Technische Universität Chemnitz

– Lehrmaterial –

Versuch QoS and/or fair Queuing

© Jan Horbach

25. November 2001

Inhaltsverzeichnis

Lernziele:
Ziel dieses Versuches ist es, die Möglichkeiten des Linux-Kerns hinsichtlich Quality of Service kennenzulernen und anzuwenden. Dabei sollen die verschiedenen Verfahren vorgestellt und praktisch eingesetzt werden.

1. Versuchsvorbereitung

1.1. Was ist QoS?

Der Begriff **Quality of Service (QoS)** besitzt viele verschiedene Bedeutungen und wird von verschiedenen Seiten unterschiedlich benutzt, was zu einiger Verwirrung geführt hat. Deshalb wurde 1997 auf dem Next Generation Internet (NGI) Workshop in Virginia eine (sehr allgemeine) Definition geprägt: QoS differenziert Traffic und Services, d.h. verschiedene Classes of Service (CoS) werden unterschiedlich behandelt.

Man stelle sich z.B. folgendes Szenario vor: In einem Campusnetz teilen sich eine Vielzahl von Studenten eine relativ begrenzte Netzanbindung. Zu bestimmten Zeiten wollen so viele Studenten das Netz nutzen, dass ein vernünftiges Arbeiten nicht mehr möglich ist. Ein Ansatz wäre die Anschaffung neuer Netztechnik, um die verfügbare Bandbreite zu erhöhen. Das könnte sogar soweit gehen, das Netz zu "over-engineeren", d.h. eine Bandbreite zur Verfügung zu stellen, die mindestens so groß wie die Summe der Bandbreiten pro Nutzer ist. Aus Kostengründen wird das aber so gut wie nie gemacht, sondern reale Netze sind in aller Regel "over-subscribed", wobei die Bandbreite nur für eine durchschnittliche Nutzung ausreichend ist. Falls das Netz nun aber stärker belastet wird als vorgesehen, befinden wir uns wieder bei der Ausgangssituation. Der Ausweg hieraus besteht aus einer differenzierten Behandlung von verschiedenen Traffic-Klassen. In unserem Campusnetz könnte z.B. Traffic, der von Forschung &Lehre herrührt, der Vorzug gegenüber dem diverser Netzwerkspiele gegeben werden.

Durch QoS kann z.B. auch garantiert werden, dass wichtige Management-Daten auch in einer Überlastsituation noch ankommen (indem sie bevorzugt behandelt werden), oder dass interaktive Verbindungen eine mehr oder weniger garantierte Antwortzeit bekommen, während das für simple Filetransfers nicht zwingend notwendig ist.

Was versteht man eigentlich unter den Begriffen Quality und Service?

Quality:

- verlässliche Datenlieferung, kein Datenverlust
- minimale Verzögerung

- konstante Verzögerungscharakteristika (z.B. konst. Jitter = Varianz zwischen min. und max. Verzögerung)
- effizienteste Nutzung von Netzwerkressourcen (kürzeste Entfernung usw.)

Service:

- End-zu-End-Kommunikation oder Client-Server-Applikationen
- von E-Mail bis Desktop Video, von Surfen bis Chatten
- Protokollumgebungen (IP, IPX, AppleTalk u.a.)
- hierarchisch (z.B. SAP-Typen innerhalb von IPX)

Classes of Service (CoS):

Um die verschiedenen Services unterschiedlich behandeln zu können, teilt man sie in Serviceklassen auf. Dabei werden sie z.B. nach folgenden Gesichtspunkten klassifiziert:

- Protokoll (IP, TCP, UDP, IPX, ...)
- Quell-/Zieladresse und/oder Quell-/Zielport (Flow = Adressen + Ports)
- Netzinterface der Quelle
- IP Precedence Bits (3 Bit) im TOS-Byte des IP-Headers
- IPv6: 4 Bit Priority/Class-Field + 24 Bit Flow-Label

Frage 1.1.1:
Was ist bei der Benutzung der IP Precedence Bits zur Klassifizierung von Datenströmen zu beachten, wenn die so klassifizierten Pakete ihren Weg durch das Netz nehmen? Denken Sie dabei daran, dass die Pakete unter Umständen viele verschiedene Netzabschnitte durchlaufen, die durch Router voneinander getrennt sind, die die Precedence Bits verschieden interpretieren oder meist gar nicht beachten.

Probleme bei der Umsetzung von QoS

- Netzwerkcharakteristiken müssen vorhersagbar bleiben
- z.B. RTT (Round-Trip Time) → wichtig für TCP-Timeouts
- Filtering bevorzugt an Endpunkten → belastet Netzwerkkern nicht so sehr
- aber: dadurch auch leicht manipulierbar (keine Kontrolle während des Transfers)
- sehr eng mit Routing verbunden → evtl. dort einbauen

- aber: dynamisches Routing in Abhängigkeit von Last noch sehr unausgereift
- Problem wegen Asymmetrie (unterschiedliche Wege für Hin- und Rückweg)
- QoS-Maßnahmen machen sich meist nur in Überlastsituationen bemerkbar
- Messungen verfälschen Ergebnis zusätzlich

Vertiefung:

Paul Ferguson, Geoff Huston:

"Quality of Service - Delivering QoS on the Internet and in Corporate Networks"
http://www.wiley.com/compbooks/catalog/24358-2.htm

Wiley Computer Publishing, 1998

Mario Lorenz: "Geordnete Wege - Traffic Control mit Linux", iX 4/2000, Verlag Heinz Heise GmbH &Co KG

Internet Protocol - Quality of Service Page
http://qos.ittc.ukans.edu/

RFC 2549 "IP over Avian Carriers with Quality of Service"
http://www.ietf.org/rfc/rfc2549.txt

1.2. Grundlegende Verfahren

Dieser Abschnitt soll einen Überblick über existierende Verfahren bieten, von denen aber nur einige wirklich für die Versuche benötigt werden.

Admission Control

Admission Control legt fest, welcher Traffic im Netz überhaupt erlaubt ist. Festgelegt wird das über Policies. Es gibt zwei Ausprägungen: die passive Admission Control, in der es den Endnutzern überlassen wird, die Policies festzulegen, und die aktive Admission Control, in der das im **Ingress-Router** (d.h. dem Router, der die Daten von einem LAN in ein größeres Netz einspeist), realisiert wird.

Traffic Shaping

Traffic Shaping hat die Aufgabe, die Menge und Rate des Traffics zu kontrollieren, der in das Netz eintritt. Ziel dabei ist es, zumindest einen Anschein von vorhersagbarem Verhalten zu erhalten.

Leaky Bucket:

- Traffic-Bursts werden zwischengepuffert und in konstanten Raten wieder abgegeben
- wenn der Puffer voll ist, kommt es zu einem Überlauf und neu ankommende Pakete werden weggeschmissen
- Implementierung in der IP- oder z.B. in der ATM-Schicht
- Nachteil: wenn genug Bandbreite vorhanden ist, wird der Traffic unnötig eingeschränkt

Token Bucket:

- der Bucket wird hier statt mit Daten vom System in bestimmten Abständen mit Tokens versorgt
- jedes Token repräsentiert eine gewisse Menge von Datenbytes (festlegbar)
- Daten können nur gesendet werden (auch in Bursts), wenn genügend Tokens im Bucket sind
- zusätzlich ist eine maximale Burst-Größe festlegbar

Eine Kombination aus Token Bucket und Leaky Bucket wird verwendet, um zu verhindern, dass ein Traffic-Flow die ganze Bandbreite für sich allein beansprucht.

