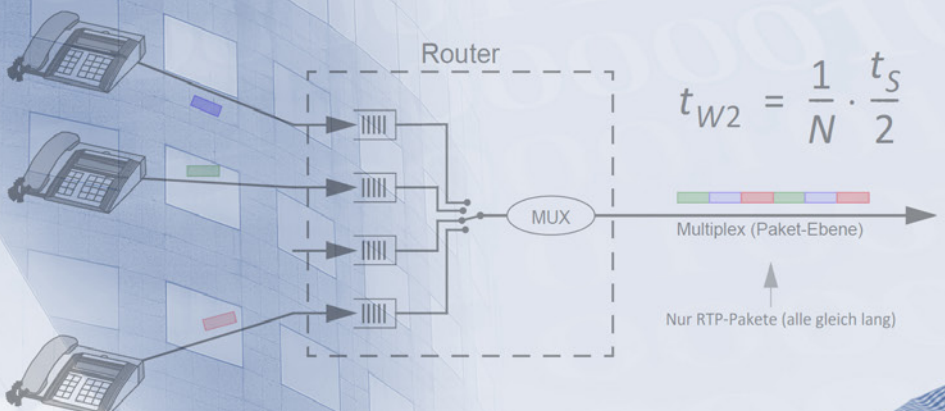
Bandbreitenberechnungen in VoIP-Systemen

VAF BUNDESVERBAND TELEKOMMUNIKATION

Ein Arbeitspapier des
VAF Bundesverband Telekommunikation e.V.

Aktualisierte und erweiterte 2. Auflage

$$= \left(0 + \frac{1}{N} \cdot \frac{t_S}{2}\right) / 2 = \frac{1}{N} \cdot \frac{t_S}{4}$$



$$t_{W2} = \frac{1}{N} \cdot \frac{t_S}{2}$$

Impressum

Bandbreitenberechnungen in VoIP-Systemen
Ein Arbeitspapier des VAF Bundesverband Telekommunikation e.V.

Autor:
Prof. Dr. Gerd Siegmund, Stuttgart

Titelseitengestaltung:
Uwe Klenner, Passau

Bildnachweis Titelseite:
shutterstock

Herausgeber:
VAF Bundesverband Telekommunikation e.V.
Otto-Hahn-Str. 16
40721 Hilden
Tel.: 02103 700-250
Fax: 02103 700-106
info@vaf-ev.de
www.vaf-ev.de

Aktualisierte und erweiterte 2. Auflage
Erstauflage 2012

Inhaltsverzeichnis

1 Einführung

Private Kommunikationssysteme bzw. Telekommunikationsanlagen (TK-Anlagen, TK-Systeme) aktueller Bauart basieren auf dem Internetprotokoll (IP). Diese Systeme werden oft in bereits vorhandene Kundennetze integriert, deren Eigenschaften mit wenigen Angaben als „VoIP-Ready“ deklariert werden. Mit den TK-Systemen werden jedoch erhöhte Anforderungen an diese Netze gestellt, um eine gewisse Dienstqualität (Quality of Service – QoS) für die Sprachübertragung sicherzustellen. Zugleich wird die Belastung dieser Netze durch die Pakete für die Sprachübertragung deutlich erhöht. Die Voice over IP (VoIP) Echtzeitkommunikation erzeugt relativ viele kleine (ca. 200 Byte) zusätzlichen IP-Pakete. Die Anzahl der transportierten Pakete beeinflußt die Netzbelastung und diese die Qualität der Netze, insbeondere auch die der Sprachübertragung. Für viele Anwendungen wie Web-Aufrufe und eMail ist das meist kein Problem. Für die Echzeitübertragung wie Sprache sind solche (auch nur geringen) Einschränkungen der Qualität unmittelbar spürbar. Die klassischen Datenanwendungen basieren meist auf dem Transportprotokoll TCP, das im Fall von Fehlern oder Paketverlust eine Wiederholung der Pakete vorsieht. Echtzeitverbindungen basieren auf UDP, bei dem keine Wiederholungen möglich sind. Speziell die Sprachübertragung erfordert zudem einen vollständigen und regelmäßigen Empfang von IP-Paketen. Bei größeren Verzögerungen (ca. 150 ms – die genaue Zeit hängt von der Größe des Empfangsspeichers ab) tritt ein Paketverlust ein, weil die Pakete einfach zu spät den Empfänger erreichen. Solche Störungen und Beeinflussungen der Übertragungsqualität durch Paketverlust treten typischer Weise sporadisch und kaum reproduzierbar auf. Eine einfache Mischung von Sprache und Daten in einem Netz bedeutet praktisch immer, dass es zu sporadischen Störungen in der Verständigung kommen kann – dies ist auch praktisch unabhängig von der Übertragungsgeschwindigkeit des Netzes.

