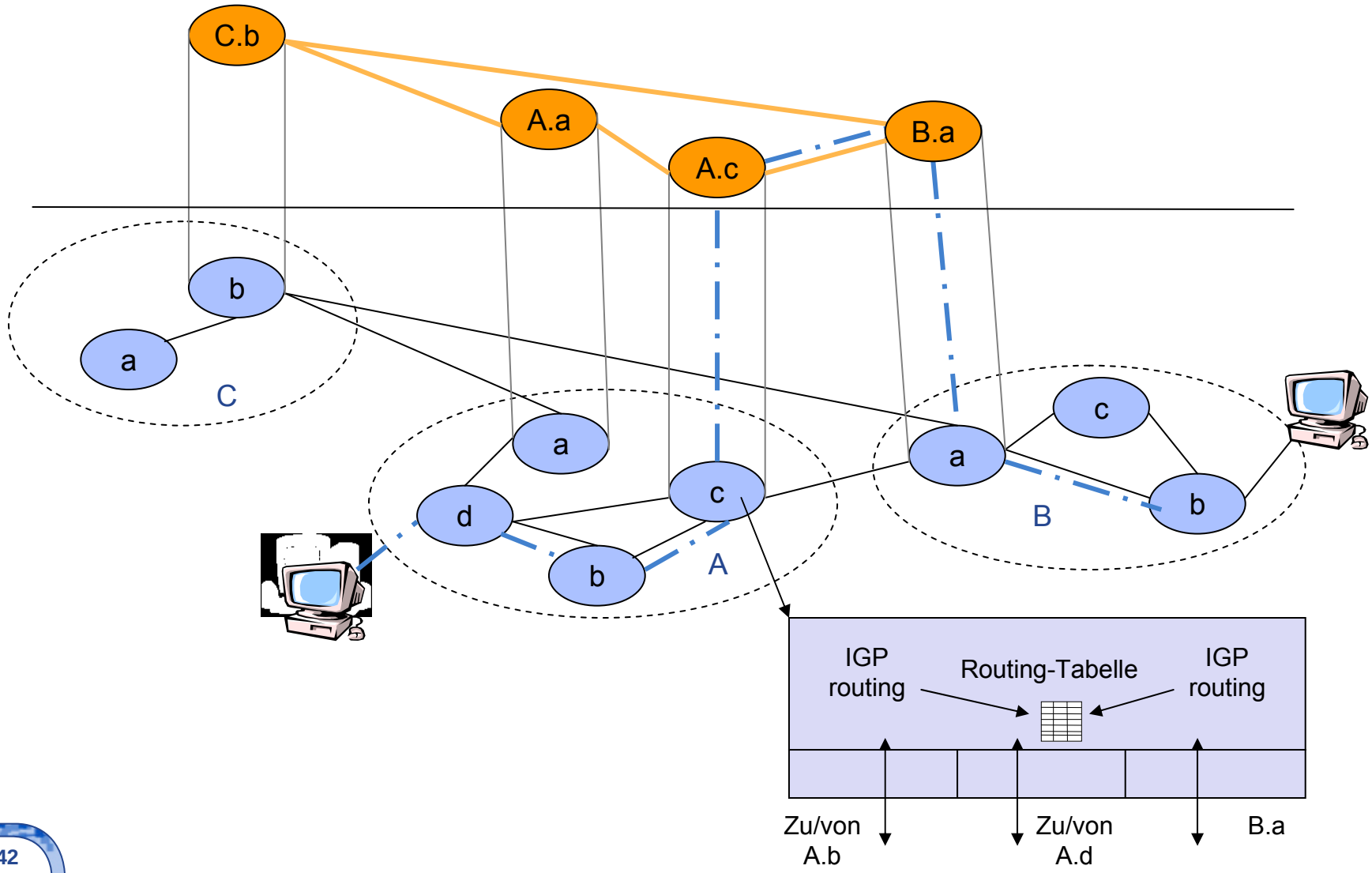
- Verschiedene Klassen Autonomer Systeme
  - je nach „wirtschaftlicher Position/Bedeutung“ (Tier-x)
  - nach Funktion (Transit / Stub)



Stub-AS
(„Verbraucher“, Provider, kein Transit)

Multihomed-Stub-AS
(Anschluss an mehrere Provider)

Routing mit IGPs und EGPs



Weshalb IGPs und EGPs?

- Policy
  - Politische Frage: welcher Transit-Verkehr darf das Autonome System passieren?
  - EGP: Policies werden vom Provider ausgewählt
  - IGP: eine Organisation, wenig Policies erforderlich
- Skalierbarkeit
  - EGP: weitere Ebene der Abstraktion; Größen der Routing-Tabellen und Anzahl der Updates können reduziert werden, da Ausfälle innerhalb eines Autonomen Systems meist verborgen bleiben können
  - IGP: bessere Stabilität
- Leistungsfähigkeit
  - EGP: Policies sind erforderlich und wichtiger als Leistungs-Metriken
  - IGP: Konzentration auf Leistungs-Metriken

3.3.3 Interior Gateway Protocols

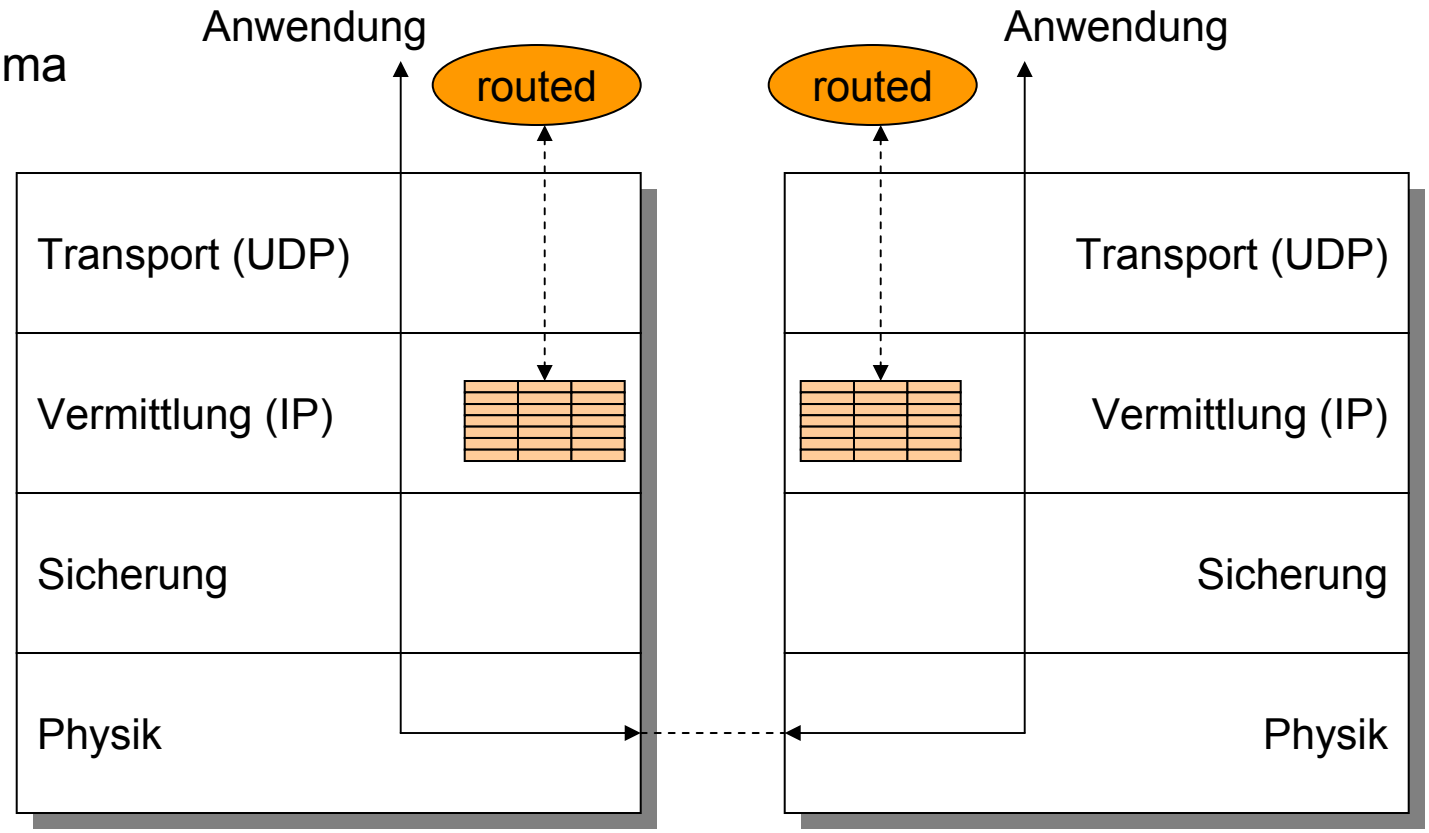
- Im Internet werden heute die folgenden IGPs eingesetzt: [Huit00]
  - RIP (Routing Information Protocol)
    - Distanz-Vektor-Algorithmus verwendet
  - OSPF (Open Shortest Path First)
    - Link-State-Algorithmus verwendet
  - werden beide im Folgenden besprochen
- Außerdem noch im Einsatz sind u.a.:
  - IS-IS (Intra-Domain Intermediate System to Intermediate System Routing Protocol) [HoWa06]
    - Ursprung als ISO-Standard 10589
    - Link-State-Algorithmus verwendet
    - Für IP eingesetzt bei großen Providern
  - EIGRP (Enhanced Interior Gateway Routing Protocol) [Cisc05]
    - CISCO proprietär
    - Distanz-Vektor-Algorithmus verwendet
    - „Weiterentwicklung“ von RIP

3.3.3.1 Das Routing Information Protokoll (RIP)

- Eines der ersten Routingprotokolle im Internet [BeGa91, Kuro07]
  - Wurde in BSD-Unix implementiert und als „routed“ (route management deamon) installiert
- Sehr einfaches Protokoll das wenig Konfigurationsaufwand erfordert
- Dokumentiert in
  - RFC 1058 im Juni 1988 (RIPv1 – Version 1) [Hedr88]
  - Verbesserungsvorschläge waren bereits enthalten
    - Adressieren die generellen Probleme von Distanz-Vektor-Algorithmen: Schleifen, count-to-infinity
    - Vorschläge: split horizon etc.
  - RFC 2453 (RIPv2 – Version 2) [Malk98]

Implementierung von RIP

- Schema



- Routing-Tabellen werden von einem Anwendungsprozess verwaltet
- RIP-Routing-Nachrichten werden über UDP versendet
  - Nicht zuverlässig
  - Periodisches Versenden (Advertisements)

Funktionsweise von RIP

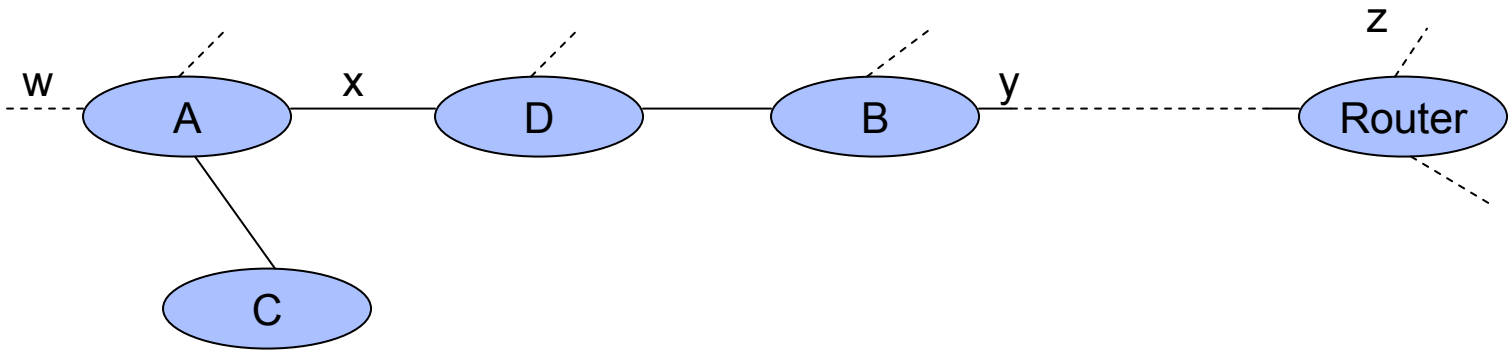
- Verwendete Routing-Metrik
  - Distanz entspricht der Anzahl Hops auf dem Weg zum Ziel. Distanz repräsentiert den Hop-Count
  - Eingeschränkter Wertebereich: 1-15
    Wert 16 entspricht „Unendlich“
    - Größerer Wertebereich wäre z.B. bei Count-to-Infinitiy problematisch
- Transport von Routing-Nachrichten
  - Verwendung von UDP und IP
    - Sowohl in Netzen mit Broadcast-Fähigkeit als auch ohne diese

Funktionsweise von RIP

- Routing-Nachrichten werden entsprechend der folgenden Regeln versendet:
  - Periodisch, normalerweise alle 30 Sekunden
    - Wurde ein Eintrag für 180 Sekunden oder mehr nicht aufgefrischt, so wird der Wert auf „Unendlich“ gesetzt und der Eintrag damit ungültig
  - Wenn sich eine Route geändert hat (als „Triggered Update“ bezeichnet)
    - Verwendung einer Ratenlimitierung – zufällige Verzögerung zwischen 1 und 5 Sekunden – um Netzlast zu reduzieren
- Vorgehensweise bei eingehenden Routing-Nachrichten
  - Eintrag noch nicht vorhanden und Metrik nicht „Unendlich“
    ➔ Neuer Eintrag in Routing-Tabelle
  - Eintrag mit größerer Metrik vorhanden
    ➔ Eintrag entsprechend modifizieren
  - In allen anderen Fällen wird die Routing-Nachricht ignoriert

Beispiel zur Funktionsweise von RIP

- Beispielszenario
  - Verbindungslinien repräsentieren jeweils komplette Netze



- Routing-Tabelle für Router D

Zielnetz	Nächster Router	Anzahl Hops
W	A	2
Y	B	2
Z	B	7
X	-	1
...	...	...