Preferential Queuing

Priority Queuing:

- höherpriorisierte Pakete werden vor niedrigpriorisierteren in die Output-Queue eingeordnet
- langsamer im Vergleich zum FIFO Queuing (da jedes Paket analysiert und ggf. umgeordnet werden muss)
- im schlimmsten Fall kommen niedrigpriorisierte Pakete nie bzw. "zu spät" an

Class-Based Queuing (CBQ):

- eine Variante von Priority Queuing mit mehreren Output-Queues
- Bandbreite wird hierarchisch (z.B. als Baum) auf Klassen je nach Bedarf verteilt
- Fairness: jede Klasse wird behandelt, auch niedrigpriorisierte Pakete kommen "pünktlich" an
- in der Praxis wird das durch die unterschiedliche Vergabe von Ressourcen geregelt
- verringerte Latenz gegenüber Priority Queuing, skaliert aber auch noch nicht gut

Weighted Fair Queuing (WFQ):

- Traffic mit niedriger Datenmenge wird bevorzugt behandelt
- Traffic mit höherer Datenmenge kann nicht sämtliche Ressourcen für sich beanspruchen
- jeder Traffic-Flow wird nach Datenmenge in die entsprechende Queue eingeordnet
- skaliert ebenfalls nicht (Rechenoverhead), außerdem zu statisch (wenig beeinflussbar)

Stochastic Fair Queuing (SFQ):

- Datenströme werden durch eine Hashbildung über Quell- und Zieladresse auf mehrere Queues verteilt, die gleichmäßig entleert werden (Round Robin)
- die Hash-Funktion wird in bestimmten Zeitintervallen geändert, um Hash-Kollisionen (verschiedene Verbindungen teilen sich dieselbe Warteschlange) zu minimieren
- einfacher als WFQ, aber nicht so leistungsfähig
- optimal in Verbindung mit CBQ einsetzbar

Clark-Shenker-Zhang (CSZ):

- ähnlich CBQ, aber präziser, dafür aber auch weniger flexibel und weniger effizient
- stellt wirklich garantierte Services zur Verfügung (garantierte Verzögerungen und Jitter)
- für jeden garantierten Service werden WFQ-Ströme erzeugt
- Rest der Bandbreite wird Dummystrom zugewiesen

Selective Forwarding

Das Grundprinzip des **Selective Forwarding** ist es, wichtige Daten auf einem schnellen Pfad zu befördern, unwichtigere Daten jedoch auf einem langsameren. Eine wichtige Rolle dabei spielen eine deterministische Pfadauswahl mit geringster Latenz, Jitter-Kontrolle u.a. Dazu gibt es eine Reihe von Verfahren (die aber selten eingesetzt werden):

TOS Routing:

- Routing in Abhängigkeit vom TOS-Byte:
- Bit 0-2: Routine, Priority, Immediate, Flash, Flash Override, CRITIC/ECP, Internetwork Control, Network Control
- Bit 3: Normal/Low Delay
- Bit 4: Normal/High Throughput
- Bit 5: Normal/High Reliability

Routing Information Protocol (RIP) und **Open Shortest Path First (OSPF)**:

- OSPF: kürzeste Wege nach Dijkstra (mit QoS im TOS-Byte)
- kompliziert im Vergleich zum zielbasierten Forwarding
- Ausweg: Policy-Based Routing (abhängig von Quelle statt Ziel), aber zu langsam

QoS Routing (QoSR):

- Absicht, QoS direkt in das Routing zu integrieren
- Probleme: asymmetrisches Routing (Hin- und Rückweg unterschiedlich), Overhead

Multi-Protocol Label Switching (MPLS):

- wird auch als Layer 2.5 bezeichnet (zwischen den Layern 2 und 3)
- das Routing erfolgt basierend auf mitgeführten Labels (20 Bit)
- bei ATM werden die Labels in VP/VC umgesetzt
- IP Precedence Bits als CoS in Label aufgenommen (3 Bit)

TCP-Mechanismen

TCP hat die Eigenschaft, in einer Überlastsituation, d.h. wenn Pakete des Datenstroms verloren gehen, die Datenrate herunterzufahren, sprich: die Pakete langsamer zu senden. Es existieren zwei Modi: der Slow-Start-Modus, bei dem die Datenrate nach dem Herunterfahren so weit wie möglich verdoppelt wird, und den Congestion-Avoidance-Modus, bei dem das Ansteigen linear erfolgt. Aber was passiert nun, wenn viele TCP-Ströme auf der Leitung fließen? Alle entdecken in etwa gleichzeitig eine Überlastsituation und alle gehen gleichzeitig in den Slow-Start-Modus. Während dieser Zeit ist die Leitung so gut wie frei (wenn man von anderen Protokollen absieht), es entsteht also ein Leerlauf. Die TCP-Ströme verdoppeln nun ihre Rate so lange, bis sie wieder eine Überlastsituation feststellen usw. Um dieser Fluktuation vorzubeugen, wurden folgende Verfahren entwickelt:

Random Early Detection/Drop (RED):

- verhindert den o.a. Überlastkollaps durch zufälliges Wegwerfen von Paketen (abhängig vom Füllstand der Queue)
- die Wahrscheinlichkeit eines Wegwerfens (Drops) ist festlegbar
- Ziel: Vermeidung der Situation, dass alle TCP-Flows gleichzeitig Überlast entdecken und in den Slow-Start-Modus gehen
- Effekt: TCP-Flows werden zu verschiedenen Zeitpunkten langsamer, Leerlauf und evtl. sogar eine echte Überlastung werden verhindert
- ist sehr effizient und fair, d.h. alle Daten werden gleich behandelt (keine CoS)
- Problem: UDP-Daten werden nicht erfasst, kommen aber im Multimedia-Bereich sehr oft vor

Weighted RED (WRED):

- auch als **Guaranteed Rate I/O (GRIO)** bezeichnet
- je höher die Precedence, desto geringer ist die Wahrscheinlichkeit eines Drops
- aktive Admission Control: z.B. über Token-Bucket-Schwellen (Tresholds)
- wenn Schwelle überschritten ->niedrigere Precedence (Treshold Triggering)

Generalized RED (GRED):

- mehrere Drop-Precedences festlegbar
- mehrere Drop-Wahrscheinlichkeiten, jeder Queue zugeordnet

Integrated Services (IntServ)

Integrated Services (IntServ) besitzt drei primäre Ziele:

1. die Services genau zu definieren
2. Application Service (End-To-End Ansprüche), Router Scheduling (welche Informationen sollen den einzelnen Routern verfügbar gemacht werden) und Link-Layer Interfaces ("Subnetze") zu definieren
3. Router Validation zu entwickeln, um sicherzustellen, dass Services auch zur Verfügung gestellt werden ->keine erhöhten zusätzlichen Anforderungen an Router

QoS tritt hier in einer ähnlichen Bedeutung auf wie bei ATM (erreichte Bandbreite, Paketverzögerung, Paketverlustraten usw.), findet aber im Gegensatz zu ATM in OSI-Layer 3 statt in Layer 2 statt. Es existieren 5 Schlüsselkomponenten:

1. QoS-Anforderungen: Serviceklassen und Traffic Control (Packet Scheduler, Classifier, Admission Control, Resource Reservation)
2. Resource-Sharing-Anforderungen: Link-Sharings, Weighted Fair Queuing (WFQ)
3. Erlaubnis, bestimmte Pakete wegzuschmeißen (Pakete, die nicht der Admission Control unterliegen oder durch Flags)
4. Festlegungen für Usage Feedback (Accounting Data)
5. **Resource Reservation Protocol (RSVP)**: dynamische QoS-Anforderungen

Differentiated Services (DiffServ)

Bei den **Differentiated Services (DiffServ)** existieren verschiedene **Queuing Disciplines (QDisc)** (z.B. TBF und RED), die an Serviceklassen gebunden werden. Die Klassen sind in einer Baumstruktur abgelegt, wobei sich die Kinder von den Eltern bei Bedarf Bandbreite borgen können, wenn dort noch Kapazität vorhanden ist. Im Gegenzug können die Kinder nicht benötigte Bandbreite den Eltern zur Verfügung stellen. Den Blättern im Baum kann wieder eine QDisc zugewiesen werden, was DiffServ ziemlich leistungsfähig macht, denn so können verschiedene CoS durch unterschiedliche Algorithmen behandelt werden. Die Klassifikation erfolgt über Filter: generische Filter (z.B. routingbasiert) oder spezielle Filter (z.B. RSVP- und U32-Classifier).