Viele Netze verfügen daher sinnvollerweise über spezielle QoS-Maßnahmen. Eine dieser Maßnahmen ist die Priorisierung der Echtzeitinformationen gegenüber den anderen IP-Paketen, ein Beispiel hierfür ist Differentiated Services (DiffServ). Eine andere Möglichkeit ist die Einrichtung von VLAN und im Bereich von Standortvernetzung zudem der Einsatz von MPLS-Systemen (MPLS: Multiprotocol Label Switching) oder Software-defined Networking (SDN). Mitunter trifft man auch auf die Ansicht, dass eine Überdimensionierung der Netze für die Unterstützung von Echtzeitdiensten völlig ausreichend sei. Welche Maßnahmen im Netz ergriffen wurden, ist für den Anbieter bzw. den Integrator der TK-Systeme häufig nicht ersichtlich. Dies zeigen auch Beispiele von typischen Ausschreibungstexten in diesem Bereich. In Ausschreibungen heißt es mitunter sogar nur „Das Netz ist VoIP-Ready“. Diese Aussage kann bedeuten, dass eine Priorisierung des VoIP-Verkehrs vorgenommen wird, dass VLAN- oder MPLS-Systeme eingesetzt werden oder dass Überkapazitäten im Netz für den zusätzlichen VoIP-Verkehr vorhanden sind. Diese verschiedenen Maßnahmen sind aber absolut nicht gleichwertig. In den folgenden Berechnungen wird gezeigt, dass auch eine deutliche Überdimensionierung Probleme bereiten kann oder das eine vorhandene 2,048-Mbit/s-Strecke mit einer durchschnittlichen Belastung von 5 % für die klassische Datenkommunikation zwischen 7 und 24 VoIP-Kanäle transportieren kann, je nachdem, welche Maßnahmen im Netz ergriffen wurden.

Dieses weite Spektrum zeigt, dass ohne Kenntnis der jeweiligen QoS-Maßnahmen oder ohne Beachtung der spezifischen Eigenschaften des Netzes Vorhersagen über die zu erwartenden Eigenschaften der VoIP-Installation bzw. Verkehrsberechnungen für den konkreten Fall unmöglich sind. Dies gilt auch, obwohl selbst Installationen ohne QoS-Maßnahmen oft recht gut funktionieren. Denn im laufenden Betrieb dieser Systeme sind „unerwartet“ auftretende Störungen in den Gesprächen durch stark verzögerte oder fehlende VoIP-Pakete vorprogrammiert. Allein die Überdimensionierung eines Netzwerks kann keinen reibungslosen Betrieb garantieren. Etwas besser sieht es in Systemen mit einer strikten Bevorrechtigung der VoIP-Pakete aus. Idealerweise sollten aber SDN, MPLS oder VLAN mit QoS-Maßnahmen auf Hardwarebasis als Transportnetz eingesetzt werden. Für den Transport der VoIP- und der Daten-Pakete werden hier getrennte logische Kanäle verwendet. Durch den Einsatz dieser getrennten, virtuellen Kanäle kann die größte Anzahl von VoIP-Kanälen in einem gegebenen System unterstützt werden. Zudem zeigen die Berechnungen, dass nur mit dieser Technik geringe Verzögerungszeiten bei Pakettransport zuverlässig eingehalten werden können.