Beispiel zur Funktionsweise von RIP

- Neue RIP-Information von A wird 30 Sekunden später von D empfangen

Zielnetz	Nächster Router	Anzahl Hops
Z	C	4
W	-	1
X	-	1
...	...	...
...	...	...

- Neue Routing-Tabelle in D

Zielnetz	Nächster Router	Anzahl Hops
W	A	2
Y	B	2
Z	A	5
...	...	...
...	...	...

Verbesserungen in RIPv2

- Trotz der bekannten grundsätzlichen Einschränkungen wurde an einer verbesserten Version von RIP gearbeitet. U.a. die folgenden Verbesserungen wurden vorgenommen:

 [Malk98, BaAt97]

- Authentisierung
  - In RIPv1 kann jeder einfach Routing-Nachrichten senden und wird damit als Router wahrgenommen
    - Sehr problematisch beim Versenden „falscher“ Routinginformation (z.B. Null-Werten)
  - In RIPv2 ist eine Authentisierung vorgesehen
    - Zunächst nur als einfaches Passwort (RFC 2453)
    - Dann auch mit MD5 verschlüsseltem Passwort (RFC 2082)
- Einsatz von Multicast
  - RIPv1 versendet Routing-Nachrichten via Broadcast. Sie werden auch von allen Endsystemen empfangen
  - RIPv2 versendet Routing-Nachrichten via Multicast. Hier werden nur die Router angesprochen
- … die grundsätzlichen Beschränkungen von Distanz-Vektor-Algorithmen bleiben erhalten!
- Aufgrund seiner Nachteile wird RIP heute kaum noch praktisch eingesetzt
  - (RIP RIP)

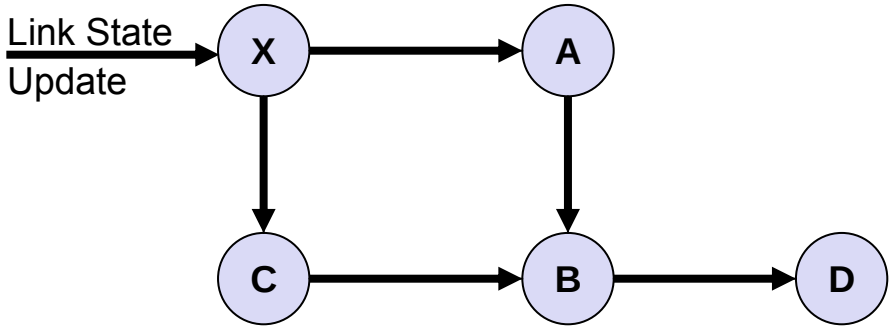
3.3.3.2 Das „Open Shortest Path First“ Routing-Protokoll


[Kuro07, Moy98]

- OSPF ist ein Link-State-Protokoll (RFC 2328)
- Es bietet dedizierte Unterstützung für
  - die Trennung von Endsystem und Netzwerk
    - Subnetzmaske gibt Aufschluss, ob es sich um ein Endsystem oder ein Subnetz handelt
  - sog. Broadcast-Netzwerke (z.B. Ethernet)
  - sog. Nicht-Broadcast-Netzwerke (z.B. X.25)
  - die Aufteilung von großen Netzwerken in kleinere Bereiche (Areas)
- OSPF verwendet sog. „Hello“-Nachrichten um seine Nachbarn zu entdecken
- Routing-Nachrichten werden mittels Multicast an die anderen OSPF-Router versendet
  - Eine Routing-Information, die neu ist, wird an 224.0.0.5 gesendet
  - Die Änderung einer Routing-Information wird an 224.0.0.6 gesendet

OSPF-Fluten

- Flooding-Algorithmus
  - Empfangen einer Dateneinheit. Überprüfen des Eintrags in der Datenbasis mit der Sequenznummer der Dateneinheit
    - falls Eintrag noch nicht vorhanden, dann Hinzufügen und Multicasten (an die Gruppe „AllSPFRouters“ und mit Reichweite 1 Hop)
    - falls Eintrag vorhanden und neue Sequenznummer größer als alte, dann Überschreiben und Multicasten
    - falls Eintrag vorhanden und neue Sequenznummer niedriger als alte, dann keine Änderung
    - falls Sequenznummern gleich sind, wird nichts unternommen
- Wichtig: Routing-Informationen müssen zuverlässig ausgetauscht werden
  - Hierzu verfügt OSPF über Quittungen pro Übertragungsabschnitt, Zeitgeber, Prüfsumme, ...

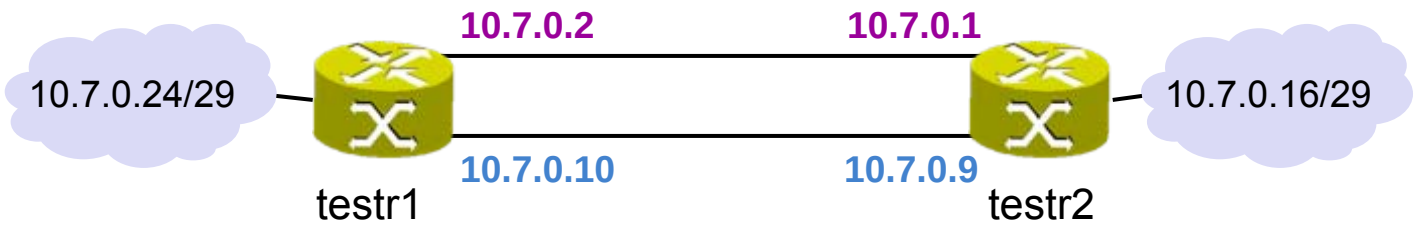


OSPF-Routing-Tabelle

```
testr1#sh ip route ospf
Protocol/Route type codes:
  I1- ISIS level 1, I2- ISIS level2,
  I- route type intra, IA- route type inter, E- route type external,
  i- metric type internal, e- metric type external,
  O- OSPF, E1- external type 1, E2- external type2,
  N1- NSSA external type1, N2- NSSA external type2

  Prefix/Length     Type      Next Hop        Dist/Met            Intf
----------------- ------- --------------- -------------- --------------------
10.7.0.16/29       O-I     10.7.0.1        110/2          FastEthernet1/2
                           10.7.0.9        110/2          FastEthernet1/3
```

```
testr2#sh ip route ospf
Protocol/Route type codes:
  ....
  Prefix/Length     Type      Next Hop        Dist/Met            Intf
----------------- ------- --------------- -------------- --------------------
10.7.0.24/29       O-I     10.7.0.2        110/6          FastEthernet1/1
                           10.7.0.10       110/6          FastEthernet1/2
```

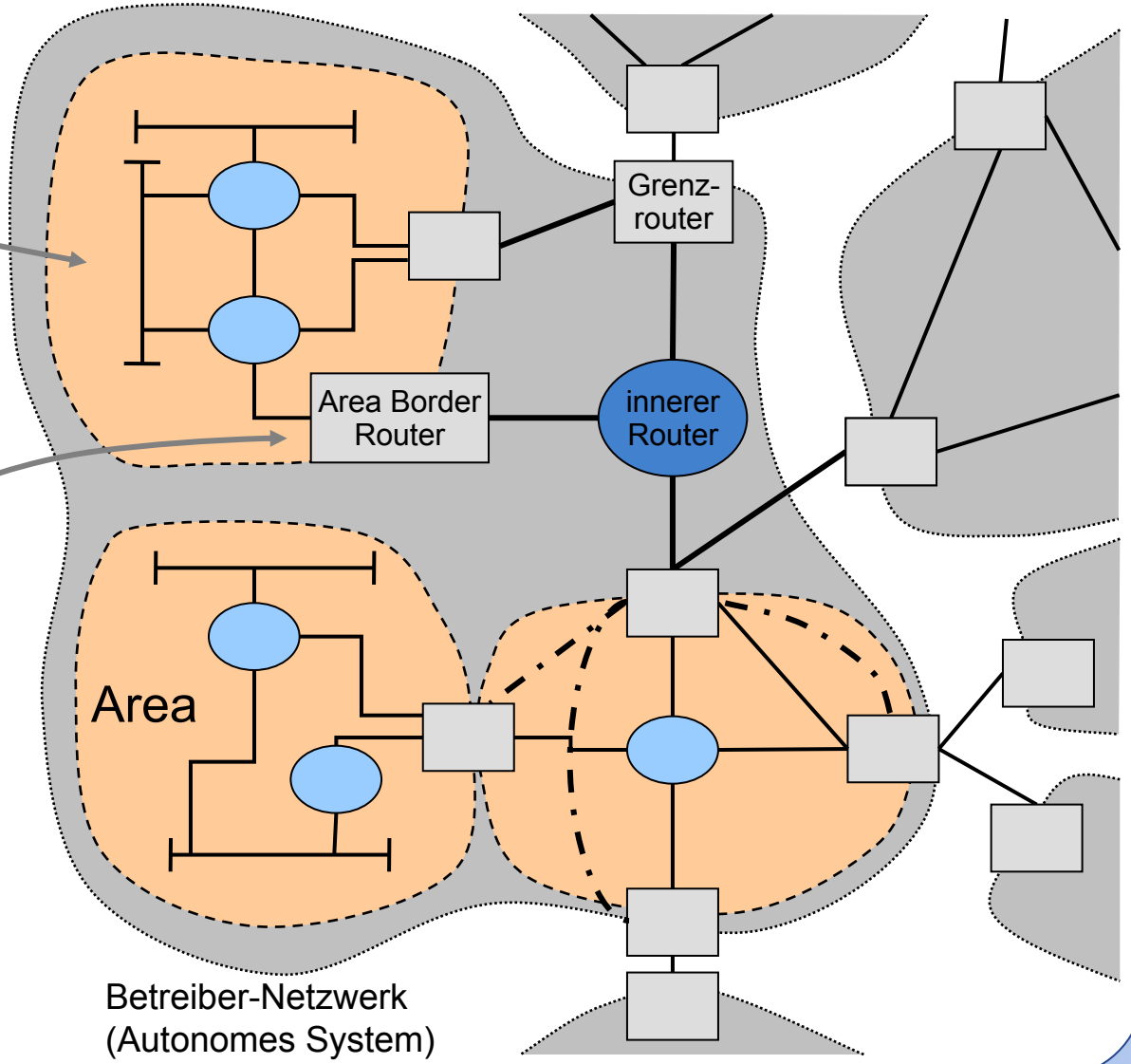


OSPF – Eigenschaften

- Authentifikation von Routing-Nachrichten
  - Routing-Nachrichten werden als Link State Advertisements (LSAs) bezeichnet
  - LSAs werden nur von berechtigten Knoten angenommen
    - Ein „fehlkonfigurierter“ Router kann so keinen Schaden anrichten
- Lastausgleich
  - Mehrere Routen zum gleichen Ziel sind möglich
  - Diese Routen kosten gleich viel
    - Symmetrische Lastverteilung
  - Unterschiedliche Routen in Abhängigkeit vom Inhalt des IP-Kopffeldes „*Type of Service*” (ToS)

OSPF – Hierarchie

- Zusätzliche Hierarchie (OSPF-Area)
  - Router werden in Gruppen unterteilt
    - Skalierbarkeit
  - So genannte *Area Border Router* verbinden eine solche Gruppe mit anderen
    - Reduzierung der Anzahl von Routing-Nachrichten



3.3.4 Exterior Gateway Protocols

- Grundlage

[Huit00]

  - Aufteilung großer Netze in „Autonome Systeme“ (AS)
    - Ansonsten Anzahl der Einträge in der Routing-Tabelle und Menge der ausgetauschten Routing-Information nicht skalierbar mit Netzgröße
  - Autonome Systeme haben i.d.R. nur Routing-Informationen über sich
  - In jedem Autonomen System gibt es zumindest ein ausgezeichnetes Zwischensystem, das als Schnittstelle zu anderen Autonomen Systemen dient
- Vorteil
  - Skalierbarkeit
    - Größe der Routing-Tabellen ist abhängig von der Größe des AS
    - Änderungen von Einträgen in den Routing-Tabellen werden nur innerhalb eines Autonomen Systems weitergegeben
  - Autonomie
    - Internet = Netz von Netzen
    - Routing kann im eigenen Netz kontrolliert werden
      - Im Autonomen System ein einheitliches Routing-Protokoll
      - Routing-Protokolle der ASe müssen nicht identisch sein

Das Border Gateway Protokoll (BGP)

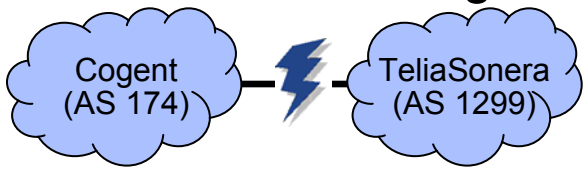
- BGP ist das bedeutendste EGP
  - Basis des heutigen internetweiten Routings
  - Weltweiter Einsatz zwischen allen autonomen Systemen

- Pfad-Vektor-Protokoll
  - Erweiterung zum Distanz-Vektor
  - BGP verbreitet keine Metriken wie Kosten etc. sondern Pfade
    - Pfade garantieren Schleifenfreiheit
  - Ausschlaggebend für die Wegewahl sind die Vorgaben des Netzbetreibers (Policies)
    - Z.B. Wahl möglichst wirtschaftlicher Wege, Beachtung von vertraglichen Vereinbarungen, etc.