Die Festlegung der Parameter geschieht meist nur in den Endpunkten oder beim Provider. Um die Klassifizierung der Flows mitzuführen, wird das **TOS-Byte** verwendet, was hier allerdings als **Differentiated Services Code Point (DSCP)** bezeichnet wird. In

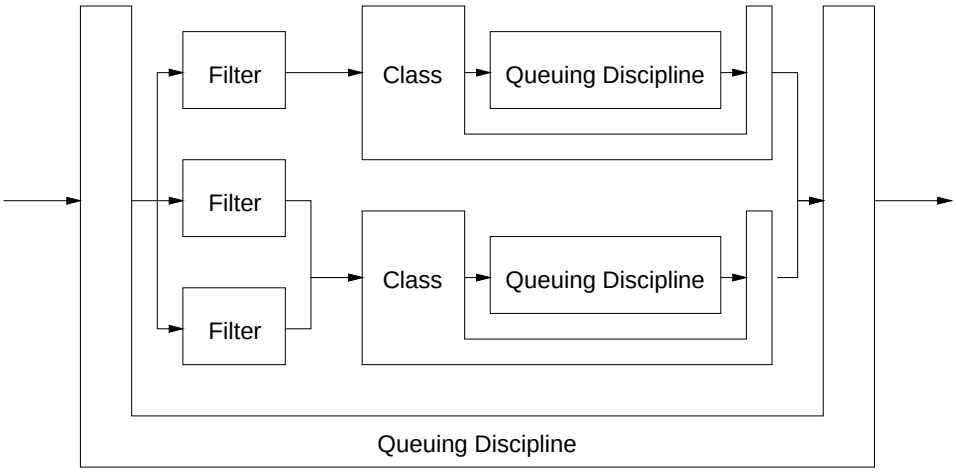

Abbildung 1.2-1.: Queuing Discipline mit mehreren Klassen

den Routern wird dann meist nur noch der DSCP ausgewertet, die Daten entsprechend behandelt und die Flows am Ende evtl. neu markiert, d.h. der DSCP angepasst. Linux unterstützt die Behandlung von DSCP's erst ab dem Kernel 2.3.x, weswegen sie für die Versuche noch keine Rolle spielen.

Die DSCP-Werte stammen aus drei Pools: Pool 1 (xxxxx0) ist für die Standard-Aktionen reserviert, Pool 2 (xxxx11) und Pool 3 (xxxx01) sind zu experimentellen Zwecken oder lokalem Gebrauch gedacht, wobei aber Pool 3 auch irgendwann zur Ergänzung von Pool 1 herangezogen werden kann. Der DSCP-Wert "xxx000" dient als Class-Selector CP, der Standard-CP ist "000000".

Man unterscheidet drei grundlegende Serviceklassen:

- **Assured Forwarding (AF)**: gesicherte Verbindungen, die in 4 verschiedene Klassen eingeteilt werden, von denen jede in unterschiedlichen Dropwahrscheinlichkeiten (z.B. über RED) vorkommt; z.B. für bevorzugte Kunden

	AF-Klasse 1	**AF-Klasse 2**	**AF-Klasse 3**	**AF-Klasse 4**
Niedrige Dropwkt.	001010 (0x0a)	010010 (0x12)	011010 (0x1a)	100010 (0x22)
Mittlere Dropwkt.	001100 (0x0c)	010100 (0x14)	011100 (0x1c)	100100 (0x24)
Hohe Dropwkt.	001110 (0x0e)	010110 (0x16)	011110 (0x1e)	100110 (0x26)

- **Expedited Forwarding (EF)**: ein so genannter Premium Service; z.B. für normale Kunden, mit DSCP = 101100 (0x2e)
- **Best Effort (BE)** bzw. **Default (DE)**: "herkömmliche" Verbindungen, in denen sich unwichtigerer Traffic die verbliebene Bandbreite nach dem Best-Effort-Prinzip teilt

Vergleicht man IntServ und DiffServ, stellt man fest, dass DiffServ wesentlich besser skaliert, was vor allem daran liegt, dass die Klassifikation der Datenströme an den Netzwerkrändern vorgenommen wird (in der Regel vom Provider) und in den Routern nur noch der DSCP ausgewertet werden muss. Als Nachteil ergibt sich daraus natürlich die fehlende Dynamik von DiffServ, da das Setup statisch erfolgt, wohingegen bei IntServ über RSVP eine dynamische Anpassung möglich ist.

Vertiefung:

References on CBQ (Class-Based Queueing)
http://www.aciri.org/floyd/cbq.html

References on RED (Random Early Detection) Queue Management
http://www.aciri.org/floyd/red.html

Differentiated Services on Linux
http://icawww1.epfl.ch/linux-diffserv/

IP Quality of Service: An Overview
http://www.ittc.ukans.edu/~rsarav/ipqos/ip_qos.htm

A Study of IP QoS
http://www.ittc.ukans.edu/~iyer/Mainpage.html

QoS Routing (qosr) - Charter
http://www.ietf.org/html.charters/qosr-charter.html

Integrated Services (intserv) - Charter
http://www.ietf.org/html.charters/intserv-charter.html

Resource Reservation Setup Protocol (rsvp) - Charter
http://www.ietf.org/html.charters/rsvp-charter.html

Differentiated Services (diffserv) - Charter
http://www.ietf.org/html.charters/diffserv-charter.html

RFC 1633 "Integrated Services in the Internet Architecture: An Overview"
http://www.ietf.org/rfc/rfc1633.txt

RFC 2205 "Resource ReSerVation Protocol (RSVP) - Version 1 Functional Specification"
http://www.ietf.org/rfc/rfc2205.txt

RFC 2212 "Specification of Guaranteed Quality of Service"
http://www.ietf.org/rfc/rfc2212.txt

RFC 2474 "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers"
http://www.ietf.org/rfc/rfc2474.txt

RFC 2676 "QoS Routing Mechanisms and OSPF Extensions"
http://www.ietf.org/rfc/rfc2676.txt

IETF Internet Draft "A Framework for Multiprotocol Label Switching"
http://www.ietf.org/internet-drafts/draft-ietf-mpls-framework-05.txt

1.3. Realisierung in Linux

Der Traffic Shaper

Seit dem 2.0-er Kern existiert der Traffic Shaper, mit dem ein zusätzliches Device angelegt werden konnte, dessen Bandbreite dann explizit beschränkt wurde. Um den Netzverkehr über dieses Device zu leiten, mussten nur noch die Routen entsprechend gesetzt werden:

```
shapecfg attach shaper0 DEVICE
shapecfg speed shaper0 BANDWIDTH
ifconfig shaper0 HOST netmask NETMASK broadcast BROADCAST up
route add -net NETWORK netmask NETMASK dev shaper0
```

Dieses Verfahren war sehr einfach zu handhaben, aber auch sehr beschränkt in seiner Leistungsfähigkeit, so dass ab den späten 2.1-er Kernels *QoS and/or fair queueing* in den Kern integriert wurde und der Traffic Shaper kaum noch benutzt wird.