2 Das Zeitmultiplexsystem

2.1 TDM

Seit der Einführung der digitalen Vermittlungstechnik werden 64-kbit/s-Kanäle für die Übertragung der Sprachinformationen im klassischen Fernsprechnetz verwendet. Die Sprachsignale werden mit dem Pulscodemodulationsverfahren alle 125 µs abgetastet und mit jeweils 8 Bit dargestellt. Durch die Abtastung im 8-kHz-Raster und die Darstellung mit 8 Bit mit einem Codec ergibt sich eine Übermittlungsgeschwindigkeit von 64 kbit/s. In der digitalen Vermittlungstechnik wird dieser Kanal, im ISDN auch als B-Kanal bezeichnet, den beiden Kommunikationspartnern für die Dauer der Kommunikation exklusiv zur Verfügung gestellt.

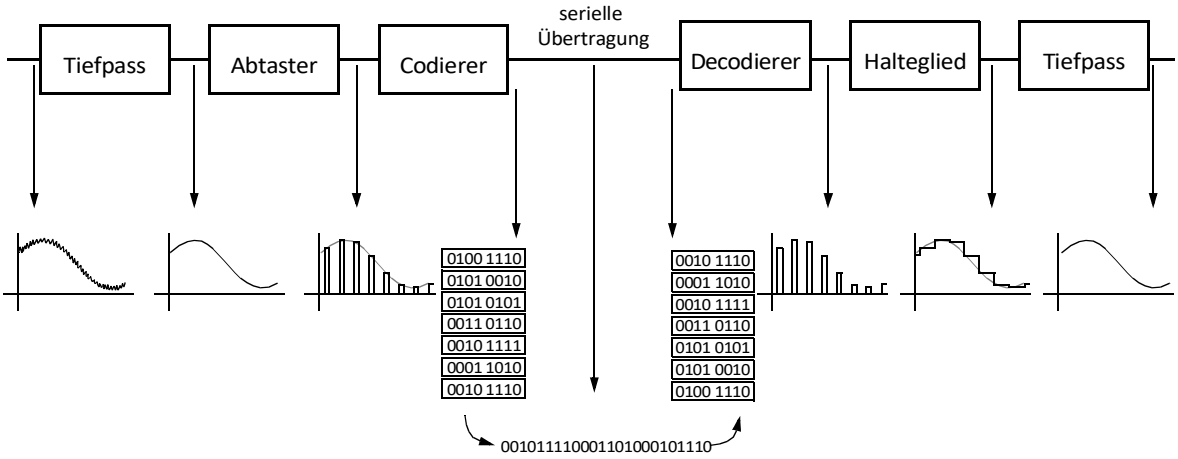


Abbildung 1: Codierung der Sprache nach G.711

Die Übertragung der Sprachinformation erfolgte im ISDN mit Zeitmultiplexsystemen (Time Division Multiplex – TDM), bei denen mehrere 64-kbit/s-Kanäle gleichzeitig übertragen werden. Beim Primärmultiplexsystem bilden beispielsweise 30 solcher 64-kbit/s-Kanäle plus einen Kanal für die Synchronisierung und ein Kanal zur Übertragung der Signalisierung (beide jeweils mit 64 kbit/s, zusammen also 32 Kanäle mit jeweils 64 kbit/s), einen Übertragungsrahmen.

In diesem Multiplex hat jeder einzelne Kanal einen festen Zeitbereich mit konstanter Wiederholrate. Andere Kanäle können diesen nicht stören oder beeinträchtigen. Der Nutzen der Multiplextechnik liegt in der gemeinsamen Verwendung einer Übertragungsstrecke. Für jede der verschiedenen Übertragungen in einem Multiplex verhält es sich, als ob sie eine eigene Leitung (mit einer begrenzten Bandbreite) für die Übertragung exklusiv hätte (Abbildung 2).

2.2 Verkehrsberechnungen (Erlang)

Das klassische Kommunikationsnetz arbeitet nach dem Prinzip eines Verlustsystems: Sind noch genügend Leitungen oder Kanäle vorhanden, können diese Verbindungen zugeordnet werden. Wenn alle Kanäle belegt sind, kommt es zu Verlust, d. h., der Verbindungswunsch kann nicht erfüllt werden.

In der folgenden Abbildung ist dieses Verhalten der TDM-Systeme dargestellt. Solange noch freie Kanäle zur Verfügung stehen, können diese auf Anfrage vergeben werden – der Durchsatz kann entsprechend gesteigert werden. Wenn alle Kanäle vergeben sind (die normalisierte Belastung ist dann genau 1), wird auch der maximal mögliche Durchsatz erreicht (der normalisierte Durchsatz ist ebenfalls gleich 1). Eine weitere Steigerung der Belastung führt nicht zur einer weiteren Durchsatzsteigerung [Sie09].