[Huit00]

Provider-Krach sorgt für Riss im Internet

- Am 14. März 2008 kappt der IP-Carrier Cogent sämtliche Verbindungen zu TeliaSonera



  - Grund für die Streitigkeiten ist bis heute unklar
    - Vermutlich ging es um wirtschaftliche Interessen und Macht
  - Ca. 12 Stunden lang konnte eine Alternativroute über Verizon genutzt werden, dann war auch diese nicht mehr verfügbar
  - Folge
    - Kunden von TeliaSonera konnten Webserver, die über Cogent angebunden waren, nicht mehr erreichen und umgekehrt
    - Besonders problematisch
      - Die europäischen Server von World of Warcraft sind ausschließlich bei TeliaSonera gehostet
      - Blizzard gab Cogent die Schuld und wartete das Ende des Streits ab

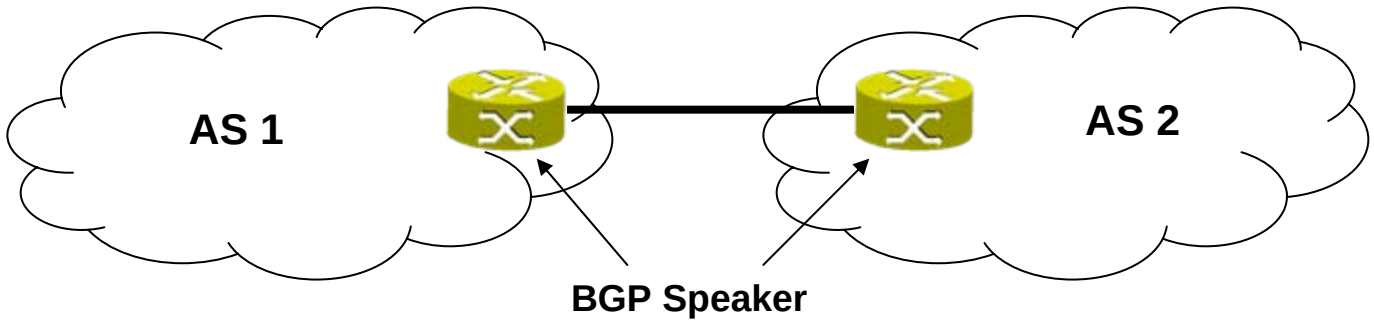
- Am 28. März 2008 funktionierte alles wieder wie gehabt
  - Eine gemeinsame Erklärung oder Stellungnahme beider Carrier ist bis heute nicht bekannt

BGP-Kommunikation

- Punkt-zu-Punkt
  - BGP wird i.A. nur mit direkten Nachbarn „gesprochen“
    - Direkte Nachbarn werden „Peers“ genannt
  - BGP nutzt TCP-Verbindungen zwischen diesen Nachbarn
- BGP-Nachrichtentypen
  - OPEN
    - Aufbau einer BGP-Verbindung zum Peer
    - Authentisierung
  - UPDATE
    - Bekanntgabe eines neuen oder Zurücknahme eines veralteten Pfads
    - Achtung: Wird nur gesendet, falls neue, beste Pfade verfügbar
      - Für bessere Skalierbarkeit (ohne Neuigkeiten bleibt BGP relativ stumm)
  - KEEPALIVE
    - Hält Verbindung aufrecht in Abwesenheit von UDPATE-Nachrichten
    - Quittung zu einem OPEN-Request
  - NOTIFICATION
    - Fehlermeldung und Abbau einer BGP-Verbindung

BGP

- BGP-Einsatz zwischen autonomen Systemen: Externes BGP (EBGP)
  - Wird zwischen den BGP-Routern (auch BGP Speaker genannt) zweier Autonomer Systeme gesprochen
  - Diese BGP-Router sollten direkt miteinander verbunden sein
  - Interne Detail-Informationen des AS werden nicht über diese BGP-Router ausgetauscht

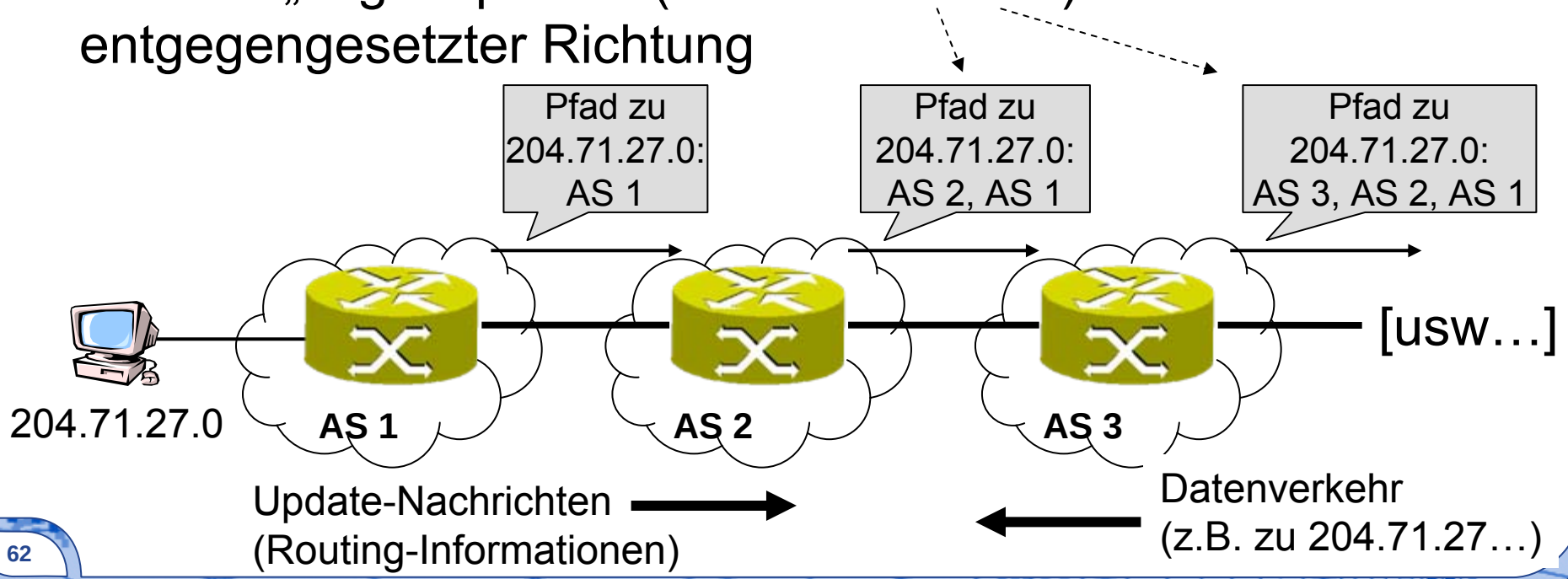


Funktionsweise BGP

- Was wird durch Update-Nachrichten verbreitet?
  - Pfade (auch Routen genannt) bestehend aus:
    - Ziel: Präfixe, Netzwerke, Netzwerk-Präfixe, IP-Adressbereiche.
    - Attribute: Pfad, Next Hop
      - Jedes durchquerte Autonome System stellt sich im Pfad voran

Classless Inter-Domain Routing

- Verkehr „folgt“ Updates (Announcements) in entgegengesetzter Richtung

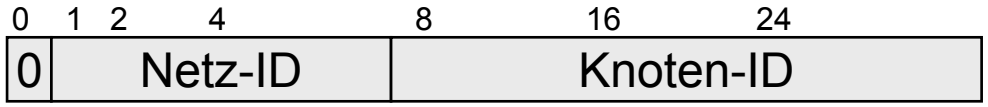


Einschub: Adressierung im Internet

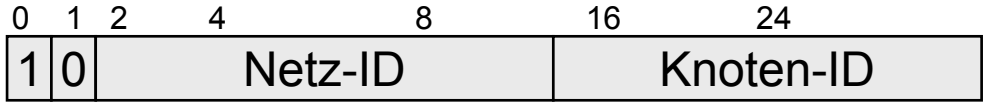
- Ziel
  - Eindeutige Identifizierung aller im Internet angeschlossenen Netz- und Endsysteme bzw. deren einzelner Schnittstellen
    - Ein Gerät kann mehrere Interfaces haben
- Vorgehensweise
  - Weltweit eindeutige Adressen auf Schicht 3: IP-Adressen
  - Einfaches, für Maschinen leicht zu verarbeitendes Format
  - IPv4
    - Adressen einer Länge von 32 Bit
  - IPv6
    - Größerer Adressraum durch Adressen von 128 Bit Länge

Einschub: Historische Adressklassen

- Ursprünglich unterstützte IP fünf verschiedene Adressklassen:
  - Class A für Netze mit mehr als 65.536 Knoten

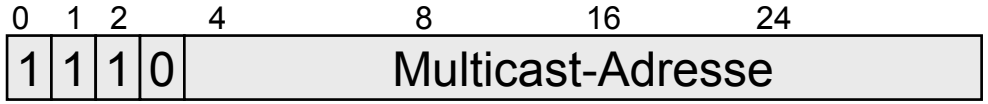


  - Class B für Netze zwischen 256 und 65.536 Knoten



  - Class C für Netze mit weniger als 256 Knoten

  - Class D für Gruppen-kommunikation (Multicast)



  - Class E, reserviert für zukünftige Anwendungen

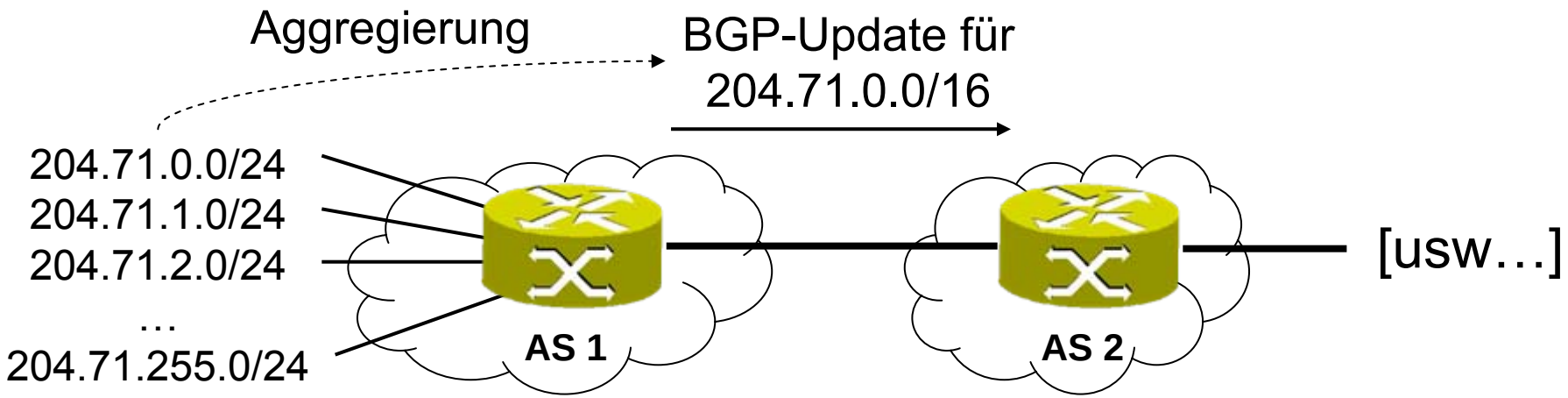
- Diese Form der Adressierung wird heute nicht mehr direkt eingesetzt
  - ⇨ Classless Inter-Domain Routing (CIDR, 1993)

Einschub: CIDR
Classless Inter-Domain Routing

- Bisher: Drei Adressklassen für Unicast, somit schlechte Ausnutzung durch ungenutzte Adressen („Verschnitt“), z.B.:
  - Größeres Netz mit mehr als 254 Komponenten benötigt Class-B-Adresse
  - Kleines Netz mit 100 IP-Adressen, benötigt Class-C-Adresse
    - 254 Adressen wären verfügbar, damit 154 ungenutzte Adressen
- CIDR
  - Ersetzen der festen Klassen durch Netzwerk-Präfixe variabler Länge
    - Beispiele
      - 129.24.12.0/14: Die ersten 14 Bits der IP-Adresse werden für die Netzwerk-Identifikation verwendet
      - 141.3.64.0/21 = 141.3.64.0 bis 141.3.71.255
    - Einsatz in Verbindung mit hierarchischem Routing:
      - Backbone-Router, z.B. an Transatlantik-Link, betrachtet z.B. nur die ersten 13 Bits; dadurch kleine Routing-Tabellen, wenig Rechenaufwand
      - Router eines angeschlossenen Providers z.B. die ersten 15 Bit
      - Router in einem Firmennetz mit 128 Hosts betrachtet 25 Bits