TC - Traffic Control

Im Paket IPRoute2 ist ein Tool zur Verwaltung von Queuing Disciplines, Klassen und Filtern enthalten: **Traffic Control (TC)**. Um neben der Benutzung dieser Applikation auch von eigenen Programmen auf QoS zurückgreifen zu können, wird eine API entwickelt. Generell muss QoS-Support im Kernel eincompiliert oder als Modul verfügbar sein (*Networking Options —>QoS and/or fair queueing*). Die benötigten Algorithmen sind ebenfalls als Modul oder in den Kernel eincompiliert auszuwählen. Weiterhin gibt es noch die Optionen *QoS/Rate Estimator*, der für QDiscs (z.B. CBQ) nötig ist, die die benutzte Bandbreite von Datenströmen abschätzen müssen, und die verschiedenen Classifier, die ausgewählt werden müssen, wenn die Klassifizierung nicht nur über das TOS-Byte erfolgen soll. Wenn man die QDiscs oder die Classifier als Module compiliert hat, sind sie explizit mit `modprobe sch_`*`qdisc`* bzw. `modprobe cls_`*`classifier`* zu laden, bevor sie verwendet werden können.

Nun soll die allgemeine Vorgehensweise zum Einrichten von QoS mittels TC erläutert werden: Zuerst muss an das gewünschte Netzinterface eine Queuing Discipline gebunden werden, nach der der Traffic (nur der ausgehende!) statt dem Einordnen in eine Standard-FIFO behandelt werden soll. Falls es sich bei dieser QDisc um CBQ (oder auch DSMark und ATM) handelt, sind nun noch Serviceklassen und die zugehörigen Classifier (Filter) zu definieren. Für die anderen QDiscs entfallen diese Schritte, hier wird der gesamte Traffic, der das Interface verlässt, nach der zugewiesenen Queuing Discipline behandelt. Im Falle von CBQ sind den Klassen (in der Regel den Blättern

des Klassenbaums) wieder QDiscs zuzuweisen, nach denen der Traffic dieser Klasse behandelt werden soll. Unterbleibt dieser Schritt, wird wieder eine Standard-FIFO dafür benutzt. Mit den Filtern wird durch eine Auswertung bestimmter Paketinformationen festgelegt, welche Traffic-Arten welchen Klassen zugeordnet werden.

> Hinweis!
> Für die Versuche steht eine vereinfachte Variante von TC zur Verfügung. Diese Übersicht ist hauptsächlich als Kommandoreferenz für weitergehende Anwendungen von QoS zu verstehen.

Zuweisung der QDisc

Mit folgendem Aufruf von TC wird eine QDisc einem Interface zugewiesen:

```
tc qdisc add dev DEVICE [ handle X: ] { root | parent CLASSID }
            [ estimator INTERVAL TIME_CONSTANT ] QDISC_KIND OPTIONS

  dev        - Netzdevice
  handle     - Handle der QDisc (eindeutige ID) der Form X:, z.B. 1:
  root       - Bindung bezieht sich direkt auf das Interface
  parent     - Bindung bezieht sich auf Parent mit CLASSID
  estimator  - Steuerung des Rate Estimator (Zeitraum zwischen Messungen
               und Zeitkonstante zur Mittelwertbildung in Sekunden)
  QDISC_KIND = { cbq | csz | dsmark | [p|b]fifo | prio | [g]red |
                 sfq | tbf | teql }
```

Folgende QDiscs sind implementiert:

- CBQ (Class-Based Queuing):

```
tc qdisc add ... cbq bandwidth BPS avpkt BYTES [ mpu BYTES ]
                 [ cell BYTES ] [ ewma LOG ]

  bandwidth - reale Bandbreite des Interfaces
  avpkt/mpu - durchschnittliche/minimale Paketgröße
  cell      - Anzahl Bytes, in der Transferzeit gemessen wird
  ewma      - Exponentially Weighted Moving-Average (Beeinflussung
              aktueller durch alte Werte bei Messungen)
```

- CSZ (Clark-Shenker-Zhang): als defekt gekennzeichnet, wird wahrscheinlich ersetzt

- DSMark (Differentiated Services Field Marker, ab Kernel 2.3.x): ruft Classifier und speichert zurückgelieferte ClassID als TC-Index, sonst den Default-Index

```
tc qdisc add ... dsmark indices NUMBER [ default NUMBER ]
                 [ set_tc_index ]

  indices      - Anzahl Einträge in der Maske-Wert-Tabelle
  default      - Default-Index
  set_tc_index - DS-Byte soll als TC-Index gespeichert werden
```

- PFIFO (Packet FIFO) und BFIFO (Byte FIFO):

```
tc qdisc add ... [p|b]fifo limit NUMBER

  limit - Größe der Queue
```

- PRIO (Priority): bis zu 16 Unterwarteschlangen ("Bänder") mit unterschiedlicher Priorität (Band 0 die höchste); Default-Scheduler von Linux nutzt 3 Bänder

```
tc qdisc add ... prio bands NUMBER priomap P1 P2

  bands   - Anzahl der Bänder
  priomap - für Zuordnung der Pakete in die Bänder
```

- RED (Random Early Detection) und GRED (Generalized RED, ab Kernel 2.3.x):

```
tc qdisc add ... red limit BYTES min BYTES max BYTES avpkt BYTES
                 burst PACKETS probability PROB bandwidth KBPS
tc qdisc add ... gred setup DPs NUMBER default NUMBER [ grio ]
tc qdisc add ... gred limit BYTES min BYTES max BYTES avpkt BYTES
                 burst PACKETS probability PROB bandwidth KBPS
                 DP NUMBER prio NUMBER

  min/max     - minimale/maximale Paketgröße
  burst       - erlaubte Anzahl von Paketen in einem Burst
  probability - Drop-Wahrscheinlichkeit (Default: 2%)
  DPs         - Anzahl Drop-Precedences
  default     - Default Drop-Precedence
  grio        - GRIO (WRED) Buffer-Sharing
  DP          - Zuordnung zu einer Drop-Precedence
  prio        - Priorität
```

- SFQ (Stochastic Fair Queuing): maximal 128 Warteschlangen

```
tc qdisc add ... sfq perturb SECS quantum BYTES

  perturb - Zeitintervall, nachdem Hash-Funktion geändert wird
  quantum - Bytes, die pro Runde übertragen werden
```

- TBF (Token Bucket Filter):

```
tc qdisc add ... tbf limit BYTES buffer BYTES[/BYTES] rate KBPS
                 [ mtu BYTES[/BYTES] ] [ peakrate KBPS ]
                 [ latency TIME ]

  limit        - Größe des TBF
  buffer/burst - Schwelle, bis zu der Bursts gesendet werden dürfen
  mtu          - Maximum Transfer Unit
  rate         - Transferrate
  peakrate     - maximale Transferrate
  latency      - Latenzzeit
```

- TEQL (True Link Equalizer): Daten über ein logisches Interface (teqlN) auf mehrere physikalische Interfaces verteilt (je nach Bandbreite der Interfaces)

```
modprobe sch_teql
tc qdisc add dev DEVICE_1 root TEQL_DEV
 :
tc qdisc add dev DEVICE_N root TEQL_DEV
ifconfig TEQL_DEV HOST [ netmask NETMASK broadcast BROADCAST ] up
route add ... TEQL_DEV
```

Ab dem Kernel 2.3.x existieren auch noch eine Ingress QDisc für ankommenden Traffic und eine ATM VC Selection QDisc.

