



5 Dienstgüte – Quality of Service

Motivation



- Multimediasysteme sollen insbesondere kontinuierliche Medien in Echtzeit erfassen, verarbeiten, übertragen und darstellen
- Die Dienstgüte (Quality of Service) besagt in welcher Güte dies geleistet wird
- Um eine gewisse Dienstgüte zu erbringen werden Betriebsmittel (Ressourcen) benötigt
- 2 grundlegende Ansätze
 - Skalierung
 - Anpassung des Datenflusses an die Gegebenheiten
 - Ressourcenreservierung
 - Betriebsmittel im gesamten Datenpfad werden reserviert



Ausgangssituation in Netzwerken



- herkömmliche Netz-Algorithmen und Protokolle zerstören isochronen Fluss der Paketströme
 - Sie erzeugen erhebliche Varianz in der Verzögerung (delay jitter)
- Dies gilt u.a. für
 - Netzzugangsprotokolle in LANs
 - z.B. CSMA/CD, TokenRing
 - Fehlersicherung durch Übertragungswiederholung
 - Flusskontrolle mit Schiebefenster (sliding window)
- „best effort“-Netze
- selten echte Unterstützung für Multicast



Dienstgüte



- Ein Dienstgüte-Modell besteht aus
 - Dienstgüte-Spezifikation
 - Dienstgüte-Kalkulation
 - Dienstgüte-Erbringung (QoS Enforcement)
- Dienstgüte beinhaltet verschiedenste Aspekte
 - Betriebssystem
 - Dateisystem
 - Kommunikationssystem
 - Kompression
 - Mediensynchronisation
 - User Interface Software
 - ...



Dienstgüevertrag



- Kontinuierliche Medien erfordern Dienstgütegarantien im Netz
- Idee: Dienstgüte-Vertrag
 - die Quelle spezifiziert den generierten Verkehr und verspricht, sich daran zu halten
 - das Netz verspricht die Übertragung mit garantierten Dienstgütemerkmalen



Echtzeitsystem vs. Multimediasystem



- Echtzeitsystem
 - DIN 44300:
„Echtzeitbetrieb ist ein Betrieb eines Rechensystems, bei dem Programme zur Verarbeitung anfallender Daten ständig derart betriebsbereit sind, dass die Verarbeitungsergebnisse innerhalb einer vorgegebenen Zeitspanne verfügbar sind“
- Echtzeitprozess
 - Ein Prozess der Resultate in einer vorgegebenen Zeitspanne liefert



Echtzeitsystem vs. Multimediasystem



- Deadline – Frist – Zeitschranke
 - Eine Deadline ist der letzte Zeitpunkt, zu dem das Resultat noch korrekt ist
 - Harte Deadline (hard real-time system)
 - Darf nie verletzt werden
 - Resultate, die zu spät kommen, sind wertlos
 - Deadlineverletzungen führen u.U. zu Schäden
 - Z.B. Kraftwerk, Space Shuttle, ...
 - Weiche Deadline (soft real-time system)
 - Darf etwas verfehlt werden
 - Resultate, die etwas zu spät kommen, sind noch akzeptabel
 - Z.B. Bahnfahrplan, Vorlesungen, und auch die meisten MM-Systeme



Echtzeitsystem vs. Multimediasystem



- Anforderungen an ein Echtzeitsystem
 - Deterministisch vorhersagbares Verhalten bzgl. der Spezifikation
 - Vorhersagbar schnelle Bearbeitung zeitkritischer Ereignisse
 - Hohes Maß an Planbarkeit (Schedulability)
 - Stabilität des System bei Überlast



Echtzeitsystem vs. Multimediasystem



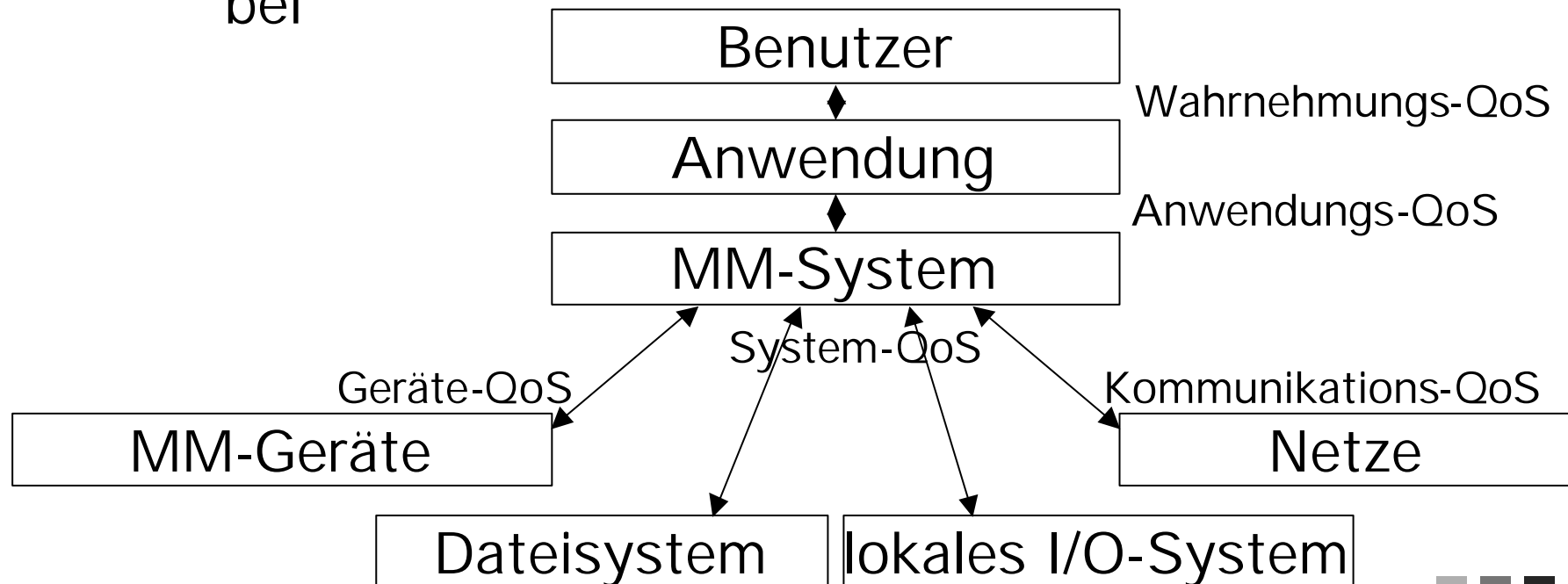
- Multimediasystem
 - Soft Real-time System
 - Charakteristika
 - Periodische Verarbeitungsprozesse
 - Große Bandbreiten
 - Ende-zu-Ende Garantien
 - Skalierung möglich
 - Kostenbasierte Fairness