Aggregierung in BGP mittels CIDR

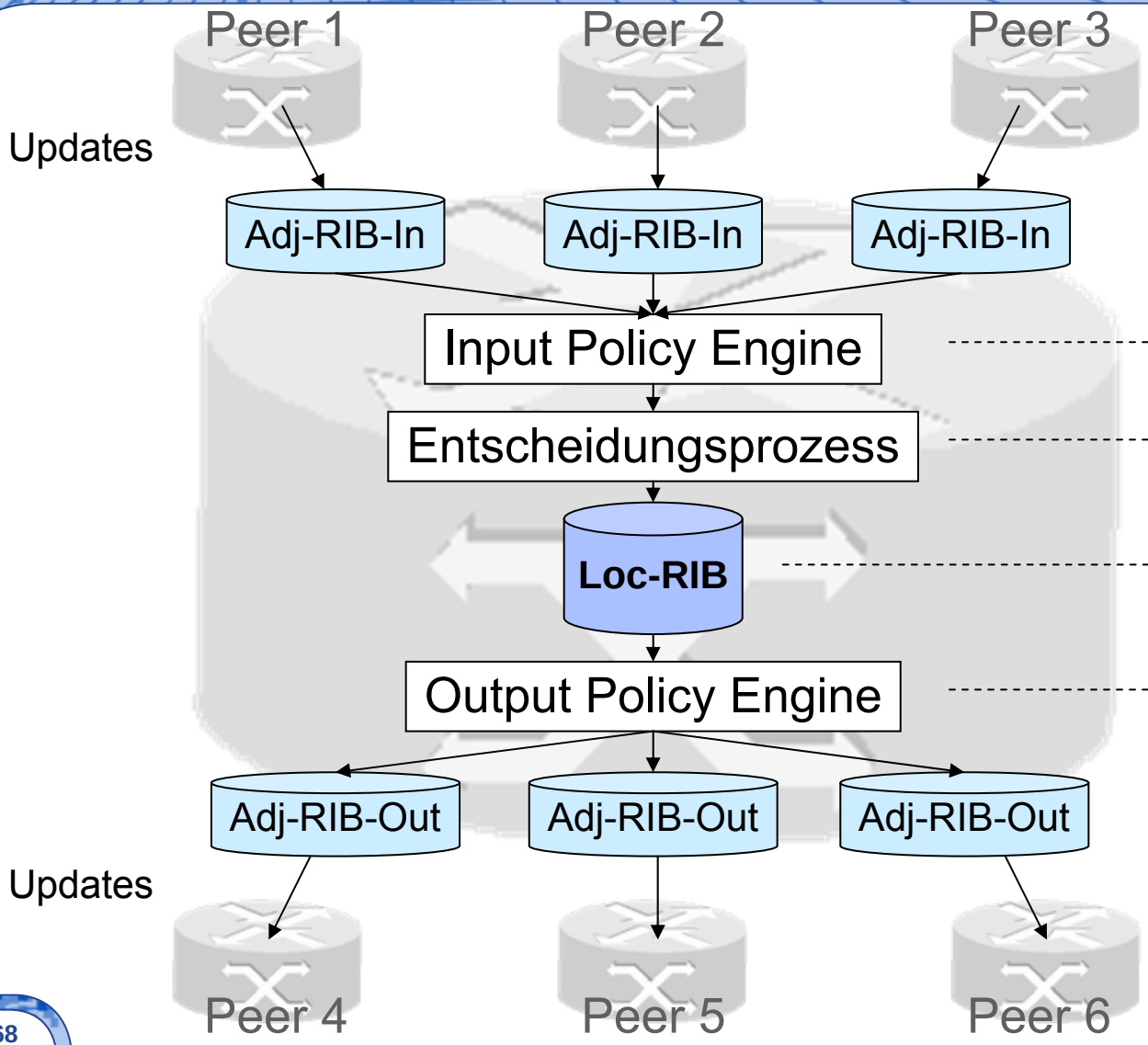
- Durch geschickte Adressvergabe können mehrere Adressbereiche durch ein einziges Präfix zusammengefasst werden
- Einsparung von Routing-Nachrichten durch diese Aggregierung
  - Verbessert die Skalierbarkeit



BGP-Verwaltung von Routing-Informationen

- Die BGP-Instanz eines Routers sammelt erhaltene und versandte Routing-Informationen in verschiedenen internen Tabellen (Routing Information Base, RIB)
  - Adj-RIB-In (*Adjacency RIB Incoming*)
    - Existiert pro Peer
    - Speichert die Informationen, die von diesem Peer empfangen wurden
  - Loc-RIB (Local RIB, *Forwarding Information Base*)
    - „Eigentliche Routingtabelle“
      - Hier sind nur die bevorzugten (= besten) Routen zu den Zielnetzen enthalten.
      - Diese Routen bilden die Forwarding Information Base (FIB)
  - Adj-RIB-Out (*Adjacency RIB Outgoing*)
    - Existiert pro Peer
    - Enthält Routen, die an diesen Peer weitergesendet wurden
- Diese Datenhaltung dient hauptsächlich der logischen Strukturierung

BGP Routing Engine



Filterung der Informationen nach definierten Regeln

Auswahl der besten Route

Verwendete Routen (Routing-Tabelle)

Filterung der Informationen nach definierten Regeln

Beispiel: Auszug aus BGP-Tabelle

```
myrouter>show ip bgp
BGP table version is 6543445, local router ID is .....
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
*  3.0.0.0  202.249.2.86                           0 7500 2516 701 80 i
*           167.142.3.6                            0 5056 701 80 i
*           195.66.224.82         23302            0 4513 701 80 i
*           195.219.96.239                         0 8297 6453 701 80 i
*           192.121.154.25                         0 1755 701 80 i
*           205.215.45.50                          0 4006 701 80 i
*           195.211.29.22                          0 5409 6667 854
*           207.172.6.173            22            0 6079 701 80 i
*           206.220.240.222                        0 10764 1 701 80 i
*>          157.130.185.17                         0 701 80 i
*           157.22.9.7                             0 715 701 80 i
*  4.40.228.0/23    206.220.240.222                0 10764 11537 5661 13755 i
*                   203.181.248.242                0 7660 11537 5661 13755 i
*                   193.0.0.56                     0 3333 1103 […] 13755 i
*                   134.55.20.229                  0 293 11537 5661 13755 i
*>                  198.32.8.252                   0 11537 5661 13755 i
```

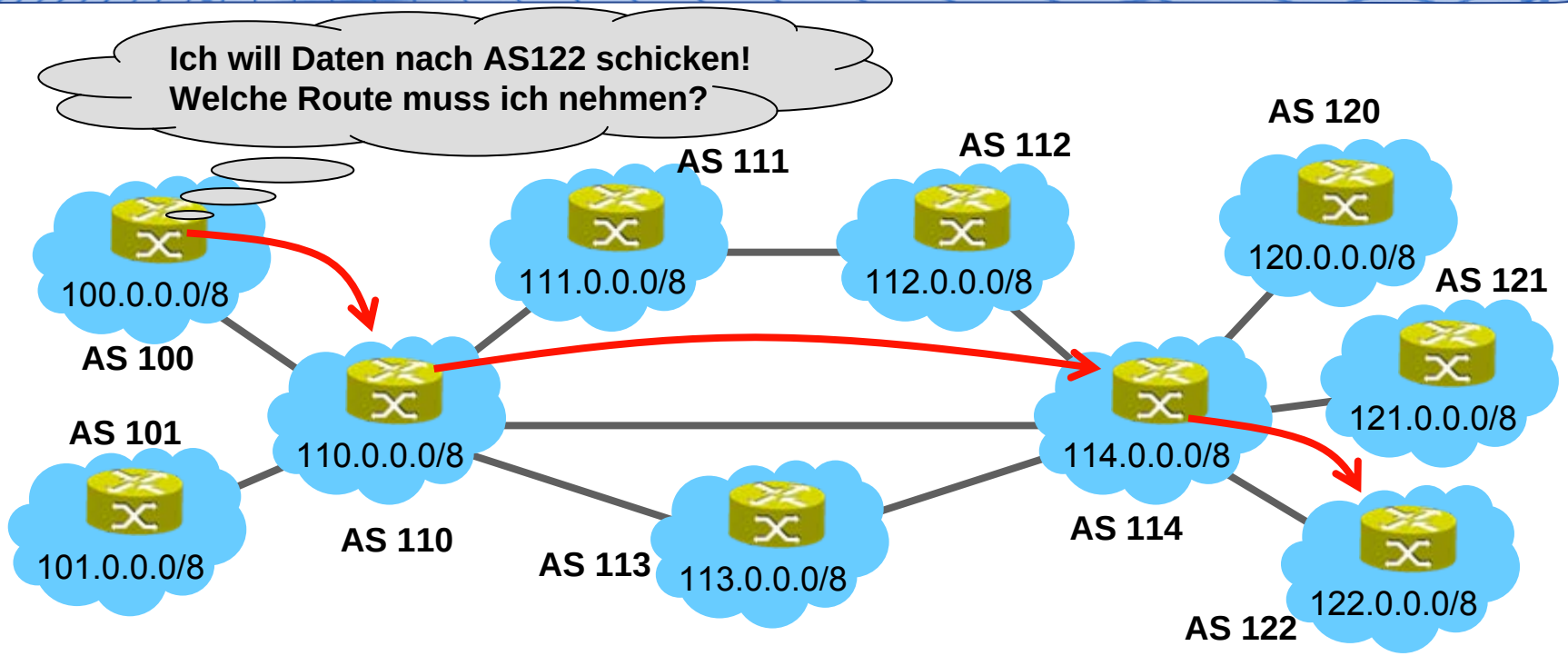
AS Pfad

Ziel AS

Beste Route

Zielnetz

Beispiel zu BGP – Topologie



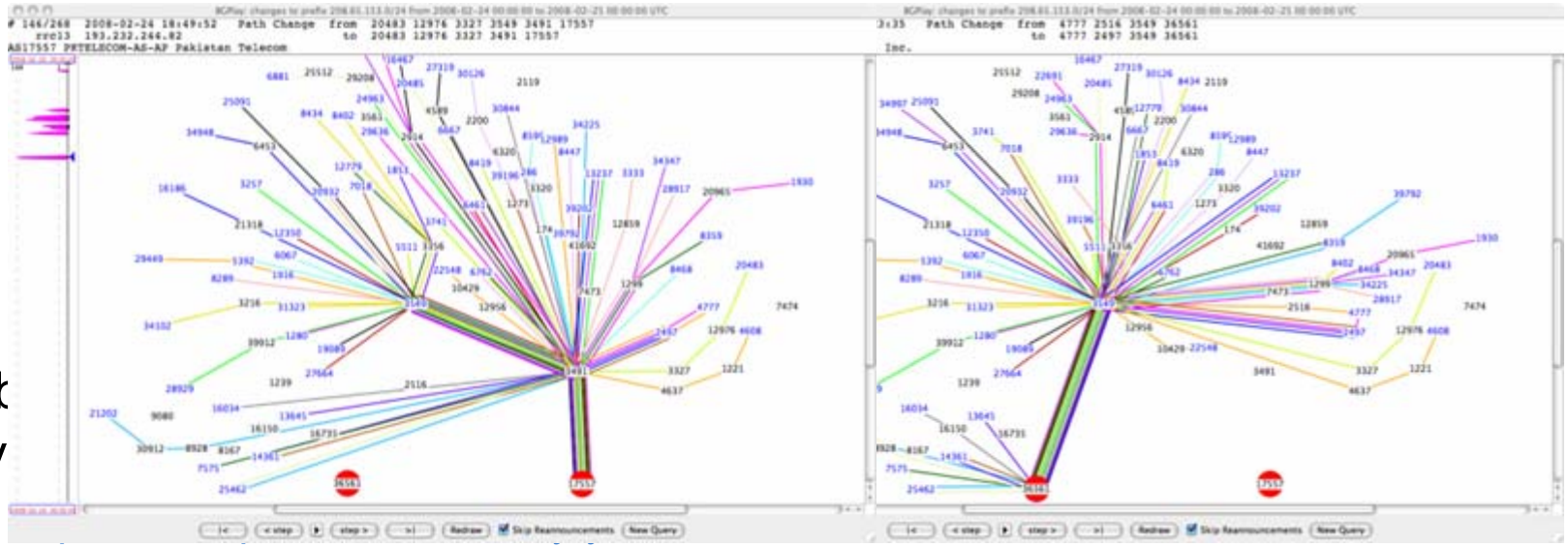
Routing-Tabelle AS100

Routing-Tabelle AS110

Routing-Tabelle AS114

Network	Next Hop	Metric	LocPrf	Weight	Path
*> 112.0.0.0	10.1.1.112		0		0 112 i
*	10.1.1.110				0 110 111 112 i
…					
*> 122.0.0.0	10.1.1.122		0		0 122 i

YouTube-Ausfall

- Am 24. Februar 2008 beschließt die pakistanische Regierung die landesweite Sperrung von YouTube
  - Pakistan Telecom (AS 17557) gibt über BGP das Präfix 208.65.153.0/24 bekannt, welches eigentlich YouTube (AS 36561) gehört
  - Dies verbreitet sich innerhalb weniger Minuten weltweit
    - Die Folge: YouTube-Aufrufe werden direkt nach Pakistan geleitet
      ➔ YouTube ist nicht mehr erreichbar
  - 
- Beob
  Serv
  -

Die Routen zu YouTube zeigen auf Pakistan Telecom (AS 17557)