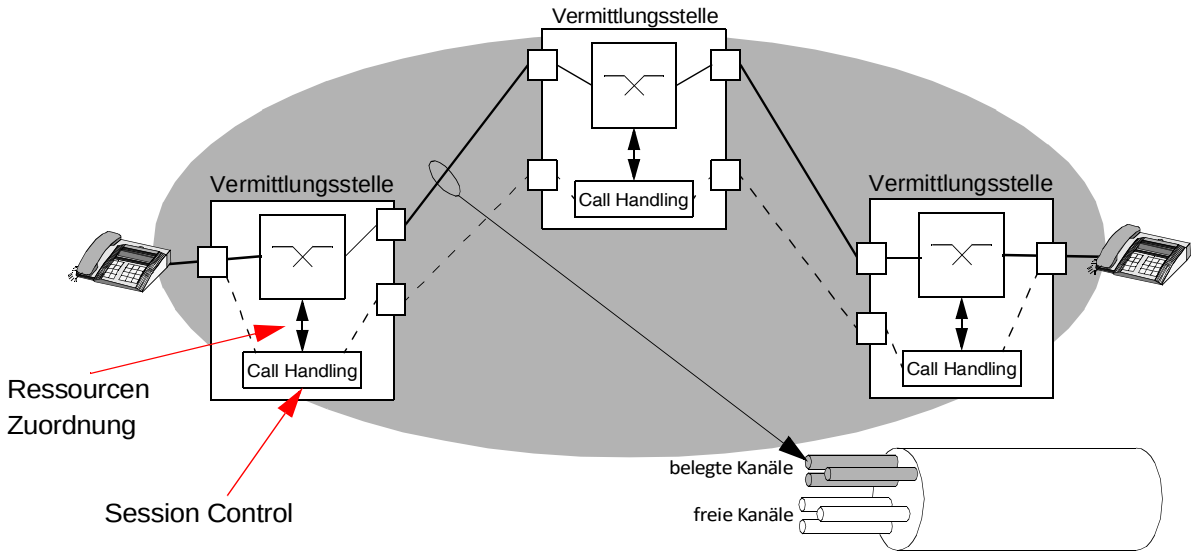


Abbildung 2: Nutzkanäle im ISDN

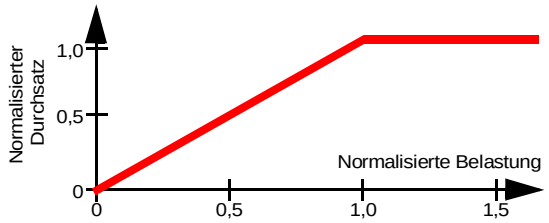


Abbildung 3: Mit der Belastung steigt der Duschsatz solange noch freie Kanäle verfügbar sind

Die Verkehrstheorie ermöglicht mithilfe der Statistik die Berechnung der notwendigen Kanäle oder Leitungen, die für eine bestimmte Verkehrsbelastung (das Verkehrsangebot) und einen bestimmten, akzeptablen Verlust (z. B. < 1 %) notwendig sind.

Als Beispiel ein Ausschnitt aus einem Ausschreibungstext:

Für ein TK-System mit 1000 Telefonen und einem Verkehrswert von 0,025 Erl (80 % intern, 20 % extern) und 1000 Telefonen mit einem Verkehrswert von 0,030 Erl (50 % intern, 50 % extern)soll die Auslegung der externen Leitungen betrachtet werden. Machen Sie einen Vorschlag für die externe Anbindung für einen maximalen Verlust von 1 %. Das Netz ist VoIP-Ready.

Hinweis: Der Verkehrswert ist eine Zeiteinheit wird aber in der Einheit Erlang, kurz Erl angegeben. Agner Krarup Erlang war ein dänischen Mathematiker der seine ersten Arbeiten zu der Warteschlangentheorie 1909 veröffentlichte.