Dienstgütemodell



- Dienstgüte kennzeichnet das definierte, kontrollierbare Verhalten eines Systems bezüglich quantitativ messbarer Parameter
 - Verschiedene Schichten tragen zur Dienstgüte bei



Dienstgüte-Parameter



- quantitative (funktionale) QoS-Parameter
 - Genauigkeit, Verzögerung, Fehlerrate, etc.
- qualitative (nicht-funktionale) QoS-Parameter
 - Zuverlässigkeit, Stabilität, Fehlertoleranz, Kosten, etc.
- Beispiele für Parameter
 - Wahrnehmungs-QoS
 - Visuelle Wahrnehmbarkeit
 - Akzeptierte Synchronisationsdrift
 - Anwendungs-QoS
 - Medienqualität, z.B. Fernsehqualität
 - Medieneigenschaften



Dienstgüte-Parameter



- Weitere Beispiele für Parameter
 - System-QoS
 - CPU Rate
 - Speicherbedarfe
 - Kommunikations-QoS
 - Paketgrößen und –raten
 - Bandbreite, Durchsatz
 - Verzögerung, Jitter
 - Paketverlust
 - Geräte-QoS
 - Samplerate
 - Auflösung



Dienste-Klassen



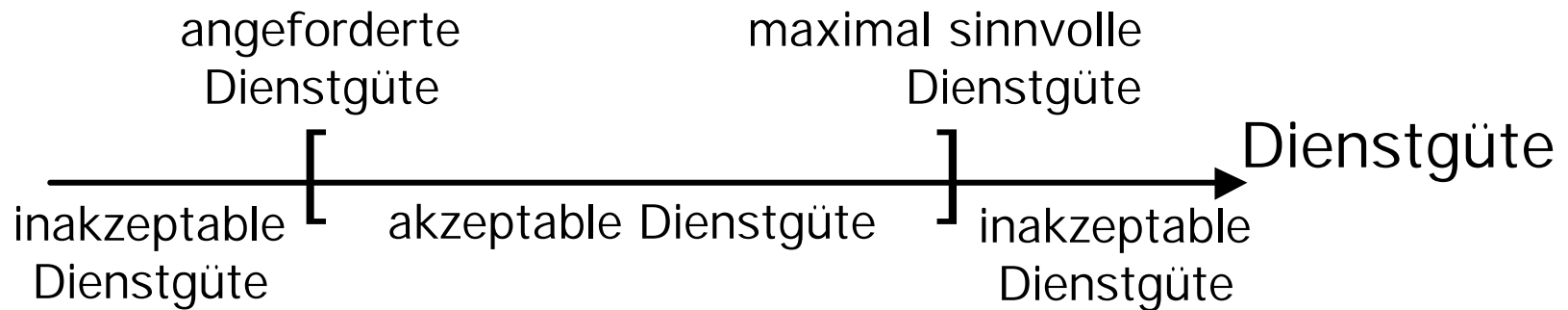
- Garantierte Dienste
 - Wert- oder Intervallangabe der QoS-Parameter
 - Deterministisch
 - $QoS = P$ bzw. $QoS_{min} \leq P \leq QoS_{max}$
 - Statistisch
 - Werte sind statistische Größen, z.B. Fehlerrate
- Vorhersagbare Dienste
 - Die Werte der QoS-Parameter beruhen auf "historisch" erhobenem vorherigen Verhalten
 - Der Dienst verspricht sich wie früher zu verhalten
 - Die Historie wird in Betracht gezogen entweder
 - Von Anfang an, oder
 - In einem gleitenden Zeitfenster
- Best-Effort Dienste
 - Der Dienst "tut so gut er kann"



QoS-Intervalle



- Die Werte der QoS-Parameter resultieren in
 - Akzeptablen Regionen
 - Inakzeptablen Regionen



- Dies kann auch multidimensional sein
 - Z.B. Framerate und Auflösung als Dimensionen bei Video



Betriebsmittel



- Betriebsmittel sind Systemkomponenten, die von Prozessen zur Bearbeitung bzw. zum Datentransfer benötigt werden
- Klassifikation nach Funktionalität
 - Aktive Betriebsmittel
 - Erfüllen ihre Aufgabe aktiv
 - Z.B. CPU, Netzadapter
 - Passive Betriebsmittel
 - Erfüllen ihre Aufgabe passiv, in dem sie einen Dienst anbieten
 - Z.B. Speicher, Frequenzspektrum