Die Routen zu YouTube zeigen wieder auf das korrekte AS 36561

3.3.4.1 Probleme und Herausforderungen

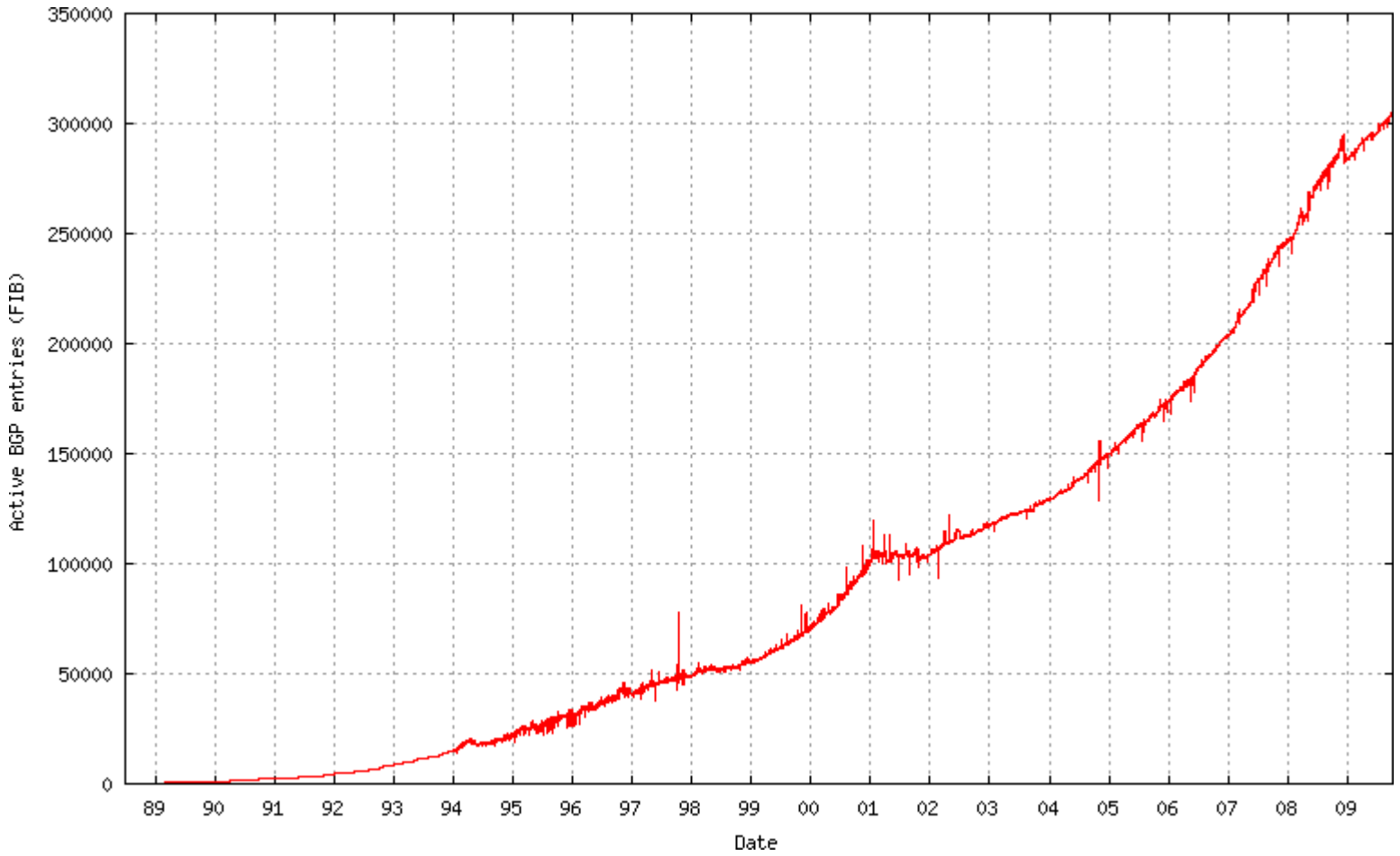
- Diese Probleme von BGP stellen gegenwärtig große Herausforderungen dar:
  - Aufrechterhaltung der Skalierbarkeit
    - Zunehmende Deaggregierung von Routing-Informationen
      - Beispielsweise aufgrund des Multi-Homing von ASen
    - Wachstum der Routing-Tabellen
    - Zunehmende Dynamik der Routing-Änderungen
  - Gestiegene Anforderungen an das Internet
  - Sicherheitsprobleme
  - Entwicklung zukünftiger Inter-Domain-Routingprotokolle

Probleme von BGP – Multi-Homing

- Was ist Multi-Homing?
  - Ein Autonomes System am Rand des Internets (zum Beispiel das BelWue) wird über mehr als ein Autonomes System an das Internet angeschlossen (zum Beispiel Telekom und Colt)
- Warum „multi-homen“ Autonome Systeme?
  - Ausfallsicherheit der Internet-Anbindung
  - Kosten (wichtiger Verkehr über teuren Uplink, anderer Verkehr über günstigen Uplink)
- Wo liegt das Problem?
  - Aggregierung der Präfixe wird aufgebrochen
  - Änderungen der bevorzugten Route müssen schlimmstenfalls im ganzen Internet propagiert werden
- Mögliche Abhilfe:
  - NOPEER-Attribut: Schränkt die Propagierung von Änderungen am Rand des Internets ein

[Hust04]

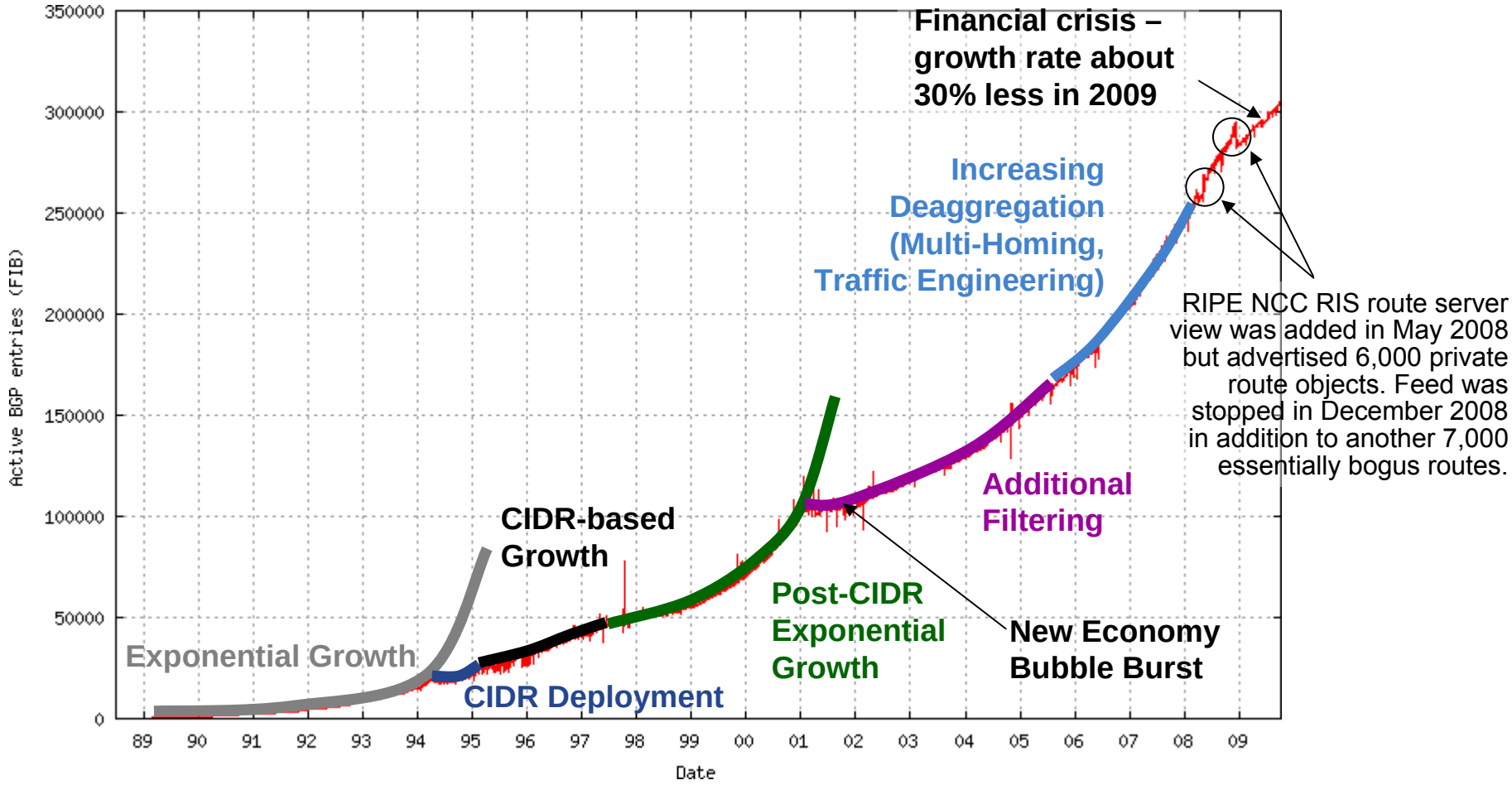
Größe von Routing-Tabellen (IPv4)



FIB: Forwarding Information Base
BGP: Border Gateway Protocol

Quelle: http://bgp.potaroo.net/

Wachstum von Routing-Tabellen



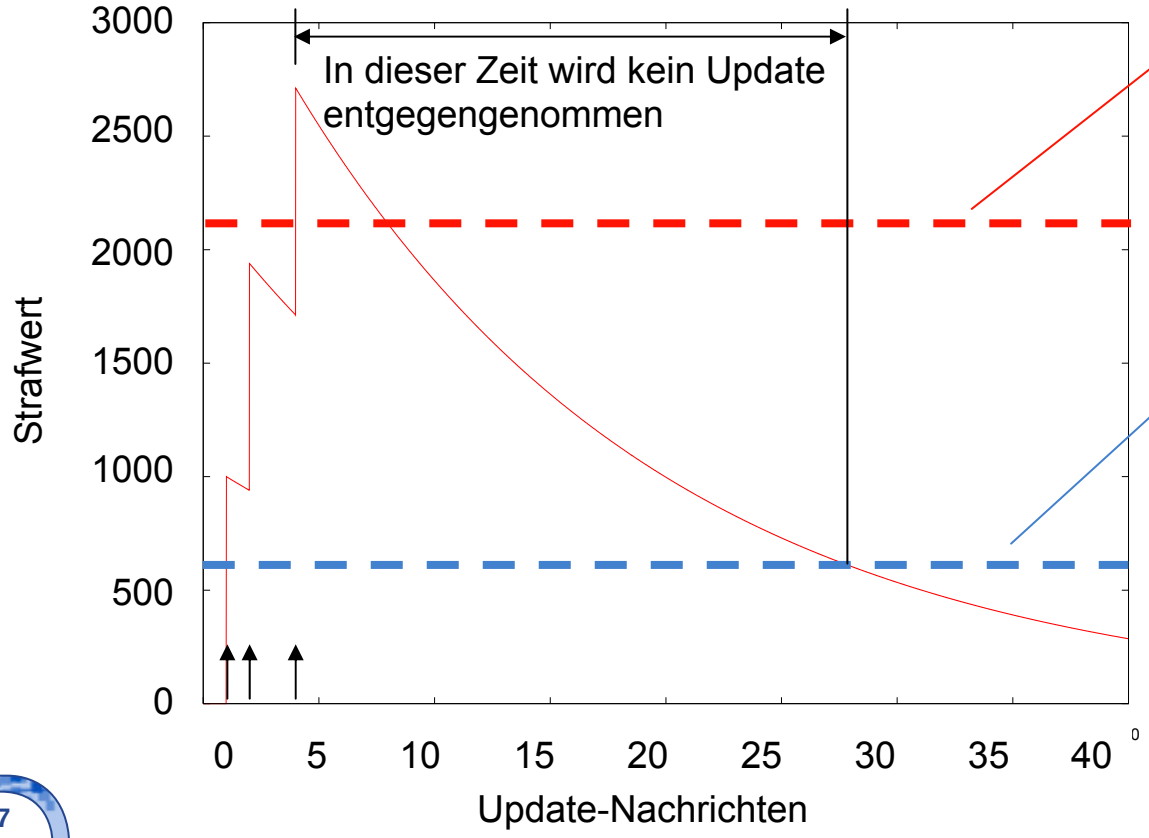
Ursachen: Zunehmendes Multi-Homing und Verkehrslastlenkung über BGP
Auswirkung: Viele kleine Präfixbereiche (>/20) werden propagiert
Als Gegenmaßnahme filtern Betreiber stärker unnötige Details

Probleme von BGP – Zunehmende Dynamik von Routing-Änderungen

- Problem
  - Die Nummerierung der Autonomen Systeme ist flach und nicht untergliedert
  - Autonome Systeme unterscheiden sich in ihrer Funktion im Internet:
    - Stub-ASe bzw. Multi-Homed-ASe (Autonome Systeme am Rand des Internets)
    - Transit-ASe (Autonome Systeme, die den Verkehr der Stub-ASe weiterleiten)
  - BGP kennt keine Hierarchie, wie kann die Propagierung von Änderungen der Topologie „sinnvoll“ eingeschränkt werden?
- Auswirkungen
  - Aggregierung von Präfixen nimmt aufgrund von Multi-Homing und Traffic-Engineering weiter ab [MeZF07]
  - Vermaschung der Autonomen Systeme untereinander wird dichter [Hust06]
  - Routen der Routing-Tabelle ändern sich häufiger
- Mögliche Abhilfe
  - Route Flap Damping
    - Temporäres Unterdrücken von instabilen Routen [Vill98]

Route Flap Damping

- Temporäres Unterdrücken der Änderungen von instabilen Routen
- Pro Update wird ein Strafwert erhöht (hier 1000 Punkte pro Update)
- Der Strafwert fällt exponentiell mit der Zeit wieder ab
- Kann zu Konnektivitätsverlust führen!



Ab diesem Schwellenwert werden die Updates unterdrückt

Ab diesem Schwellenwert werden Updates wieder angenommen

Probleme von BGP – Steigende Qualitätsanforderungen

- Steigende Anforderungen von Anwendungen an die Übertragungsqualität
  - VoIP, Skype usw. benötigen geringen Jitter und kleine Verzögerungen
    - Verzögerung und Jitter sind hierbei abhängig von der Route bzw. den Routenwechseln
    - Auswirkungen: Was passiert, wenn eine Route nicht stabil ist?
      - Sprachverbindung klingt blechern oder hat Aussetzer
      - Abbruch der VoIP-Sitzung bei ca. 2% Paketverlust (abhängig vom Codec)
- BGP mangelt es an Mechanismen, qualitativ hochwertige Routen zu wählen
  - Benötigt würden beispielsweise Routen mit geringem Paketverlust, hoher Bandbreite, geringer Verzögerung, etc.
  - BGP bietet lediglich Länge des AS-Pfads als Metrik
- Abhilfen
  - Andere BGP-Metriken bisher nicht absehbar
  - Reduzierung der Routenänderungshäufigkeit:
    - Route Flap Damping – kann zu Verbindungsverlust führen
    - NOPEER Attribut – hilft nur bei Rand-Autonomen Systemen
  - Ansätze helfen alle nur bei Teil-Problemen

Sicherheit des Inter-Domain-Routings

- Problem
  - Netzbetreiber verdienen mit der Bekanntgabe von Routeninformationen Geld
    - A schickt an B eine Routeninformation, der Datenverkehr fließt dann von B nach A
  - Wie kann im Inter-Domain-Bereich …
    - das Routing-Protokoll (die Routing-Sitzung) geschützt werden?
    - die Informationen in Routing-Nachrichten geschützt werden?
    - die Authorisierung der Weitergabe von Routeninformationen sichergestellt werden?
  - Welche Instanz oder Instanzen werden von allen Netzbetreibern (USA, Europa, Asien, …) als vertrauenswürdige Instanz akzeptiert?
    - Beispielsweise als Vertrauensanker für kryptographische Zertifikate, Signaturen, etc.