Klassenhierarchie aufbauen

Falls als QDisc CBQ, DSMark oder ATM verwendet wurde, werden im zweiten Schritt die Klassen für die verschiedenen Traffic-Arten eingerichtet und in einem Baum angeordnet, indem beim Anlegen einer Klasse immer der Parent-Knoten mit angegeben wird. Innherhalb des CBQ-Klassenbaums kann wie bei DiffServ Bandbreite zwischen den einzelnen Klassen verborgt werden. Die Ratenangaben beim Anlegen der Klassen dienen hierbei nur zur Aufteilung der Bandbreite, eine explizite Begrenzung muss durch TBF erfolgen.

```
tc class add dev DEVICE classid CLASSID { root | parent CLASSID }
             QDISC_KIND OPTIONS

  classid    - ID der Form X:Y (X = ID der QDisc, Y = ID der Klasse),
               Root-Klasse hat X:0, andere z.B. X:1, X:2, ...
  QDISC_KIND = { atm | cbq | dsmark } (andere QDiscs sind klassenlos)
```

- CBQ (Class-Based Queuing): siehe auch DiffServ im vorigen Abschnitt

```
tc class add ... cbq bandwidth BPS rate BPS [ avpkt BYTES ]
                 [ mpu BYTES ] [ allot BYTES ] [ weight RATE ]
                 maxburst PACKETS [ minburst PACKETS ]
                 [ prio NUMBER ] [ cell BYTES ] [ ewma LOG ]
                 [ bounded ] [ isolated ]
                 [ estimator INTERVAL TIME_CONSTANT ]
                 [ split CLASSID ] [ defmap MASK/CHANGE ]

  rate      - zugewiesene Bandbreite
  maxburst  - maximale Anzahl von Paketen in einem Burst
  minburst  - minimale Anzahl von Paketen in einem Burst
  bounded   - Klasse darf sich keine Bandbreite borgen
  isolated  - Klasse teilt sich keine Bandbreite mit Nicht-Kindern
  allot     - MTU + Größe des MAC-Headers
  weight    - Gewicht der Klasse (optional, proportional zu "rate")
  prio      - Priorität der Klasse (zwischen 1 und 8, 1 = höchste)
  cell      - Anzahl Bytes, in der Transferzeit gemessen wird
  split     - eingesetzter Classifier gilt nur für Klassen mit
              entsprechendem Split-Wert
  defmap    - bei ungematchten Paketen wird TOS-Byte mit Maske
              ausgewertet
```

- DSMark (Differentiated Services Field Marker, ab Kernel 2.3.x): ändert den DSCP

```
tc class change ... dsmark mask MASK value VALUE

  mask + value - Berechnung des DSCP: DSCP = (DSCP &  MASK) | VALUE
```

Klassifizierung der Pakete

Schlussendlich müssen nur noch die einzelnen Pakete ihren Klassen (auch als Flows bezeichnet) zugeordnet werden, indem bestimmte Informationen in den Paketen ausgewertet werden. Die Filterregeln werden der Reihe nach durchgegangen, bis eine der Regeln zutrifft, so dass der Traffic einer Klasse zugeordnet werden kann. Ansonsten wird er als Best-Effort-Traffic behandelt.

```
tc filter add dev DEVICE [ handle FILTERID ] { root | parent CLASSID }
              [ prio PRIO ] [ protocol PROTO ]
              [ estimator INTERVAL TIME_CONSTANT ] FILTER_TYPE OPTIONS

  parent      - Handle der QDisc
  handle      - Handle des Filters (siehe die einzelnen Classifier)
  prio/pref   - Priorität des Filters
  PROTO       = { ip | ... }
  FILTER_TYPE = { route | fw | u32 | rsvp[6] | tcindex }
```

Folgende Klassifizierungsmöglichkeiten stehen zur Auswahl:

- über das Routing-Subsystem: Mit `ip route` wird eine Route gesetzt und dieser eine Nummer zugewiesen, die dann vom Filter an einen Flow, d.h. an ein Blatt im Klassenbaum, gebunden wird.

  ```
  ip route add NET_ADDRESS [ via GATEWAY ] dev DEVICE realms REALM
  tc filter add...route [ from REALM | fromdev DEVICE | fromif TAG ]
                  [ to REALM ] [ flowid CLASSID ] [ police POLICE ]

    flowid - Blatt im Klassenbaum
    POLICE = rate BPS buffer BYTES[/BYTES] burst BYTES[/BYTES]
    (s. TBF) [ mpu BYTES[/BYTES] ] [ mtu BYTES[/BYTES] ]
             [ peakrate BPS ] [ avrate BPS ] [ index NUMBER ]
             [ ACTION ]
    ACTION = { reclassify | drop | continue }
  ```

- über das Firewall-Subsystem: Mit `ipchains` wird eine Firewallregel erzeugt und mit einer Nummer markiert, die der Filter dann wieder an den Flow binden kann (über das Handle!).

  ```
  ipchains -A output -s NET_ADDRESS -m NUMBER
  tc filter add ... handle NUMBER fw [ flowid CLASSID ]
                    [ police POLICE ]
  ```

- über den U32-Classifier: Mit den Match-Anweisungen wird der Paketkopf ausgewertet, und durch die FlowID geschieht die Zuordnung zu einer Klasse. Ein Match-Statement bezieht sich auf ein Byte (u8), ein Wort (u16) oder ein Doppelwort (u32) mit entsprechender Bitmaske (max. 0xFF, 0xFFFF oder 0xFFFFFFFF bzw. relevante Bits bei IP-Adressen).

```
tc filter add ... [ handle X:Y:Z ] u32 [ match SELECTOR ]
                  flowid CLASSID [ offset mask MASK shift NUMBER ]
                  [ hashkey mask MASK at NUMBER ] [ link HTID ]
                  [ ht HTID ] [ police POLICE ] [ sample SAMPLE ]
tc filter add ... handle X: u32 divisor DIVISOR

  divisor  - Hashtable mit DIVISOR Slots und Handle X: erzeugen
  ht       - Zuordnung zu Y-tem Slot der Hashtable X (HTID=X:Y:)
  offset   - Offset in der Hashtable
  hashkey  - Key in der Hashtable
  SELECTOR = SAMPLE SAMPLE ...
  SAMPLE   = { ip | ip6 | udp | tcp | icmp | u{8|16|32} }
             FIELD [ TYPE MASK ]
```

	FIELD	**TYPE**	**MASK**	**Beschreibung**
ip	src	<ipaddr>/BITS		Quelladresse
	dst	<ipaddr>/BITS		Zieladresse
	tos, dsfield,	<u8>	0xFF	TOS-Byte, DSCP bzw.
	precedence	<u8>	0xFF	IP-Precedence
	nofrag,			Bits zur
	firstfrag,			Steuerung der
	df, mf			Paketfragmentierung
	ihl	<u8>	0xFF	Länge des Headers
	protocol	<u8>	0xFF	Layer4-Protokolltyp
	sport	<u16>	0xFFFF	Layer4-Quellport
	dport	<u16>	0xFFFF	Layer4-Zielport
	icmp_type	<u8>	0xFF	ICMP-Typ
	icmp_code	<u8>	0xFF	ICMP-Code
udp	src	<u16>	0xFFFF	UDP-Quellport
	dst	<u16>	0xFFFF	UDP-Zielport
tcp	src	<u16>	0xFFFF	TCP-Quellport
	dst	<u16>	0xFFFF	TCP-Zielport
icmp	type	<u8>	0xFF	ICMP-Typ
	code	<u8>	0xFF	ICMP-Code

- über den RSVP-Classifier: Hier kann die Bandbreite dynamisch mittels RSVP je nach Anforderung vergeben werden (Bandwidth on Demand):

```
tc filter add ... handle X:Y rsvp ipproto PROTOCOL session
                  DST[/PORT | GPI ] [ sender SRC[/PORT | GPI ] ]
                  [ classid CLASSID ] [ police POLICE ]
                  [ tunnelid ID ] [ tunnel ID skip NUMBER ]

  ipproto - Protokoll
  GPI     - Generalized Port Identifier:
            { flowlabel NUMBER | spi/ah SPI | spi/esp SPI |
              u{8|16|32} NUMBER mask MASK at OFFSET }
  SPI     - Source Port ID
```

- über den TC-Index-Classifier (ab Kernel 2.3.x):

```
tc filter add ... [ handle DSCP ] tcindex [ classid CLASSID ]
                  [ mask MASK ] [ shift NUMBER ]
                  [ pass_on ] [ fall_through ]

  mask + shift - Keyberechnung: KEY = (TC-Index » NUMBER) &  MASK
  pass_on      - wenn handle != DSCP, dann nächsten Filter suchen
  fall_through - versucht, eine neue Klasse zu erzeugen, wenn es
                 keine zu KEY passende gibt
```