Betriebsmittel



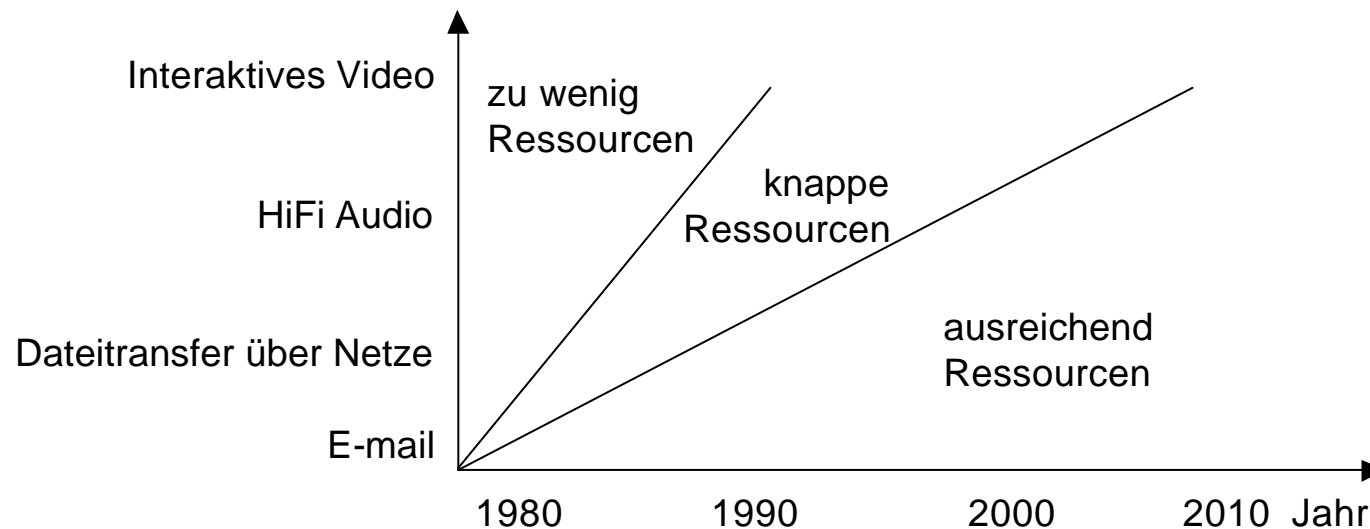
- Klassifikation nach Nutzung
 - exklusive Betriebsmittel
 - Zu einem Zeitpunkt kann nur ein Prozess das Betriebsmittel nutzen
 - Meist sind dies die aktiven Betriebsmittel
 - gemeinsame Betriebsmittel
 - Zu einem Zeitpunkt können mehrere Prozesse das Betriebsmittel nutzen
 - Meist sind dies die passiven Betriebsmittel
- Klassifikation nach Anzahl
 - Einfachbetriebsmittel
 - Im System kommt das Betriebsmittel nur einmal vor
 - Mehrfachbetriebsmittel
 - Im System kommt das Betriebsmittel mehrfach vor
- Alle Betriebsmittel verfügen über den quantitativen Parameter Kapazität



Verfügbarkeit von Betriebsmitteln



- Multimediasysteme mit AV-Verarbeitung erreichen oft die Kapazitätsgrenzen
 - Ziel ist es den besten Dienst (Service) bei möglichst niedrigen Kosten anzubieten
 - D.h. man muss Ressourcenmanagement betreiben



Beziehung QoS und Ressource



- Die Dienstgüte nach der Nutzung eines Betriebsmittels hängt von dessen Kapazität ab
 - Die Dienstgüte davor ist besser oder gleich der Dienstgüte danach



Netz-Ressourcen und QoS-Parameter



- Ressourcen im Netz beeinflussen QoS-Parameter
- Ein großer Puffer beim Empfänger (playout buffer) erlaubt die Kompensation einer höheren Varianz in der Verzögerung
 - aber auf Kosten einer größeren absoluten Verzögerung



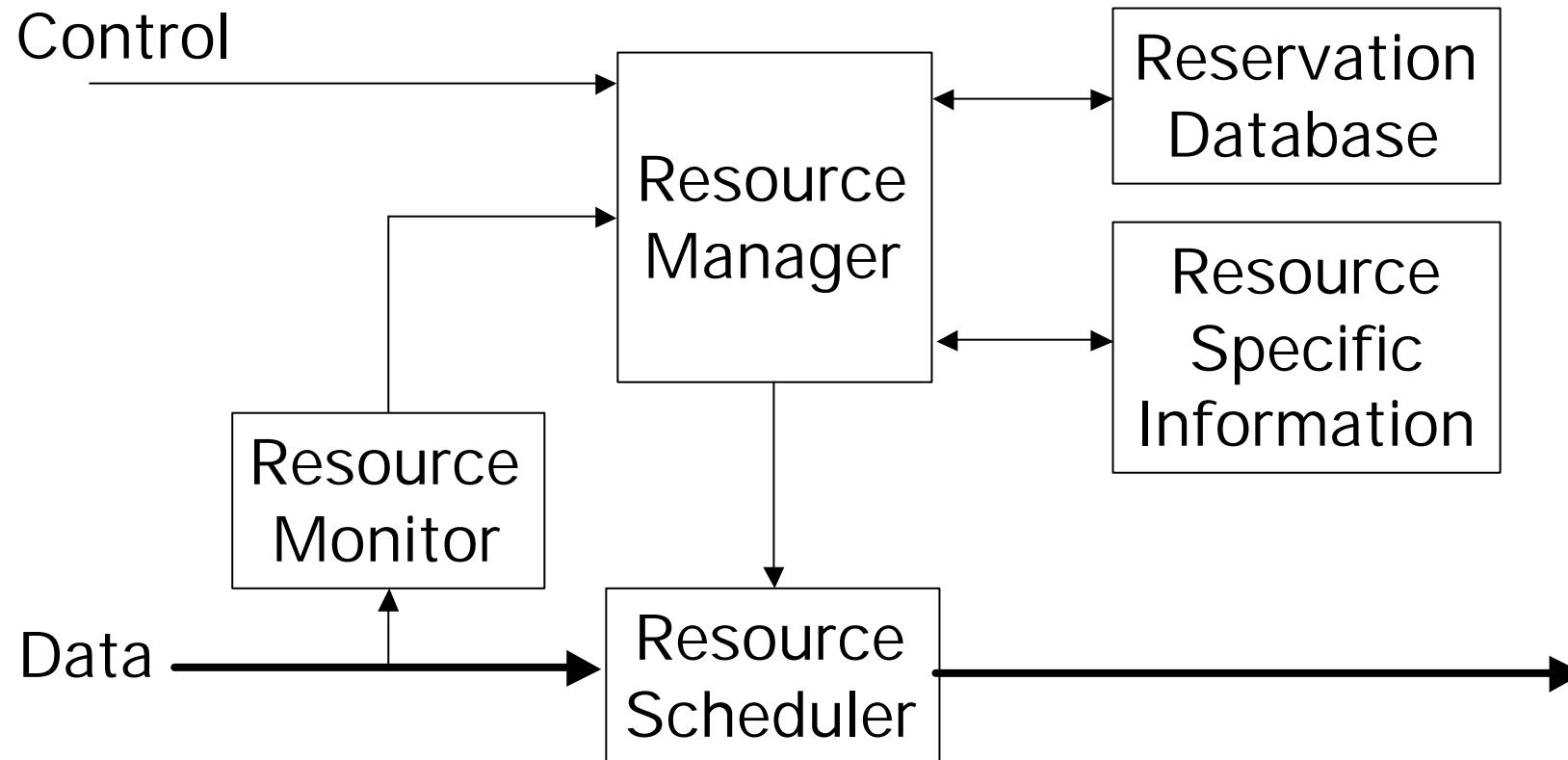
Netz-Ressourcen und QoS-Parameter



- bei zunehmender Pufferauslastung in einem Router steigt die mittlere Wartezeit der Pakete
 - neue Verbindungen nur dann akzeptieren, wenn für alle existierenden Verbindungen Verzögerungsgrenzen eingehalten werden können (Connection Acceptance Control)
- CPU-Leistung eines Routers entscheidet über die Maximalzahl und maximale Datenrate der gleichzeitigen Verbindungen