Sicherheit des Inter-Domain-Routing

- Lösungsansätze
  - Sicherung der Routeninformationen
  - Erweiterungen von BGP zur Absicherung:
    - soBGP (secure origin BGP) von Cisco
    - S-BGP (secure BGP) von BBN-Technologies
    - Prinzipiell ähnliche Ansätze
      - Authentifizierung und Autorisierung von Absendern
      - Integritätsschutz und Verschlüsselung des Inhalts von Routing-Informationen
    - Dennoch zahlreiche Unterschiede:
      - Zum Beispiel Übertragung von Zertifikaten in-band oder out-of-band
  - Sicherung der Routing-Sitzung durch IPSec und TCP mit MD5
- Umsetzung scheitert unter anderem an
  - Fehlendem Leidensdruck bei den Netzbetreibern
  - Fortwährenden Diskussionen zwischen Cisco (R. White) und BBN-Technologies (S. Kent) bei der IETF

[Hust05,Hust05a]

Neue Inter-Domain-Routingprotokolle

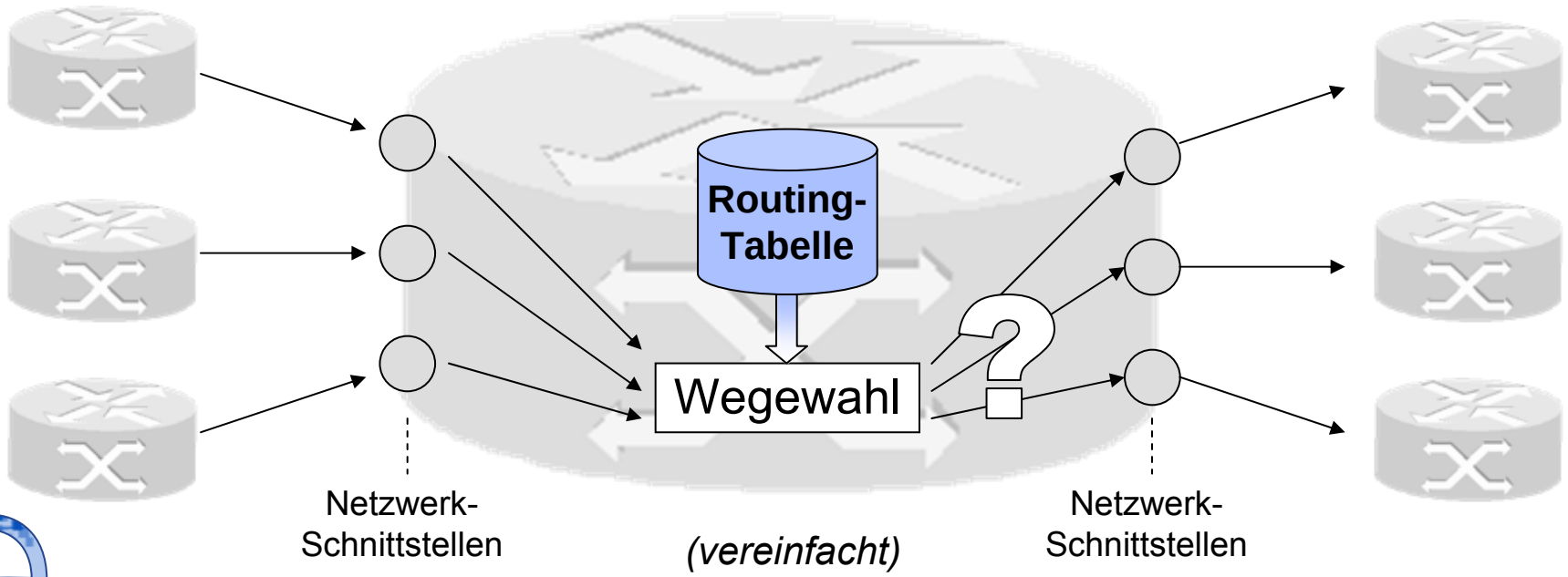
- Problem:
  - Routing-Konvergenz im Inter-Domain-Bereich nicht schnell genug
    - ~ 15 Minuten und länger
    - In dieser Zeit können Schleifen auftreten (Micro Loop Problem)
  - Common sense: Es wird ein neues Routing-Protokoll benötigt, ABER:
    - Kein Flag Day möglich!
    - Muss alles können, was BGP bisher schon kann (und am besten noch mehr)
    - Muss stabil sein, darf keine Probleme machen … eierlegende Wollmilchsau?
- Lösungsansätze:
  - IRTF (Internet Research Task Force) diskutiert in zwei Gruppen seit 2002 über ein Dokument mit Anforderungen an ein neues Routing-Protokoll bzw. eine neue Routing-Architektur
  - Immer wieder neue Vorschläge auf Konferenzen (z.B. Metarouting und HLP[1] Sigcomm 2005), ABER Netzwerkoperatoren sind „sehr zaghaft“

[1] HLP – Hybrid Link-State and Path-Vector Protocol

[Grif05, Shen05]

3.3.5 Weiterleitung im IP-Router

- Aufgaben der Weiterleitung (Forwarding)
  - Wegewahl mit Hilfe der Routing-Tabelle
    - Identifikation des Next-Hop
  - Weitersenden an diesen Next-Hop über entsprechende Schnittstelle
- Herausforderung:
  - Soll mit möglichst hoher Geschwindigkeit geschehen



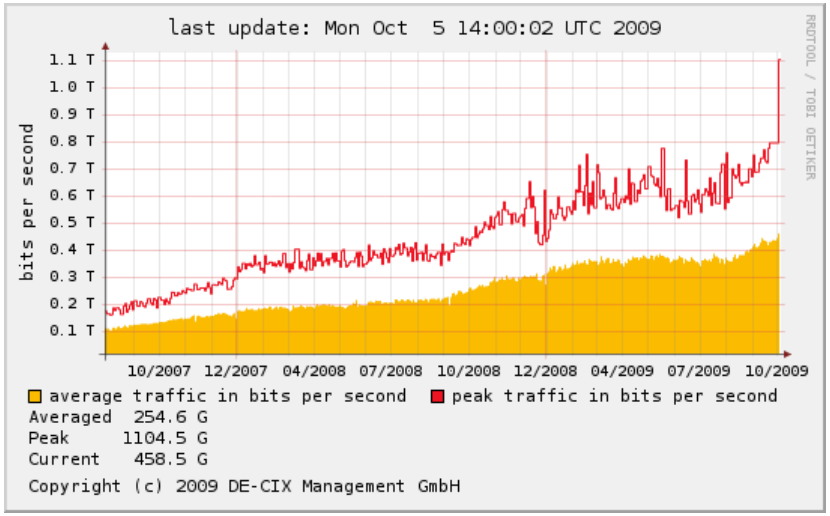
Einschub: Herausforderung schnelles Forwarding

- Kaum Zeit pro Dateneinheit
  - z.B. Ciscos Hochleistungsrouter Carrier Routing System CRS-1
  - bis zu 1,2 TBit/s pro Chassis
    - = Gehäuse mit Zentraleinheit, schnellem Bus (Backplane) und Einschüben für Netzwerkkarten
    - Kosten Chassis: knapp 500.000$
    - Kosten für eine Schnittstellenkarte: etwa 100.000-650.000$ (je nach Ausführung)
      - (Aber: Teils hohe Rabatte üblich)
  - 40 GBit/s pro Karteneinschub
    - Nur wenige dutzend Nanosekunden Bearbeitungszeit pro Dateneinheit

Einschub: Herausforderung schnelles Forwarding

- Zukünftige Skalierbarkeit
  - am Beispiel des deutschen Austauschpunktes De-Cix
    - Vier Standorte in Frankfurt (Main)
    - 250 ISPs
    - 200+ GBit/s-Ports, 350+ 10Gbit/s-Ports
    - 1105 GBit/s Peak Throughput
    - 255 GBit/s im Durchschnitt
    - 2,0 TBit/s angeschlossene Übertragungskapazität
  - Über De-Cix ausgetauschte Datenmenge wächst stetig weiter

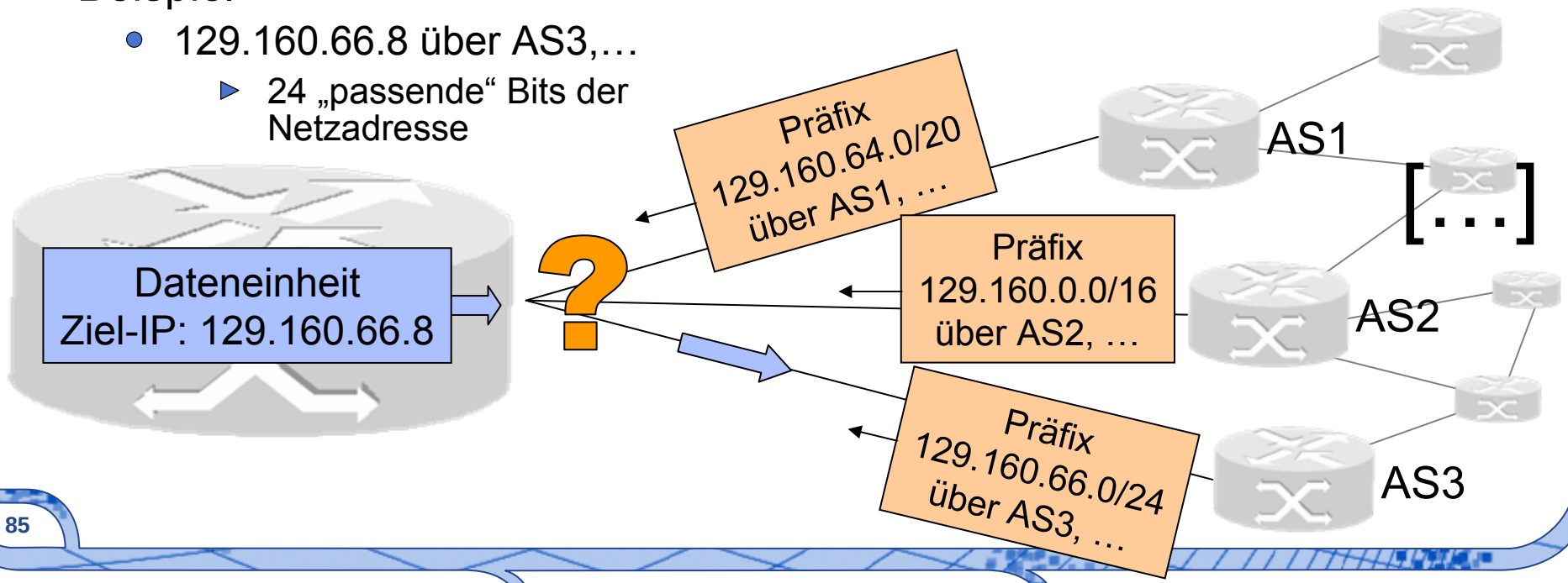
DE-CIX



[http://de-cix.de]

Longest Prefix Matching

- Problem
  - Wohin weiterleiten, falls mehrere Präfixe in der Routing-Tabelle auf eine Zieladresse passen (matchen)?
- Regel
  - Spezifischsten Eintrag wählen (most specific)
    - Größte Anzahl Bits des Netzwerkteils sollen übereinstimmen
  - Longest Prefix Matching
- Beispiel
  - 129.160.66.8 über AS3,…
    - 24 „passende“ Bits der Netzadresse



3.3.5.1 Präfix-Suche

- Problem
  - Wie können IP-Dateneinheiten effizient nach Netzwerken bzw. Sub-Netzwerken klassifiziert werden?
  - Unter Beachtung der Vorgabe „longest prefix match first“
- Generelles Vorgehen:

Nr.	Ziel-Präfix	Port
1	0110*	0
2	01101*	1
3	0110110*	2
4	01101000*	3

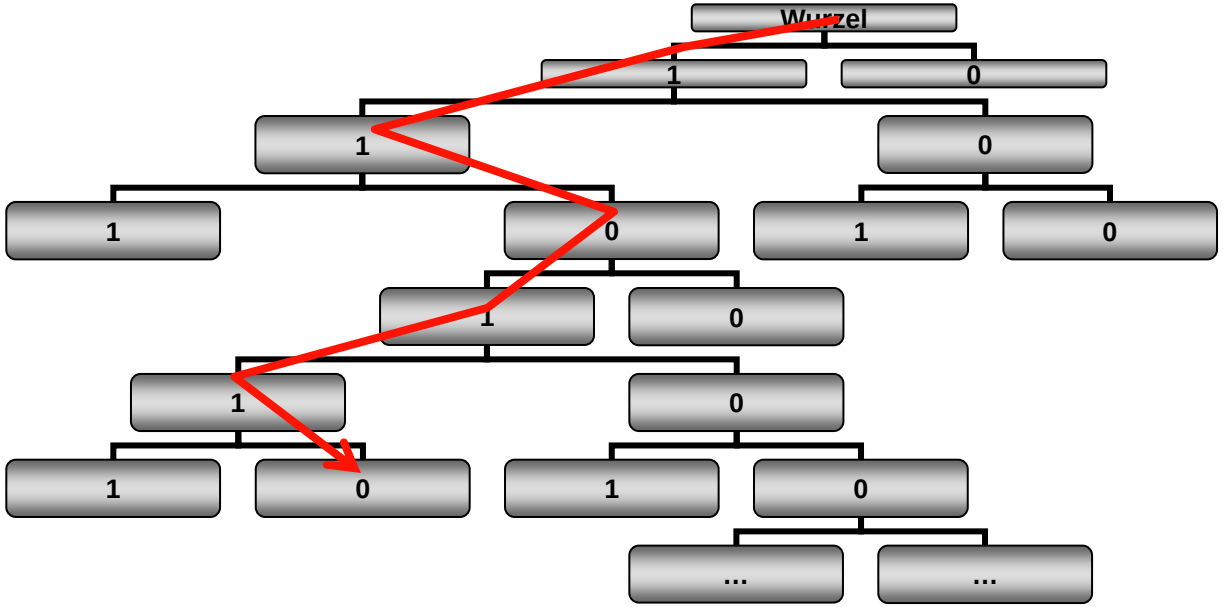
IP-Dateneinheit mit Ziel 109.21.33.9 ➔ 0110 1101 . 0001 0101 . 0010 0001 . 0000 1001