Eingerichtete QoS anzeigen

```
tc [ -s ] qdisc  { show | ls } dev DEVICE [ ingress ]
tc [ -s ] class  { show | ls } dev DEVICE [ root | parent CLASSID ]
tc        filter { show | ls } dev DEVICE [ root | parent CLASSID ]

  -s - statistische Angaben (Anzahl behandelter Pakete etc.)
```

Vertiefung:

Linux - Advanced Networking Overview
http://qos.ittc.ukans.edu/howto/

Linux 2.4 Advanced Routing HOWTO
http://www.linuxdoc.org/HOWTO/Adv-Routing-HOWTO.html

Linux QoS Support
http://www.ittc.ukans.edu/~rsarav/projects/networking/ipqos/diffoverview/

Differentiated Services on Linux
ftp://lrcftp.epfl.ch/pub/linux/diffserv/misc/dsid-01.txt

IP QoS Efforts
http://qos.ittc.ukans.edu/slides/

An API for Linux QoS Support
http://www.ittc.ukans.edu/~pramodh/courses/linux_qos/mainpage.html

Linux Iproute2, Traffic Control &Friends
http://defiant.coinet.com/iproute2/

README zu IPRoute2 und TC
file:///usr/doc/packages/iproute/README.iproute2+tc

Traffic Shaper For Linux
file:///usr/src/linux/Documentation/networking/shaper.txt

Linux-Kernelquellen
file:///usr/src/linux/net/sched/

2. Versuchsumgebung

2.1. Versuchsaufbau

Für die Versuche stehen vier Rechner zur Verfügung, die drei miteinander verbundene Netzwerke bilden. Der Rechner *athos* fungiert als Router, der die Netze verbindet.

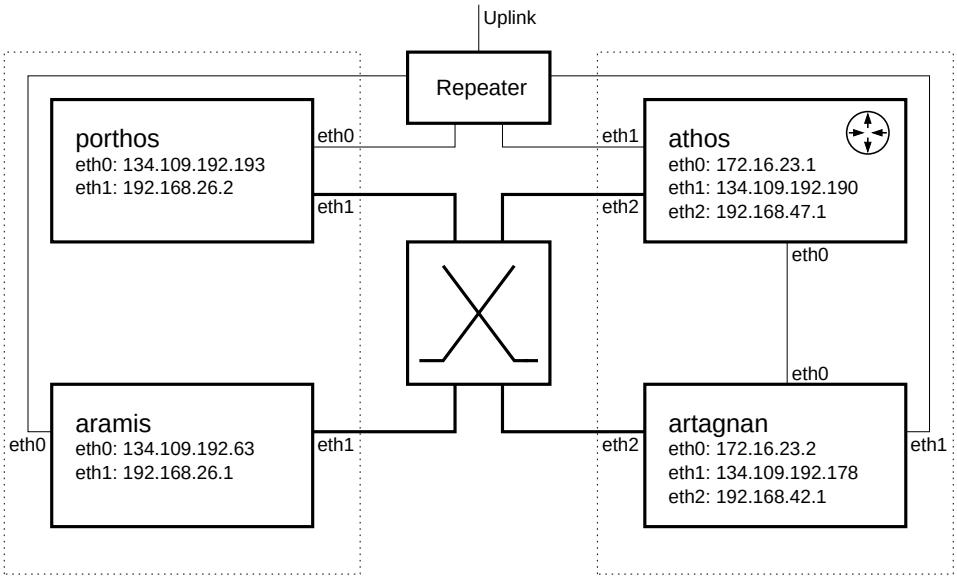

Abbildung 2.1-1.: Versuchsaufbau

Gigabit-Netz

Wie aus dem Gigabit-Versuch noch bekannt sein sollte, besteht das Gigabit-Netz aus drei VLAN's: *aramis* und *porthos* mit der Netzadresse 192.168.26/24, *artagnan* mit 192.168.42/24 und *athos* mit 192.168.47/24. Als Default-Route ist der Switch mit 192.168.xx.254 eingetragen, wobei xx die Nummer des Subnetzes darstellt. Der Switch ist so konfiguriert, dass er Pakete defaultmäßig an *athos* schickt, wenn sie eine unbekannte Zieladresse enthalten.Für diesen Versuch spielt das Gigabit-Netz zwar keine Rolle,

aber der Aufbau ist dennoch hier enthalten, um die Routingeinstellungen verstehen zu können.

100MBit-Netz

Alle Rechner besitzen auch eine offizielle IP aus dem 134.109.192/24-er Netz der Informatik und sind für den Versuch mit einem 10/100MBit-Repeater verbunden. Der besitzt zwar einen Uplink nach außen, dieser sollte aber bei der Versuchsdurchführung gekappt werden, um nicht das gesamte Subnetz zu beeinträchtigen.

100MBit-Direktverbindung

Zusätzlich existiert noch eine 100MBit-Direktverbindung zwischen *athos* und *artagnan* mit der Netzadresse 172.16.23/24.

Hinweis!
Da das Labor hauptsächlich für das *Gigabit Ethernet Praktikum* gedacht ist, müssen vor Beginn der Versuche noch einige Einstellungen vorgenommen werden: Auf allen Rechnern ist die Default-Route zu löschen und auf *athos* zu setzen, d.h. `route add default gw 172.16.23.1 eth0` für *artagnan* und `route add default gw 134.109.192.190 eth0` bei den beiden anderen Rechnern. Zusätzlich ist auf *artagnan* noch die Route ins 134.109.192/24-er Netz zu löschen, damit die Pakete auf dem Rückweg auch über den Router laufen. Der Aufruf von `routes set` auf einem der Rechner erledigt das für Sie. Um mit den Rechnern arbeiten zu können, benötigen Sie natürlich noch das Root-Passwort, es lautet: `Praktikum`. Zwischen den Rechnern steht Ihnen allerdings eine passwortlose SSH zur Verfügung.

2.2. Zur Verfügung stehende Software

Da die Handhabung des TC-Kommandos sehr komplex ist, gibt es für diesen Versuch eine vereinfachte Variante namens **Traffic Control Changer (TCC)**, der auch gleich die benötigten Module mit lädt und folgende Syntax besitzt:

```
tcc cbq    DEVICE HANDLE { root | PARENTID } BANDWIDTH
tcc red    DEVICE HANDLE { root | PARENTID } BANDWIDTH PROBABILITY
tcc sfq    DEVICE HANDLE { root | PARENTID }
tcc tbf    DEVICE HANDLE { root | PARENTID } RATE
tcc teql   DEVICE
tcc class  DEVICE CLASSID PARENTID BANDWIDTH RATE WEIGHT PRIO \
           [ bounded ] [ isolated ]
tcc filter DEVICE FLOWID PARENTID match SELECTOR
tcc remove DEVICE
tcc list   DEVICE [ -s ]
```

- Die ersten vier Aufrufe binden eine Queuing Discipline an ein Interface (`root`) bzw. an ein Blatt im Klassenbaum mit der ClassID PARENTID. Bei TEQL muss nur das Netzdevice angegeben werden.
- Durch `tcc class` wird eine Klasse mit der ClassID CLASSID und der Elternklasse PARENTID angelegt; handelt es sich um die oberste Klasse, ist als PARENTID X:0 anzugeben, wobei X: das Handle der QDisc darstellt.
- BANDWIDTH ist immer die reale Bandbreite des Interfaces DEVICE, RATE die gewünschte (natürlich nur kleiner oder gleich BANDWIDTH) und WEIGHT das zu RATE proportionale Gewicht (z.B. 1/10 von RATE).
- Mit `tcc filter` wird über einen SELECTOR der Traffic, der das Interface verlässt, der Klasse mit der ClassID FLOWID zugeordnet, PARENTID ist in dem Fall das Handle der QDisc (also X:0).
- Die restlichen Parameter sollten nach der Lektüre der TC-Syntax selbsterklärend sein. Für die hier nicht angegebenen TC-Parameter werden durch den TCC Standardwerte gesetzt.