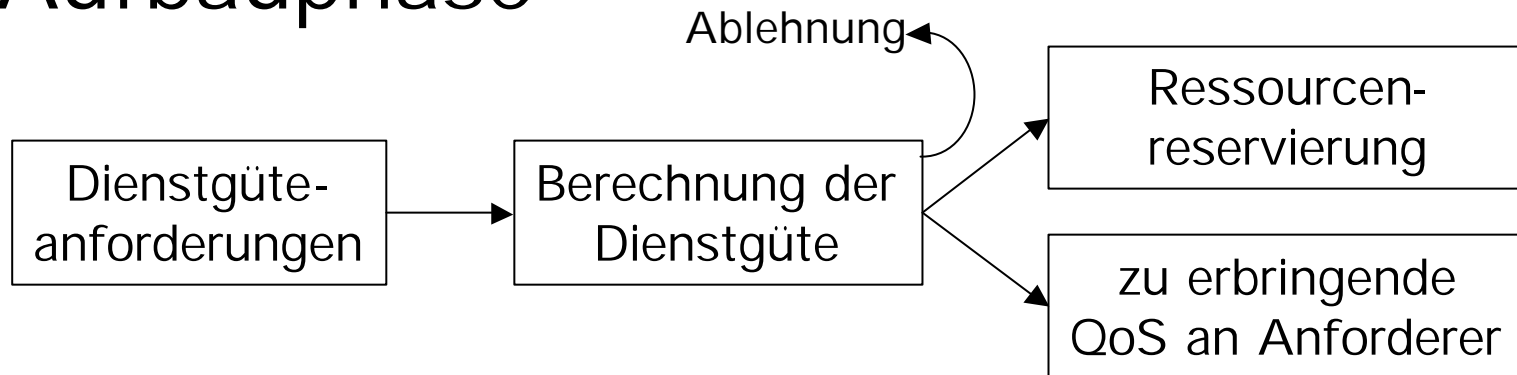
Ressourcenmanagement-Architektur



Phasen des Ressourcenmanagement



■ Aufbauphase



■ Datenbearbeitungsphase



Aufbauphase



- Definition der QoS-Parameter
 - Implizit oder explizit durch die Anwendung oder den Benutzer
- Parameter werden verteilt und ausgehandelt
- Parameter unterschiedlicher Schichten müssen übersetzt werden
 - Aufgrund unterschiedlicher Semantik bzw. Darstellung
- Abbildung der QoS-Parameter auf die Anforderungen der Betriebsmittel
- Die Betriebsmittel müssen auf dem Weg von Quellen zu Senken zugelassen, reserviert, belegt und koordiniert werden



QoS-Definition



- Verkehrsbeschreibung der Quelle
 - Verkehrsart
 - CBR, constant bit rate
 - VBR, variable bit rate
 - UBR, unspecified bit rate
 - ...
 - Stoßweises Verkehrsaufkommen (bursty)
 - mittlere Bitrate
 - maximale Bitrate
 - Gestalt der Spitzenlast



QoS-Definition



- Dienstgütemerkmale an der Netzschnittstelle
 - Verzögerung (delay)
 - Varianz der Verzögerung (delay jitter)
 - maximale Verlustrate (loss rate)



QoS-Definition



- Throughput (Durchsatz)
 - Maximale Langzeitrate = maximale Anzahl von übertragenen Dateneinheiten pro Zeitintervall
 - z.B. Pakete/Sekunde bzw. Bytes/Sekunde
 - Maximale Burst-Größe
 - Maximale Paketgröße



QoS-Definition



- Loss (Verlust)
 - Sensitivity classes
 - ignore
 - indicate
 - correct losses
 - Verlustrate = maximale Anzahl von Verlusten pro Zeitintervall
 - Verlustgröße = maximale Anzahl von konsekutiv verlorenen Paketen



QoS-Definition



- Delay (Verzögerung)
 - maximal zulässige Verzögerung zwischen Sender und Empfänger
 - maximal zulässige Verzögerungsschwankung (delay jitter) zwischen Sender und Empfänger
- QoS ist von den verfügbaren Ressourcen abhängig