2.3 Berechnung des Verkehrsangebots

Zuerst kann das Verkehrsangebot durch die Teilnehmeranschlüsse ermittelt werden:

1000 · 0,025 Erl = 25 Erl und 1000 · 0,030 Erl = 30 Erl, das macht für die externe Kommunikation dann: 25 Erl · 0,2 = 5 Erl und 30 Erl · 0,5 = 15 Erl. Zusammen sind dies 20 Erl für die externe Kommunikation.

Dieses Verkehrsangebot ist unabhängig davon, ob ein TDM- oder VoIP-System ausgelegt werden soll.

Als Vergleichswert werden zuerst die benötigten Leitungen/Kanäle für ein TDM-System ermittelt. Für ein Verkehrsangebot von 20 Erl sind 30 Leitungen (Kanäle) erforderlich, dann bleibt der Verlust unter 1 % (s. folgende Tabelle, [Sie10]). Damit reicht für die externe Anbindung in TDM-Technik genau ein Primärmultiplexsystem mit 30 Nutzkanälen.

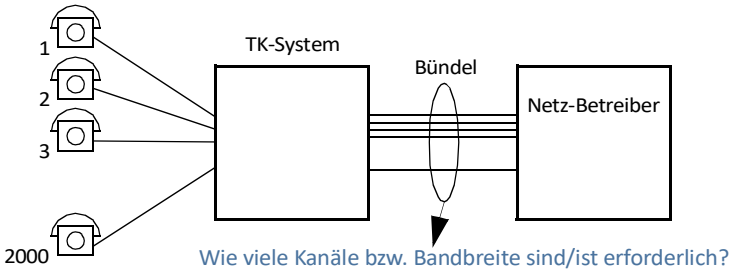


Abbildung 4: Berechnungsbeispiel für ein Bündel

	A in Erlang						
N	B = 0,1 %	B = 0,2 %	B = 0,5 %	**B = 1 %**	B = 5 %	B = 10 %	B = 20 %
23	11,5	12,3	13,4	14,5	18,1	20,7	25,3
24	12,2	13,0	14,2	15,3	19,0	21,8	26,5
25	13,0	13,8	15,0	16,1	20,0	22,8	27,7
26	13,7	14,5	15,8	17,0	20,9	23,9	28,9
27	14,4	15,3	16,6	17,8	21,9	24,9	30,2
28	15,2	16,1	17,4	18,6	22,9	26,0	31,4
29	15,9	16,8	18,2	19,5	23,8	27,1	32,6
30	16,7	17,6	19,0	**20,3**	24,8	28,1	33,8
31	17,4	18,4	19,9	21,2	25,8	29,2	35,1
32	18,2	19,2	20,7	22,0	26,7	30,2	36,3

Tabelle 1: Verkehrstabelle (Ausschnitt)

3 Grundlagen und Hintergrund

Die ISDN-Zeiten sind vorbei. ISDN wurde in den 80er Jahren eingeführt erreichte ab ca.2014 das Ende seiner geplanten 30-jährigen Betriebszeit. Für die alten ISDN-Systeme bekommt der Netzanbieter keine Ersatzteile mehr und die laufenden Systeme verlieren die einst so hohe System-Verfügbarkeit. Die Nachfolger basieren auf dem Internetprotokoll, weil die meisten der transportieren Informationen auf diesem Protokoll basieren. Anders als typische Internet-Anwendungen (World Wide Web oder eMail) erfordert die Übertragung von Sprache besondere Maßnahmen für den Transport, weil das Internet für solche Echtzeitanwendungen nicht ausgelegt wurde. Das Internet basiert auf dem Transport von Paketen, die von den jeweiligen Quellen spontan und mit unterschiedlicher Länge und Intensität erzeugt werden. Innerhalb des Netzes werden diese Pakete nacheinander transportiert und werden in Eingangsspeichern der verschiedenen Netzelemente zwischengespeichert. Sie verweilen dort, bis sie durch das Netzelement bearbeitet und weiter transportiert werden. Die Wartezeit in diesen Zwischenspeichern hängt von der Auslastung der Systeme und damit von dem Verkehr anderer Kommunikationenbeziehungen ab. Damit hängt die Laufzeit jedes einzelnen Paketes durch das Netz von der augenblicklichen Verkehrsbelastung im Netz ab. Diese Paketlaufzeiten (Delay) können von Paket zu Paket sehr stark schwanken (das wird als Jitter bezeichnet).