Hinweis!
Da die einzugebenden Kommandos immer noch recht umfangreich sind und auch leicht Fehler auftreten können, empfiehlt es sich, sie in einem Shell-Skript zu "verewigen", ehe man sie ausführt. Fügen Sie dem Protokoll bitte immer den Quelltext ihrer Skripte hinzu!

Ein einfaches Beispiel, um UDP-Traffic auf 1MBit zu begrenzen, könnte etwa so aussehen:

```
# QDisc erzeugen und an das Interface binden
tcc cbq    eth0 1:  root 100MBit

# Root-Klasse mit der ID 1:1 anlegen
tcc class  eth0 1:1 1:0  100MBit 100MBit   10MBit 8

# Klasse fuer Rest ohne QDisc und Filter
tcc class  eth0 1:2 1:1  100MBit  99MBit 9900KBit 5

# Klasse fuer UDP mit QDisc TBF und Filter (Protokoll UDP)
# erst durch TBF wird die Bandbreite endgueltig beschraenkt
tcc class  eth0 1:3 1:1  100MBit   1MBit  100KBit 5 bounded
tcc tbf    eth0 2:  1:3            1MBit
tcc filter eth0 1:3 1:0  match ip protocol 17 0xff
```

Wollte man statt UDP den gesamten ausgehenden Traffic eines bestimmten Rechners oder Netzes begrenzen (z.B. im Router), so muss man lediglich einen anderen Filter verwenden. In diesem Beispiel wird der gesamte Traffic des CSN-Subnetzes der V54 begrenzt.

```
tcc filter eth0 1:3 1:0  match ip src 134.109.96.0/22
```

Falls man diese Maßnahme im Router vornimmt, kann man durch eine analoge Vorgehensweise auf dem anderen Interface auch den eingehenden Traffic des Subnetzes begrenzen (in Wirklichkeit wird ja der in Richtung Subnetz ausgehende Traffic im Router begrenzt).

```
...
tcc filter eth1 3:3 3:0  match ip dst 134.109.96.0/22
```

Hinweis!
Zwischen den Versuchen ist auf allen benutzten Interfaces ein `tcc remove DEVICE` auszuführen und mit `tcc list DEVICE` zu überprüfen, ob auch alle QDiscs entfernt wurden. Falls das nicht der Fall sein sollte (z.B. durch Fehleingaben verursacht), ist es am einfachsten, den Rechner neu zu starten.

Um die vorgenommenen Änderungen zu messen bzw. anderweitig zu überprüfen, gibt es folgende Möglichkeiten:

FTP und SCP

Die beiden Programme zeigen die Übertragungsrate während des Dateitransfers oder danach an. Dateien zum Transfer liegen unter `/root/prak_files/`. Bei SCP können Sie mit der Option `-c blowfish` veranlassen, dass der etwas schnellere Verschlüsselungsalgorithmus *Blowfish* verwendet wird, was den Durchsatz merklich erhöht.

NetPIPE und GnuPlot

Mit NetPIPE wird eine TCP-Messreihe gestartet, bei der im Ping-Pong-Verfahren Datenblöcke gesendet werden, deren Größe stetig erhöht wird. Unter `/root/prak_messungen/` werden die Ergebnisse in Files abgelegt, die mit GnuPlot ausgewertet werden können. Um eine Messreihe zu starten, muss auf dem Empfängerrechner `netpipe recv` aufgerufen werden, und auf dem Senderrechner `netpipe send`. Bei mehreren Sendern sind auch entsprechend viele Receiver zu starten und anhand der Portnummer zu unterscheiden:

```
netpipe send HOST PORT MESSFILE
netpipe recv PORT
```

> Hinweis!
> Die NetPIPE-Messreihen nehmen relativ viel Zeit in Anspruch: ca. 5min bei 100MBit/s. Diese Zeit gilt für **jede** Messreihe; auch wenn mehrere parallel durchgeführt werden, addieren sich die Zeiten, da sich dann die einzelnen Datenströme behindern.

Zur grafischen Auswertung ist nach dem Starten von GnuPlot mittels `gnuplot` am dortigen Kommandoprompt folgendes einzugeben, um den Durchsatz (2. Spalte im File) der Messreihe über der Blockgröße (4. Spalte) abzutragen:

```
load "/root/praktikum/ds.gnu"
plot "MESSFILE1" using 4:2 w l,"MESSFILE2" using 4:2 w l, ...
```

TCP-Dump und Ethereal

Mit `tcpdump -i INTERFACE` können Sie den Traffic auf einem Netzwerkinterface überwachen. Durch die Option `-p` ist zu verhindern, dass die Karte in den Promiscous-Modus gesetzt wird, damit Sie wirklich nur den Traffic des betreffenden Rechners "erwischen". Falls Sie mit Grep bestimmte Informationen herausfiltern wollen, müssen Sie zusätzlich noch die Option `-l` angeben oder die eingebauten Filter benutzen. Das Tool Ethereal bietet eine grafische Oberfläche zur Netzwerkanalyse und -diagnose.

IfConfig

Damit können Sie u.a. feststellen, wieviele Pakete auf den einzelnen Interfaces empfangen und gesendet wurden, am besten in Verbindung mit Watch: `watch -n INTERVAL ifconfig`.

Logfiles

Falls Sie Einsicht in die Logfiles benötigen... diese liegen unter unter `/var/log/` und `/var/log/httpd/`.

Screenshots

Screenshots von grafischen Ausgaben können z.B. mit XV gemacht werden. Dazu einfach auf den Grab-Button klicken und das gewünschte Fenster auswählen.

3. Versuchsdurchführung

3.1. Versuch: Verschiedene Klassen mit CBQ

Szenario

Der Rechner *porthos* führt mittels `ping -f -s 8000 172.16.23.2` eine Flood-Ping-Attacke gegen *artagnan* durch. Von dem anderen Rechner versucht ein Nutzer mittels SCP ein File auf den angegriffenen Rechner zu kopieren.

Frage 3.1.1:
Was ist während der Attacke zu erwarten? Begründen Sie ihre Vermutungen und vergleichen Sie sie mit den praktischen Ergebnissen!

Aufgabe

Starten Sie 4 SCP-Filetransfers von *aramis* zu 172.16.23.2:

1. ohne QoS-Maßnahmen und ohne Attacke (der "Idealfall")
2. ohne QoS-Maßnahmen und mit Attacke (der "Angriffsfall")
3. mit QoS-Maßnahmen und ohne Attacke (um zu sehen, ob QoS zu Performance-Einbußen führt)
4. mit QoS-Maßnahmen und mit Attacke (der "Abwehrfall")

Begrenzen Sie dazu im Router die Bandbreite für ICMP und UDP mittels CBQ und TBF, indem Sie vom Beispiel im vorhergehenden Abschnitt ausgehen! Die Begrenzung soll dabei so ausgelegt sein, dass sich die Klassen für ICMP und UDP keine Bandbreite borgen dürfen. UDP wird hier zwar nicht genutzt, soll aber trotzdem ebenfalls begrenzt werden, da statt einem Flood-Ping z.B. auch eine bandbreitenintensive UDP-Anwendung denkbar wäre. Protokollnummern finden Sie in `/etc/protocols`. Falls Sie genügend Zeit haben, können Sie auch mit verschiedenen Paketgrößen bei `ping -f` experimentieren.

Frage 3.1.2:
Auf welchem Interface im Router müssen die QoS-Maßnahmen ergriffen werden? Begründen Sie die getroffene Wahl!