Verfahren zur Klassifikation:

Binärer Trie, Patricia Trie, Hash-Tabellen

Binärer Trie

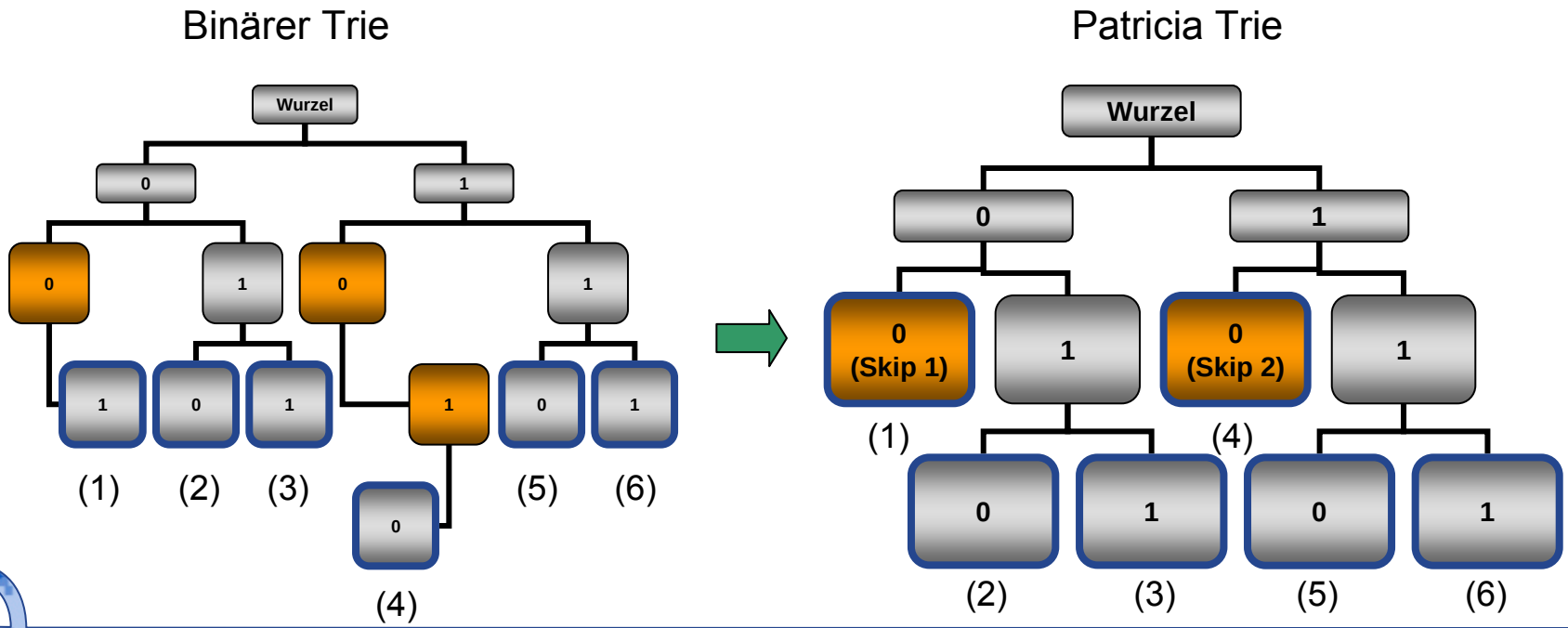
- Trie abgeleitet aus dem engl. re**Trie**val (wieder finden)
- In einem Trie repräsentiert jede Kante des Baums ein zusätzliches Bit des Netzwerks. Jeder Knoten entspricht der Bitkette, die aus der Verkettung aller Kantenbits entsteht. Der Wurzelknoten eines Tries entspricht einer leeren Zeichenkette.
  - Um Daten in komprimierter Form in einem Trie abzulegen, werden Patricia-Tries benutzt
- Beispiel:



110110 …

Patricia Trie

- Patricia-Trie (*Practical Algorithm to Retrieve Information Coded in Alphanumeric*) speichert Daten komprimiert ab
- Knoten ohne Verzweigung werden übersprungen und Anzahl der ausgelassenen Knoten vermerkt (Skip X)
- Reduzierung der Tiefe des binären Baums
- Anzahl der Knoten wird reduziert



3.4 NAT und IPv6

- Internet-Architektur basierend auf Internet Protocol Version 4 (IPv4)
  - Seit 1981 standardisiert in RFC 791
  - 32 Bit großes Adressfeld → $2^{32} \approx 4$ Milliarden Adressen
    - Wie lange noch ausreichend?
- Entwicklung von IPv6 seit 1992
  - U.a. 128 Bit großes Adressfeld
  - Allerdings …
    - Solange kein Bedarf der Nutzer an IPv6 → keine Notwendigkeit für Provider zur Umstellung der Hardware
    - Solange Provider kein IPv6 zur Verfügung stellen → keine IPv6-Software
- In der Zwischenzeit: Standardisierung der Network Address Translation
  - Adressenknappheit: Zu wenig öffentliche IPv4-Adressen
  - Bildung privater Netze („Intranets“) mit IPv4-Adressen aus einem der „privaten“ Adressräume: 10.0.0.0/8, 172.16.0.0/12, 192.168.0.0/16

3.4.1 Network Address Translation (NAT)

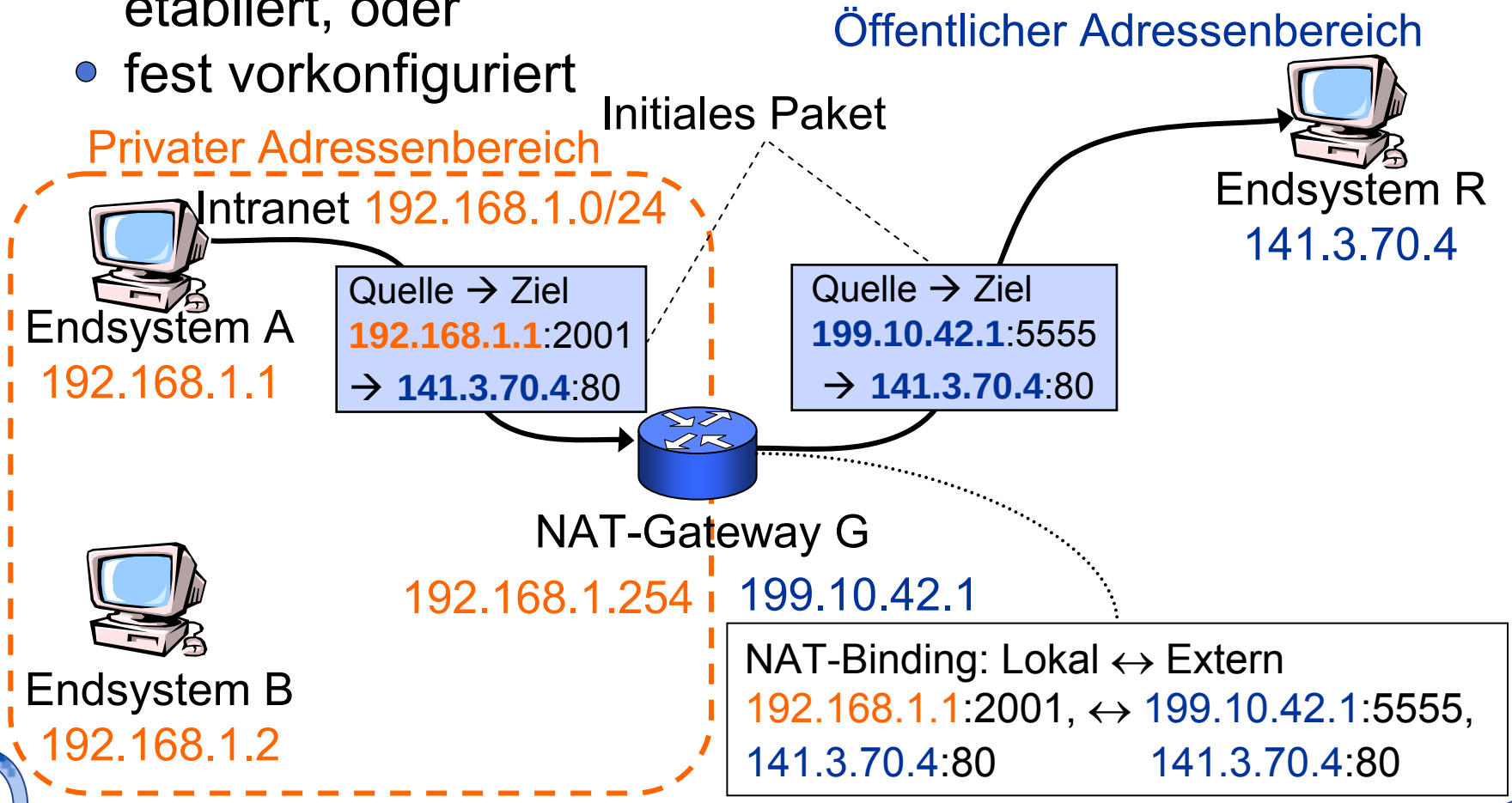
- Network Address Translation (NAT)


[SrEg01]

  - Bijektive Abbildung („Binding“) zwischen IP-Adressen (öffentlich ↔ privat)
  - Umsetzen der Adressen in IP-Dateneinheit (u. ICMP-Dateneinheit) erfolgt durch NAT-Gateway
    - Anpassen der Prüfsummen (ggf. auch für TCP/UDP) notwendig
    - Anwendungsabhängige Anpassung notwendig!
      → Application Level Gateway (ALG)
- Network Address Port Translation (NAPT)
  - Zusätzlich zu „Basic NAT“: Umsetzung von Transport-IDs: TCP/UDP-Ports bzw. ICMP Query ID
    - Bijektive Abbildung zwischen (Adresse,Port)-Tupeln
  - Adressenbedarf wird auf die (eine) öffentliche Adresse des Gateways reduziert
  - Umsetzung der TCP-/UDP-Ports nötig, um Portkollisionen aufzulösen
  - Interne Netzstruktur wird vor dem Internet verborgen
  - Erreichbarkeit von außen nur für freigeschaltete/konfigurierte Adressen

NAPT Beispiel

- Binding
  - wird durch initiales Paket aus dem Intranet heraus etabliert, oder
  - fest vorkonfiguriert



Nachteile eines NAT-Gateways

- Verbindungsloses Konzept/Soft-State
  - Wie lange gilt das Binding?
  - Zustand wird nach Timeout gelöscht (TCP: 4 min, UDP: ?)
- Spontane Erreichbarkeit von außen nicht möglich, da Binding fehlt
- Problem für Protokolle die dynamisch Nutzdatenströme auf neue bzw. von neuen Ports erzeugen
  - z.B. VoIP (SIP + RTP)
- Neuartige Protokolle müssen von NAT-Gateway unterstützt werden
  - z.B. SCTP, DCCP als Alternativen zu TCP und UDP
- Problematisch falls zwei Endsysteme hinter verschiedenen NATs
- Einschränkung der möglichen Anwendungen
  - Bruch des E2E-Prinzips
- Fragmente: NAT fehlt Addresseninformation und verwirft Paket
- Bietet keinen Sicherheitsmechanismus

3.4.2 Internet Protocol Version 6 (IPv6)

Das Internet funktioniert seit Jahrzehnten! Warum ein neues IP-Protokoll?

? ?

- **Anwachsen des Internets:** Der überwältigende Erfolg des Internets führte zu stark anwachsenden Benutzerzahlen und einer deutlichen Erhöhung der Netzbelastung sowie der angeschlossenen Geräteanzahl (Kleinstgeräte, Sensoren) → Bedarf wird eher steigen (mehr Internet-fähige Geräte)
- **Vereinfachtes Management:** Autokonfigurationsmechanismen
- **Wiederherstellung der Kohärenz:** Beseitigung von NAT und anderen Speziallösungen
  → ermöglicht Ende-zu-Ende-Sicherheit und Peer-to-Peer-Netze
- **Effizienteres Routing:** durch entsprechende inhärente Adresshierarchie
- **Hohe Datenraten:** Hochleistungsfähige Zwischensysteme benötigen geeignete Paketformate zur effizienten Bearbeitung