QoS Definitionen



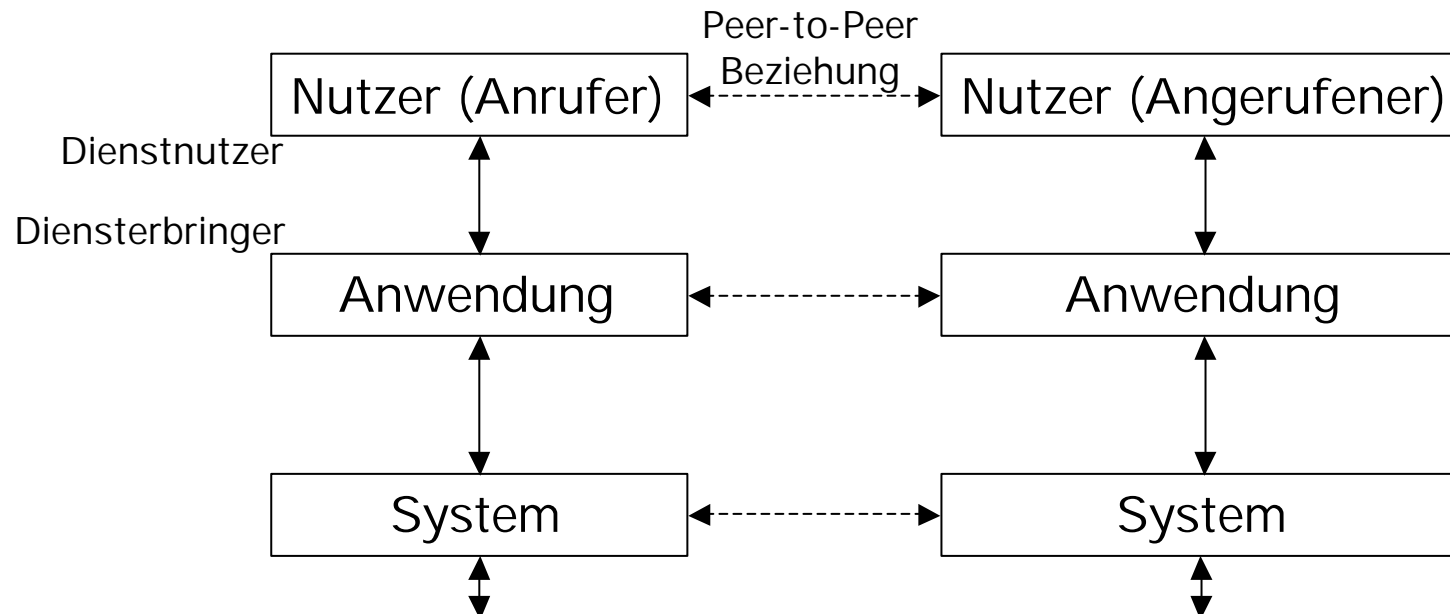
- Es gibt keine allgemein anerkannte oder standardisierte QoS-Definition
 - D.h. kein fester Satz von QoS Parametern
- Beispiele für QoS-Beschreibungen
 - IP Version 6 (IPv6)
 - Flow Label (24 bit) kennzeichnet Pakete eines Stroms
 - Priority Field (4 bit) zur Klassifizierung nach Wichtigkeit
 - ATM
 - Dienstkategorien auf der Basis von QoS-Parametern
 - CBR, UBR, rt-VBR, nrt-VBR, ABR



Dienstgüteverhandlung



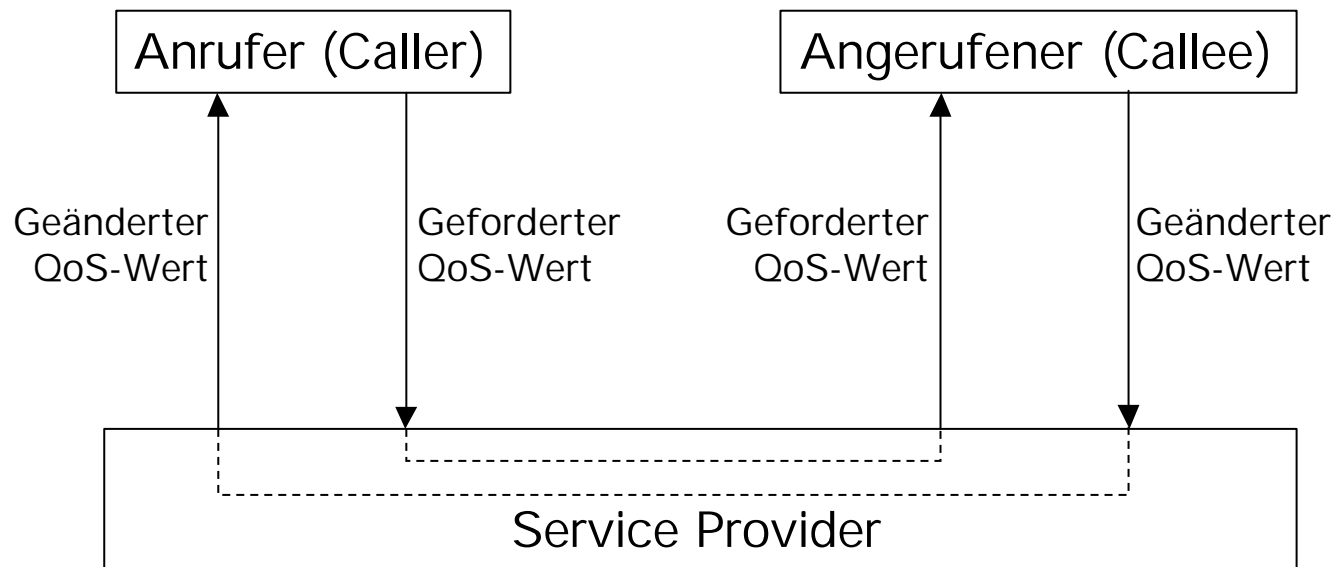
- Nach der Definition der Dienstgüteanforderungen müssen diese verteilt und ausgehandelt werden
- Modell



Dienstgüteverhandlung



- Bilaterale Peer-to-Peer-Verhandlung
 - Nur der Angerufene kann die geforderten QoS-Parameter ändern
 - Der Service Provider darf die QoS-Parameter nicht verändern



Dienstgüeverhandlung



- Bilaterale Schicht-zu-Schicht-Verhandlung
 - Die Verhandlung erfolgt ausschließlich zwischen dem Dienstonutzer und dem Erbringer
 - Z.B. lokaler Dienstonutzer und Betriebssystem
 - Z.B. Broadcast-Sender und Netz
- Unilaterale "Verhandlung"
 - Keine Modifikation der QoS-Parameter möglich
 - "take it or leave it"
 - Nutzer akzeptiert evtl. Parameter, die er aus Kapazitätsmangel nicht erfüllen kann
 - Z.B. Schwarz-Weiß-Fernseher und Farbfernsehsignal