Welche Transferraten haben Sie in den 4 Fällen erhalten? Erklären Sie die Ergebnisse!

3.2. Versuch: Borgen von Bandbreite mit CBQ

Szenario

Die beiden Rechner *aramis* und *porthos* sollen das CSN darstellen, *athos* fungiert als Gateway, und *artagnan* (172.16.23.2) ist ein Computer außerhalb des Uninetzes. Die Verantwortlichen des CSN planen nun, Bandbreitenbeschränkungen einzuführen. Der Traffic soll dabei in folgende Klassen aufgeteilt werden: SSH für das interaktive Arbeiten und eine Klasse für den restlichen Traffic (der dann u.a. FTP umfasst). Das Borgen von nicht benötigter Bandbreite soll möglich sein.

Frage 3.2.1:
Wie würden Sie die Bandbreite auf die Klassen aufteilen (prinzipiell)? Mit welchen Prioritäten? Begründen Sie!

Aufgabe

Setzen Sie im Router die geplanten Beschränkungen um! Um alle FTP-Nutzer gleich zu behandeln, soll bei der Rest-Klasse außerdem die QDisc SFQ zum Einsatz kommen. Da die Verbindungen in beide Richtungen begrenzt werden sollen, müssen die QoS-Maßnahmen auf beiden Interfaces des Routers ergriffen werden (da nur der jeweils ausgehende Traffic begrenzt werden kann). Für jedes Interface muss der SSH-Traffic in 2 Klassen unterteilt werden (SSH-Client und SSH-Daemon), wobei die Klassifizierung durch Auswertung von Quell- bzw. Zielport 22 erfolgen kann, während der Rest-Traffic z.B. durch die Netzadresse 134.109.192.0/24 (je nach Transferrichtung als Quell- oder Zieladresse) richtig zugeordnet werden kann. Denken Sie aber in dem Fall daran, die Filterregel der Rest-Klasse als letztes zu definieren, da sie sonst immer zutreffen würde, ehe die SSH-Regeln zum Zuge kommen könnten.

Frage 3.2.2:
Warum benutzt man, um den Rest-Traffic zu klassifizieren, die Netzadresse des "CSN's" und nicht die des außerhalb liegenden Netzes?

Hinweis!
Wählen Sie als Bandbreite für die SSH-Klassen nur 10MBit, da die beiden Rechner es kaum schaffen, durch SCP mehr als diese auszulasten (wegen der Verschlüsselung). Durch ein `scp -c blowfish` kommen Sie nur auf unwesentlich mehr als 32MBit, und das ist der schnellste mögliche Algorithmus. Im realen CSN stehen natürlich wesentlich mehr Rechner zur Verfügung, so dass man dort auch eine höhere Bandbreite wählen kann, bei den Versuchen würden Sie aber sonst keine sinnvollen Resultate erhalten.

Führen Sie Datentransfers per SCP und/oder FTP von einem und von beiden "CSN"-Rechnern gleichzeitig durch, um ihre getroffenen Maßnahmen zu überprüfen. Lassen Sie dabei jede Klasse erst einmal einzeln "zu Wort kommen", um zu sehen, ob das Borgen funktioniert. Danach beanspruchen Sie bitte immer mindestens 3 Klassen, damit das Netz wenigstens annähernd ausgelastet ist. Ein interessanter Fall ist es z.B., die SSH-Klassen von *eth0* und die FTP-Klasse von *eth1* zu benutzen, testen Sie aber auch andere! Messen Sie die Transferraten und fassen Sie ihre Ergebnisse zusammen! Durch `tcc list INTERFACE -s` können Sie u.a. nachschauen, welche Klasse sich Bandbreite geborgt hat (Angabe hinter "borrowed").

3.3. Versuch: Vermeidung von Slow Starts mit RED

Szenario

Die Rechner *aramis* und *porthos* senden gleichzeitig einen TCP-Datenstrom an *artagnan* (172.16.23.2). Da die Verbindung damit überlastet ist, werden ständig Kollisionen festgestellt, was dazu führt, dass die TCP-Flows in den Slow-Start-Modus geraten. Als Ausweg soll hier die QDisc RED dienen.

Frage 3.3.1:
Was erwarten Sie von dieser Maßnahme?

Aufgabe

Führen Sie obige QoS-Maßnahme durch! Sinnvolle Ergebnisse lassen sich dabei aber nur in einem belasteten Netz erzielen. Initiieren Sie dazu bitte mittels `ping -f 172.16.23.2` ein Flood-Ping von *athos* aus.

Frage 3.3.2:
An welchen Stellen könnten Sie RED einsetzen, d.h. in welchen Rechnern bzw. mit oder ohne CBQ? Wie beurteilen Sie die einzelnen Varianten?

Wählen Sie eine oder mehrere Varianten aus und messen Sie mit FTP vor und nach dem Einsatz von RED die erreichte Bandbreite, indem Sie von beiden Rechnern gleichzeitig ein File zu *artagnan* kopieren. Ein guter Wert für die Drop-Wahrscheinlichkeit ist z.B. 0.2. Falls Sie genug Zeit finden sollten, können Sie aber auch noch andere testen. Vergleichen Sie die praktischen Ergebnisse mit Ihren Erwartungen und interpretieren Sie sie!

Hinweis!
Machen Sie möglichst mehrere Messungen, um auszuschließen, dass es sich nur um normale Schwankungen handelt, und bilden Sie jeweils den Mittelwert. Außerdem sollten die Files gleich groß sein, damit die Transfers gleich lang dauern!

3.4. Versuch: Kanalbündelung mit TEQL

Szenario

Der Rechner *artagnan* besitzt 3 Netzinterfaces, von denen 2 (*eth0* und *eth1*) gemeinsam genutzt werden sollen, um eine höhere Bandbreite zu erreichen.

Frage 3.4.1:
Welche anderen Möglichkeiten fallen Ihnen ein, um beide Leitungen gleichzeitig nutzen zu können (weniger in Endknoten, sondern eher in großen Routern eingesetzt)?

Aufgabe

Benutzen Sie die Queuing Discipline TEQL, um den Traffic auf die beiden Interfaces zu verteilen! Nach dem Hinzufügen an beide Interfaces geben Sie dem Device *teql0* bitte die selbe IP wie einem der beteiligten Interfaces, aber mit einer Netzmaske von 255.255.255.255 und einer Broadcast-Adresse, die der IP entspricht. Löschen Sie die Routen auf die beteiligten Interfaces und leiten Sie allen Traffic über *teql0*!

Frage 3.4.2:
Warum wird man bei diesem Szenario TEQL eher auf dem Server als im Client einsetzen?

Rufen Sie von *athos* aus eine Webseite von *artagnan* ab und lauschen Sie auf *athos* mit `tcpdump -p -l -i INTERFACE | grep http` oder auch mittels `tcpdump -p -i INTERFACE port http` an beiden Interfaces, um zu sehen, welchen Weg die Webseite nimmt. Da TEQL manchmal erst in einer Überlastsituation zu greifen scheint, funktioniert das ab und zu nicht auf Anhieb. Setzen Sie dann bitte von *athos* aus das Netz mit `ping -f -s 16000 artagnan` unter Last und probieren Sie es noch einmal. Benutzen Sie auf *artagnan* `watch -n 2 ifconfig`, um festzustellen, wieviele Pakete auf welchen Interfaces empfangen und gesendet werden. Wenn Sie dann noch Zeit haben, können Sie auch noch eine NetPIPE-Messreihe starten (ohne Flood-Ping!), um zu sehen, ob sich die effektive Bandbreite erhöht hat. Optimal wäre es dann natürlich, wenn Sie auf *athos* ebenfalls TEQL einsetzen würden - mit einer bis auf die IP identischen Konfiguration. Was haben Sie für Resultate erhalten?