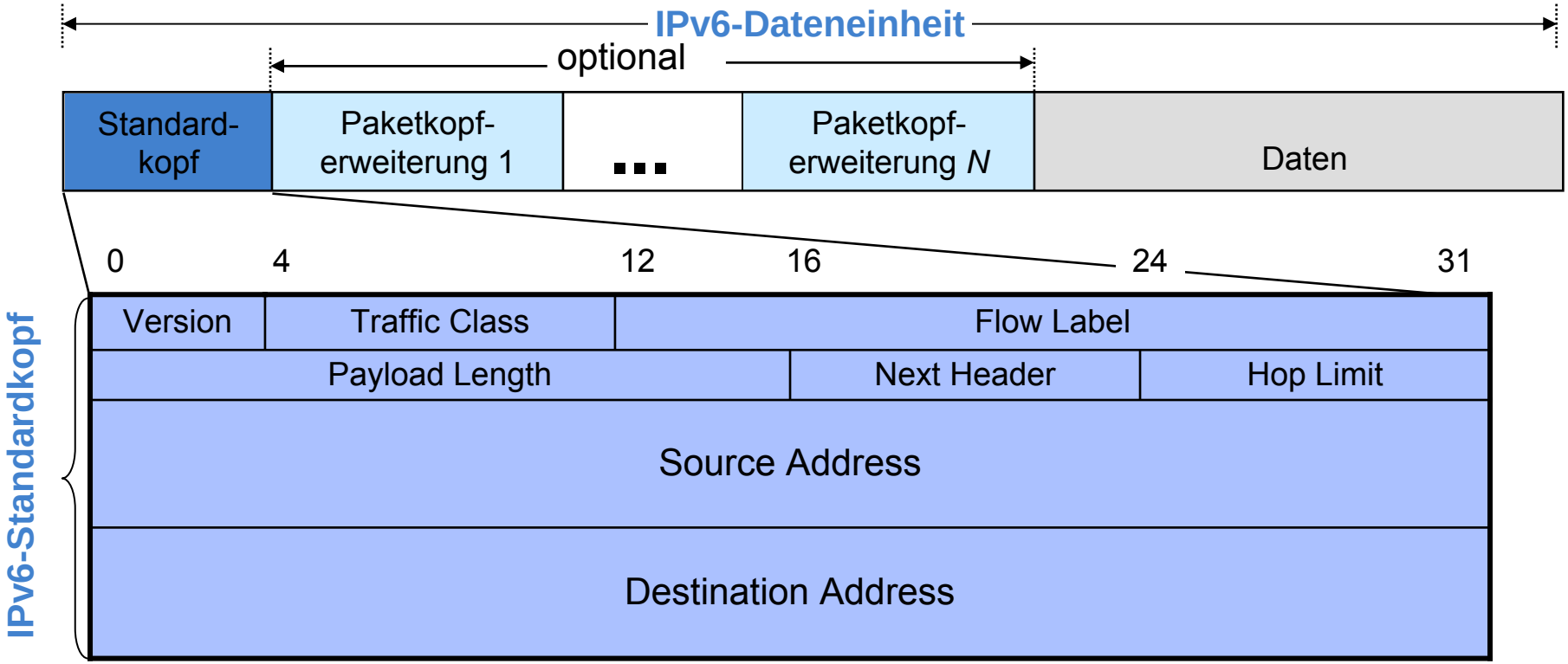
Neuerungen in IPv6

- Erweiterte Adressierung
  - Erhöhung der Adresslänge von 32 Bit auf 128 Bit
  - Einführung von Anycast-Adressen (Kommunikation zum Mitglied einer Gruppe)
  - Jede Schnittstelle besitzt Link-Local-Unicast-Adresse
  - Multicast-Adressen enthalten Reichweite (Scope)
    - Kein „Missbrauch“ des Hop Limit/TTL-Feldes mehr wie bei IPv4
- Schnelle Bearbeitung in Routern durch vereinfachtes Paketformat
  - Standard-Paketkopf mit fester Länge und nur 8 Feldern (13 bei IPv4)
  - Verschieben von Optionen in flexible Paketkopferweiterungen
  - Keine (Kopf-)Prüfsumme → UDP-Prüfsumme jetzt zwingend
  - Keine Hop-by-Hop-Fragmentierung
    → Nur Ende-zu-Ende-Fragmentierung plus Path-MTU-Discovery

Neuerungen in IPv6

- ICMPv6 und Neighbor Discovery
  - Zuvor getrennte Protokolle direkt in ICMP integriert
    - Adressauflösung (ARP) und Gruppenverwaltung (IGMP)
  - Erkennung doppelter Adressen und Detektion von Ausfällen
  - Erkennen des nächsten Routers sowie des Netzwerk-Präfixes
- Automatische Systemkonfiguration
  - Zustandslos: Präfix über Router Advertisements plus EUI-64-Interface-ID
    - Stateless Autoconfguration
  - Zustandsbehaftet: traditionelles DHCP (Dynamic Host Configuration Protocol)
- Bessere Unterstützung mobiler Systeme (MobileIPv6)
  - Bewegungserkennung und Adressenzuweisung durch automatische Systemkonfiguration
  - Die Option Binding Update im Destination-Options-Header ermöglicht die direkte Umleitung der IP-Dateneinheiten an den aktuellen Standort
- Und vieles mehr …

Neuerungen in IPv6

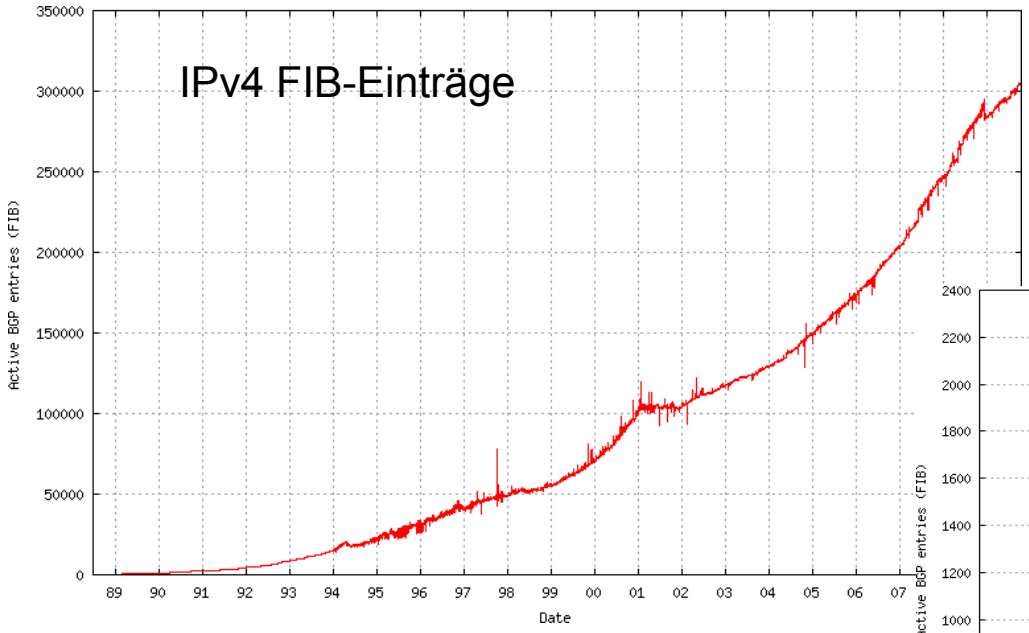


- Allerdings auch Nachteile
  - Zusatzaufwand durch Header nun mindestens 40 Bytes (IPv4: 20 Bytes)
    - z.B. IP-Telefonie: 20% statt 11% Overhead bei 8 Bit/8 KHz unkomprimiert
  - Adressen noch weniger handhabbar → DNS zwingend
  - Nicht abwärtskompatibel (IPv4 ist keine IPv6-Variante)

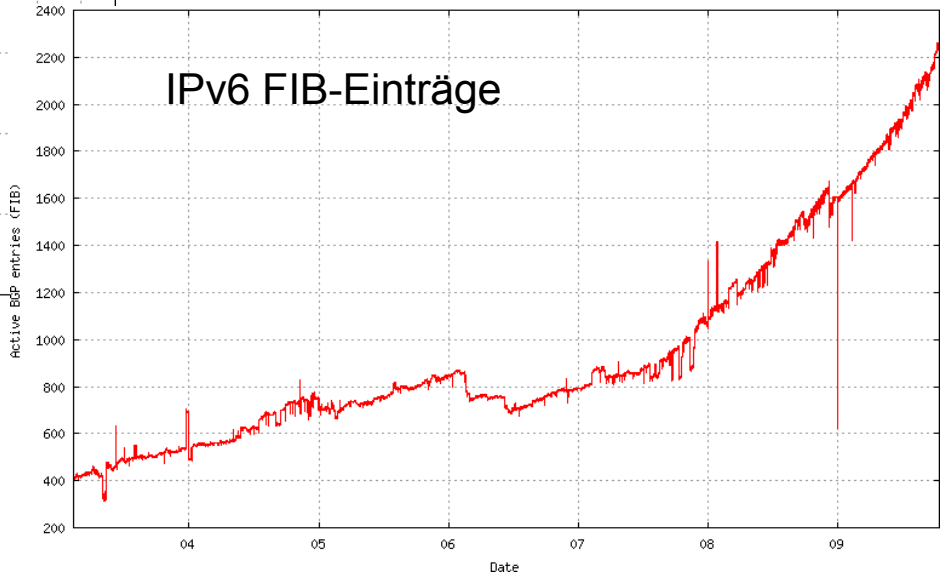
Aktueller Stand der IPv6-Verbreitung

- Seit 1998 bereits *Draft-Standard*

[DeHi98]



Stand: 12.10.2009



- Aktive BGP Einträge für IPv4- und IPv6-Adresspräfixe
  - 305173 IPv4 FIB-Einträge
  - 2262 IPv6 FIB-Einträge
- Ausführliche Informationen zu NATs und IPv6 in Vorlesung Next Generation Internet

3.5 Literatur

Bücher

[BeGa91] D. Bertsekas, R. Gallager; Data Networks; Prentice-Hall, 2nd Edition, 1991
Kapitel 5

[HaPh00] S. Halabi, D. McPherson; Internet Routing Architectures; Cisco Press, 2nd Edition, 2000
Schwerpunkt Inter-Domain-Routing, praktischer Einsatz von BGP

[Huit00] C. Huitema; Routing in the Internet; Prentice-Hall, 2nd Edition, 2000
Guter Überblick über verschiedene Routing-Protokolle

[Kesh97] S. Keshav; An Engineering Approach to Computer Networking; Addison-Wesley, 1997
Kapitel 11: Routing

[Kuro07] J. Kurose; Computer Networking; Addison-Wesley, 4th Edition, 2007
Kapitel 4: Network Layer and Routing

[Stal06] W. Stallings; Data & Computer Communications, Pearson Prentice Hall, 8th Edition, 2006
Kapitel 12

RFCs / Drafts

[BaAt97] F. Baker, R. Atkinson; RIP-2 MD5 Authentication; IETF, RFC 2082, Jan 1997

[DeHi98] S. Deering, R. Hinden; Internet Protocol, Version 6 (IPv6) Specification; IETF, RFC 2460, Dez 1998

[Hedr88] C. Hedrick; Routing Information Protocol; IETF, RFC 1058, Jun 1988

[Hust04] G. Huston; NOPEER Community for Border Gateway Protocol (BGP) Route Scope Control; IETF, RFC 3765, Apr 2004;

Literatur

[Malk98] G. Malkin; RIP Version 2; IETF, RFC 2453, Nov 1998

[MeZF07] D. Meyer, L. Zhang, K. Fall (Eds.); Report from the IAB Workshop on Routing and Adressing; IETF, RFC 4989, Sep 2007

[Moy98] J. Moy; OSPF Version 2; IETF, RFC 2328, Apr 1998

[SrEg01] P. Srisuresh, K. Evegang; Traditional IP Network Address Translator (Traditional NAT); IETF, RFC 3022, Jan 2001

[Vill98] C. Villamizar; Route Flap Damping; IETF, RFC 2439, Nov 1998

Vertiefende Literatur

[Grif05] T. Griffin; Metarouting; ACM SIGCOMM 2005; Philadelphia, USA; http://www.sigcomm.org/sigcomm2005/paper-GriSob.pdf

[KhZi89] A. Khanna, J. Zinky; The Revised ARPANET Routing Metric; ACM Computer Communication Review, Vol 19, Issue 4, Sep 1989, pp. 45-56

[McFR78] J. McQuillan, G. Falk, I. Richter; A review of the Development and Performance of the ARPANET Routing Algorithm; IEEE Trans. on Communications, Vol. 26, Issue 12, Dec 1978, pp. 1802-1810

[McRR80] J. McQuillan, I. Richter, E. Rosen; An Overview of the new routing algorithm for the ARPANET; ACM Computer Communication Review, Vol. 25, Issue 1, Jan 1995, pp. 54-60 (orig. published Nov 1979)

Literatur

[McRR80a] J. McQuillan, I. Richter, E. Rosen; The new Routing Algorithm for the ARPANET; IEEE Trans. on Communications, Vol. 28, Issue 5, May 1980, pp. 711-719

[SaRe84] J.H. Saltzer, D.P. Reed, and D.D. Clark; End-to-End Arguments in System Design; ACM Trans. On Computer Systems, Vol 2, Number 4, November 1984, pp 277-288

[Shen05] S. Shenker; HLP: A Next Generation Inter-Domain Routing Protocol; SIGCOMM 2005; Philadelphia, USA; http://www.sigcomm.org/sigcomm2005/paper-SubCae.pdf

Internet-Links

[Cisc05] Cisco; Enhanced Interior Gateway Routing Protocol; Sep 2005; http://www.cisco.com/warp/public/103/eigrp-toc.pdf

[HoWa06] Ch. Hopps, D. Ward; IS-IS for IP Internets; IETF, Working Group isis; http://www.ietf.org/html.charters/isis-charter.html

[Hust05] G. Huston; An Operational Perspective on BGP Security; IETF 63; Aug 2005; http://www.ietf.org/old/2009/proceedings/05aug/slides/grow-2.pdf

[Hust05a] G. Huston; Securing Inter-Domain Routing; The ISP Column; March 2005; http://www.potaroo.net/papers/isoc/2005-03/route-sec-2-ispcol.pdf

[Hust06] G. Huston; Aggregation Reports; http://bgp.potaroo.net/as1221/bgp-active.html

[WaAt05] D. Ward, A. Atlas; IP Fast Reroute: Overview and Things We Are Struggling to Solve; NANOG Feb 2005; Las Vegas USA; http://www.nanog.org/mtg-0501/ward.html

3.6 Aufgaben

1) Was ist ein Distanz-Vektor?
2) Wie funktioniert der Link-State-Algorithmus?
3) Schicht 3 des Internet-Referenzmodells ist zwar die Vermittlungsschicht, aber arbeiten auch Routing-Protokolle auf dieser Schicht?
4) Nennen Sie zwei Intra-Domain- (IGP) und ein Inter-Domain-Routingprotokoll (EGP)
5) Wie unterscheiden sich Intra- und Inter-Domain-Routingprotokolle?
6) Skizzieren Sie die Funktionsweise von BGP
7) Wie funktioniert in Zusammenhang mit BGP das CIDR-Verfahren?
8) Nennen Sie Probleme und Herausforderungen von BGP
9) Warum sind Internet-Routing-Tabellen meist sehr groß und was sind Gründe für deren schnelles Wachstum?