Dienstgüteverhandlung



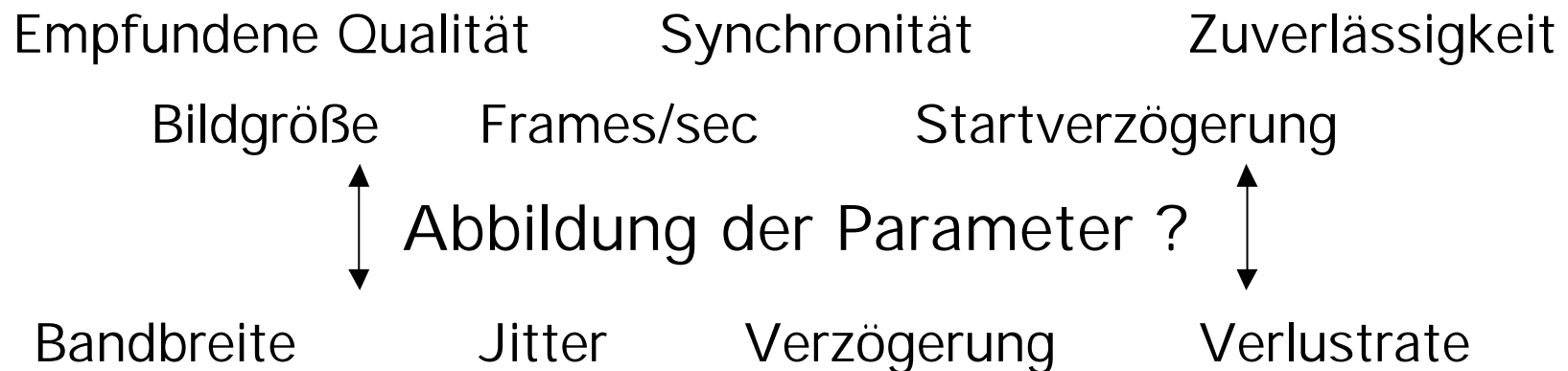
- Hybride Verhandlung
 - Notwendig bei heterogenen Empfängern (Broadcast/Multicast)
 - Der Sender verhandelt mit dem Netz bilateral mit den Empfängern unilateral
- Trilaterale Verhandlung
 - wechselseitiger Informationsaustausch, um bei mehreren Teilnehmern zu gleichen QoS-Werten zu kommen



Übersetzung der Dienstgüte



- Die QoS-Parameter der unterschiedlichen Schichten passen nicht immer 1:1 aufeinander, da die jeweilige Abstraktion evtl. sehr unterschiedliche Semantik besitzt



Verfügbarkeitsprüfung



- Die Verfügbarkeit der angeforderten Dienstgüte muss bei den angefragten Betriebsmitteln geprüft werden (Admission Control)
 - Dies gilt insbesondere bei gemeinsam benutzten Betriebsmitteln (CPU, Netz)
- Grundregel für einen Zugangstest (Admission Test)
 - Σ (bereits benutzte Ressourcen) + neue Anfrage < Kapazität ?
- Andere Tests
 - Schedulability
 - Pufferverfügbarkeit
 - Bandbreitenverfügbarkeit
- Alle Tests haben einen engen Bezug zu einem Kostenmodell
 - Die Nutzung der Ressourcen sollte den Nutzer was kosten sonst entsteht Over Provisioning



Ressourcenreservierung



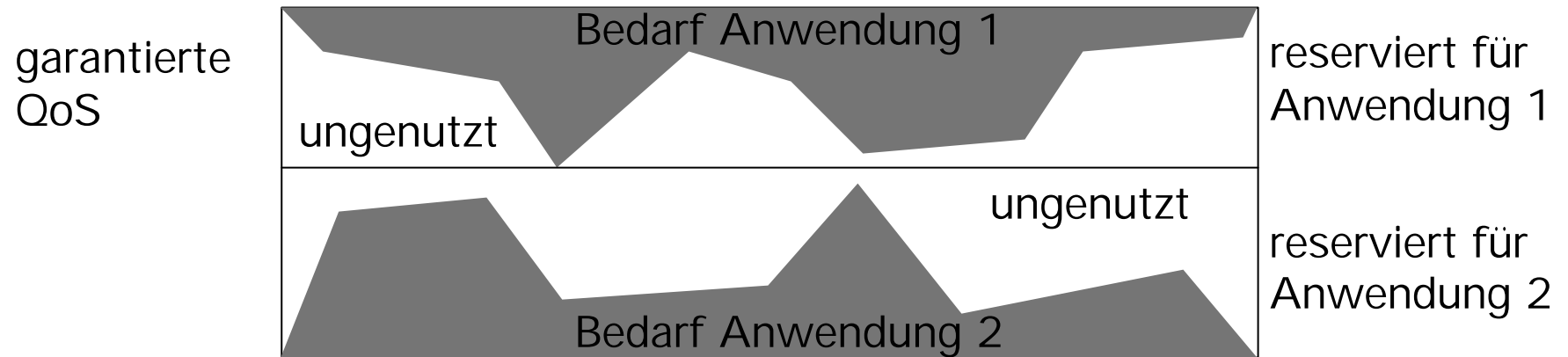
- Ohne Ressourcen zu reservieren werden Dienstgüten-Vereinbarungen nicht eingehalten werden können
- Pessimistischer Ansatz vermeidet Konflikte
→ garantierte QoS
- Optimistischer Ansatz geht von mittlerer Auslastung aus → statistische QoS
 - Denn viele Anwendungen erzeugen variable Datenraten
 - z.B. MPEG-Video, Audio mit Schweigeunterdrückung



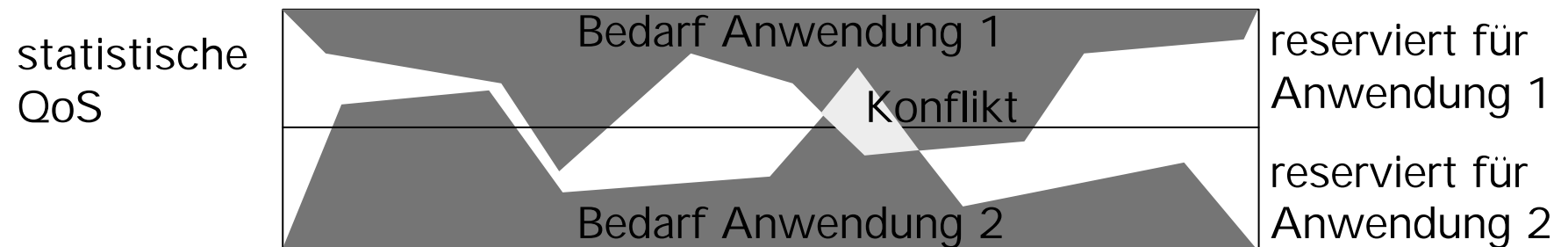
Ressourcenreservierung



■ Pessimistischer Ansatz



■ Optimistischer Ansatz



Deterministisch garantierte Dienstgüte



- 100%ige Garantie der QoS Werte
 - hard bounds
- QoS-Kalkulation basiert auf
 - harten Obergrenzen für den von der Quelle erzeugten Verkehr
 - Worst-Case-Annahmen bezüglich des Systemverhaltens



Deterministisch garantierte Dienstgüte



- Vorteile
 - QoS-Garantien auch im Worst Case erfüllt
 - hohe Zuverlässigkeit
- Nachteile
 - Überreservierung von Ressourcen
 - keine Ausnutzung des statistischen Multiplexing-Gewinns im Netz
 - unnötige Ablehnung von Reservierungsanfragen
 - harte Obergrenzen oft nicht zwingend für die Anwendung



Probabilistisch garantierte Dienstgüte



- QoS-Werte sind „soft bounds“
- QoS-Kalkulation basiert auf
 - Durchschnittswerten bzw. stochastischen Beschreibungen der Verkehrslast
 - probabilistischen Obergrenzen für das Systemverhalten



Probabilistisch garantierte Dienstgüte



- Vorteile
 - Ressourcen können statistisches Multiplexing ausnutzen
 - mehr Reservierungsanfragen können gleichzeitig berücksichtigt werden
- Nachteile
 - QoS kann zeitweise nicht voll erfüllt sein
 - schwerer implementierbar



Ressourcenreservierung



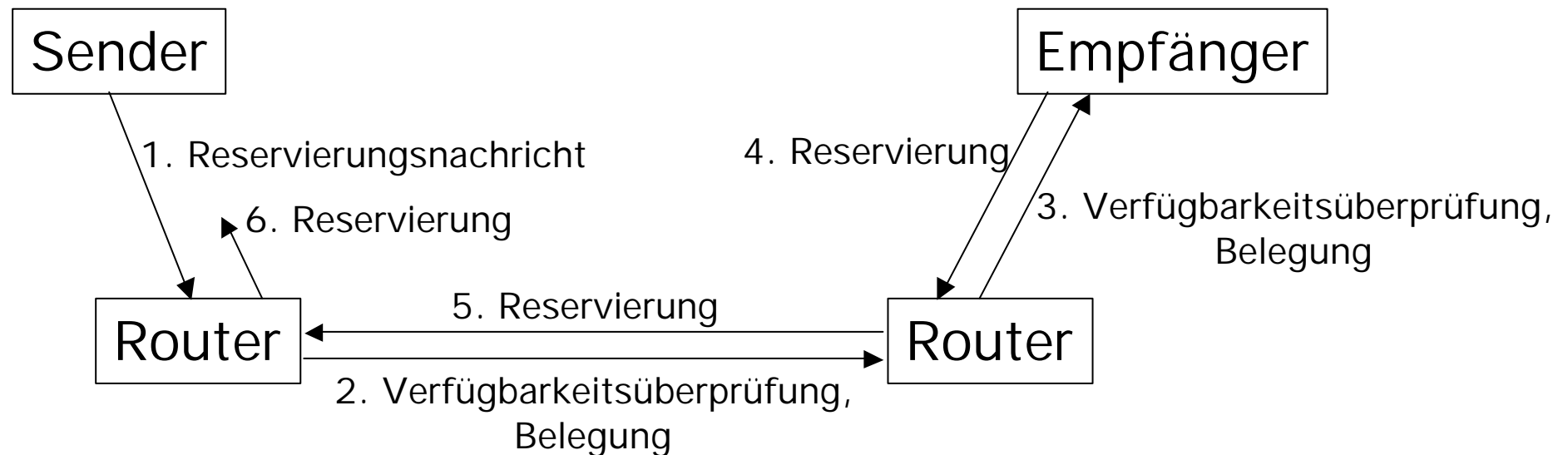
- Am Beispiel der Kommunikationsschicht lassen sich Reservierungsmodelle gut erläutern
- Prinzipiell sind die Modelle und Protokolle auch auf andere Systemkomponenten übertragbar



Senderorientierte Reservierung



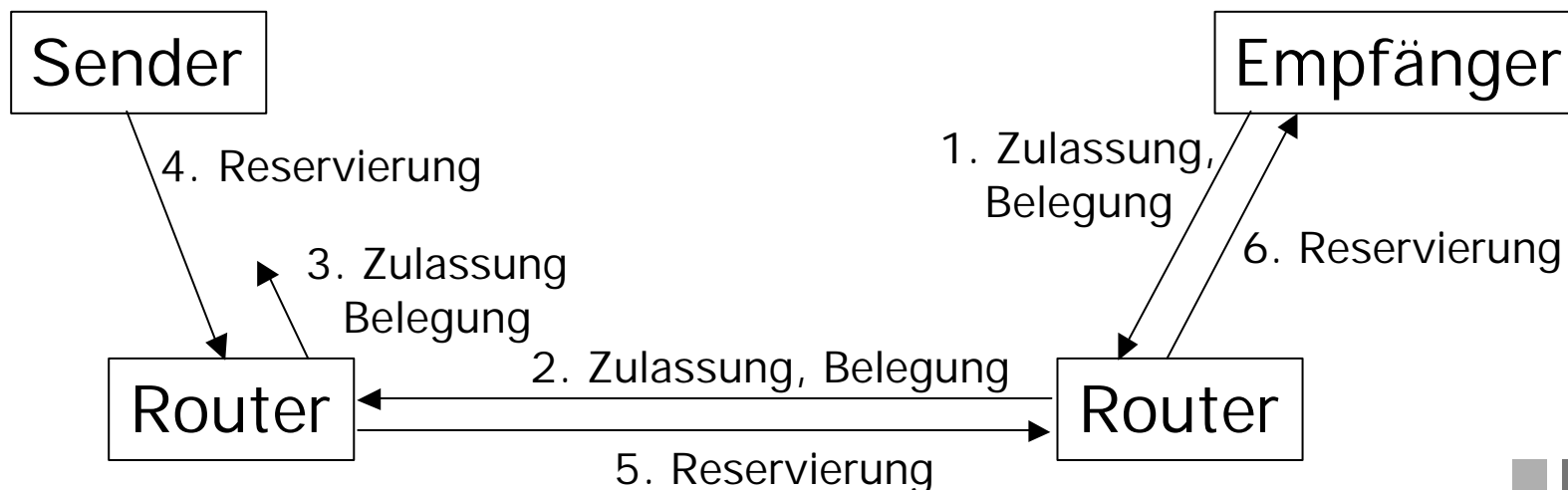
- Der Sender ist initiativ
 - Reservierung, Verhandlung und Zugang sind integriert
 - Beispiel: ST-II (stream protocol II) für einzelne Sender und mehrere Empfänger



Empfängerorientierte Reservierung



- Der Empfänger ist initiativ
 - Voraussetzung ist, dass der Sender vorab Information zur Datenquelle schickt
 - Beispiel: RSVP (resource reservation protocol) für mehrere Sender und Empfänger



Datenbearbeitungsphase



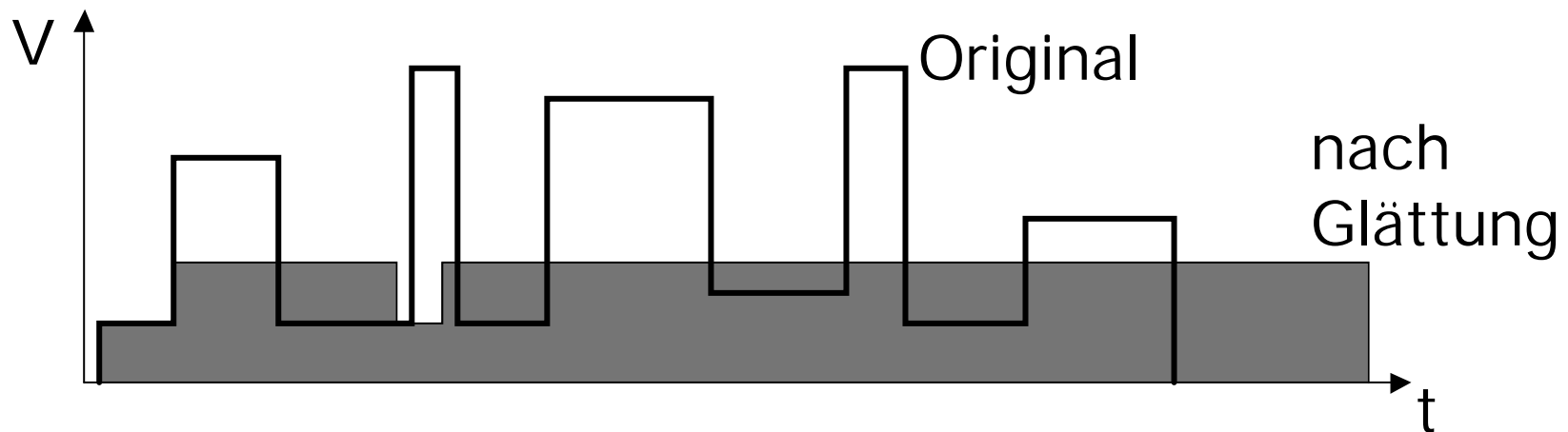
- Die ausgehandelten Ressourcenreservierungen müssen gepflegt werden
 - Zeitliche Randbedingungen
 - Speicherplatz
 - Exklusive Gerätebelegungen
 - Bandbreiten
 - Zuverlässigkeitsbedingungen
- Mechanismen dazu sind u.a.
 - Schedulingverfahren (siehe Kap. Betriebssysteme)
 - Verkehrsglättung (Traffic Shaping)
 - Rückkopplung und Adaption



Verkehrsglättung



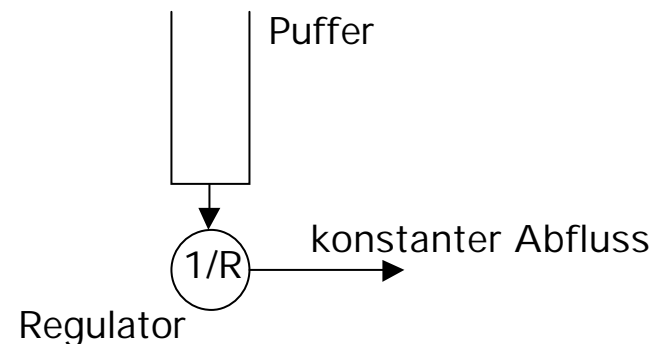
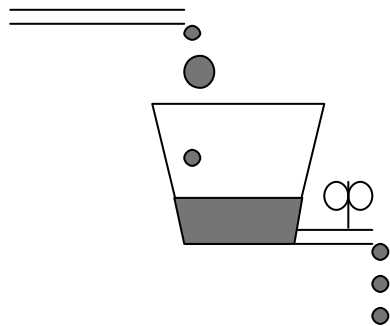
- Es wird versucht eine möglichst konstante Rate der Multimediaströme zu erreichen → Isochronität
 - Diese sind jedoch typischerweise "bursty"



Verkehrsglättung



- Leaky Bucket Verfahren
 - Der Sender platziert Pakete in einem Eimer der Größe b
 - Über ein Stellglied verlassen konstant $1/R$ Pakete den Abfluss
 - b bestimmt die mögliche Verzögerung und die maximale Kapazität vor dem Paketverlust



Rückkopplung und Adaption



- Die Last des Netzes und der Endsysteme wird von mittels Monitor gemessen
- Wenn signifikante Veränderungen auftreten, dann werden adäquate Maßnahmen ergriffen, um die Last zu reduzieren
 - Z.B. explizite Aufforderung des Senders durch Empfänger oder Netzkomponenten langsamer zu senden → Flusskontrolle
- Reaktionsmöglichkeiten sind u.a.
 - Degradation
 - Layered Transmission



QoS Architekturen



- In einer QoS-Architektur wird das Zusammenspiel der verschiedenen Ende-zu-Ende-Komponenten bzgl. Spezifikation, Verteilung, Bereitstellung von Dienstgüte definiert



QoS Architekturen



- Beispiele
 - Internet Integrated Services
 - Behandelt Flows in spezifischer Form in der existierenden Internet-Infrastruktur
 - RSVP als Signalisierungsprotokoll
 - Internet Differentiated Services
 - Definiert Serviceklassen, verhandelt Service Level Agreements und garantiert die dedizierte Behandlung der Flows, falls diese sich konform verhalten
 - IPv6
 - Dedizierte Header Fields zur Klassifikation von Flows
 - und viele (weitreichende) akademische Ansätze, die umfassend Betriebssystem- und Kommunikationssystemaspekte beinhalten