Laufzeitmessungen in lokalen Netzen können immer nur Augenblickswerte ergeben, sie treffen keine Aussagen zu dem zukünftigen Verhalten des Netzes. Das bedeutet, das ermittelte Durchschnittswerte sich bei wiederholten Messungen völlig anders darstellen können.

Durchschnittswerte verbergen die eigentlichen Probleme. Datenpakete werden im Internet nicht immer mit mit dem gleichen Abstand zueinander übertragen. In der Abbildung 5 werden innerhalb des gleichen Zeitraums (von t_0 bis t_1) gleich 5 Pakete übertragen. Die obere Übertragung erfolgt mit einer kontinuierlichen Paketrate (die Ankunftsrate der Pakete), die untere Übertragung ist „Burst-artig“.

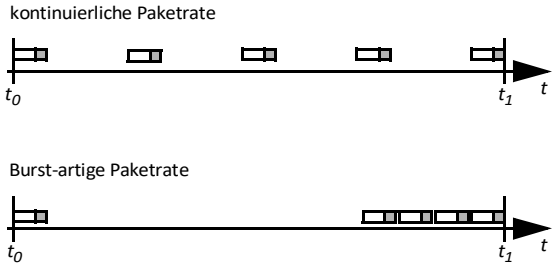


Abbildung 5: Verteilung der Ankunftsrate bei gleicher Übertragungsrate

Bei vielen, vereinfachten Betrachtungen von Systemauslastungen werden die erforderlichen, durchschnittlichen Bandbreiten einfach zusammengezählt und mit der Übertragungsrate des Netzes verglichen. Dabei werden die Burst-artigen Übertragungen der Datenkommunikation nicht berücksichtigt, genau diese machen aber die Probleme. Durch die Burst-artige Übertragung in Datennetzen füllen sich während eines Daten-Bursts die Zwischenspeicher, was größere Wartezeiten verursacht. Auch, wenn im Mittel alles passt, gibt es Probleme mit unterschiedlichen Paketlaufzeiten.

Der zeitliche Abstand zweier nacheinander gesendeter Pakete hängt von der Anzahl und Häufigkeit der Pakete im Netz, also auch von vielen anderen Verbindungen ab. In den einzelnen Abschnitten und ggf. auch im Zugangsbereich werden parallel zu den Paketen der betrachteten Verbindung auch Pakete anderer Verbindungen übertragen. Diese verwenden ganz oder teilweise die gleichen Abschnitte wie die betrachtete Verbindung. Werden Pakete verschiedener Kommunikationen über einen Abschnitt transportiert, müssen sie nacheinander den Abschnitt passieren, dies verursacht eine Wartezeit, die von der Anzahl und der Größe der anderen Pakete abhängt. Da auch die anderen Verbindungen unregelmäßig unterschiedlich große Pakete senden, ist diese Wartezeit sehr großen Schwankungen unterworfen. Der Abstand zwischen zwei Paketen vom Sender wird durch den Transport im Internet verändert und ist nicht konstant.

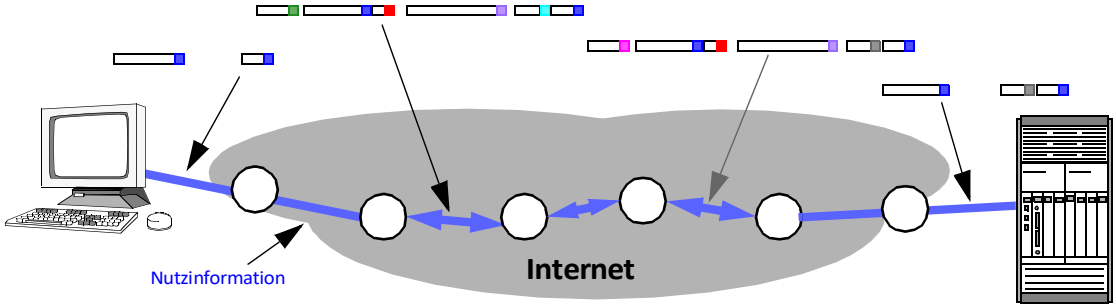


Abbildung 6: Pakete in Konkurrenz

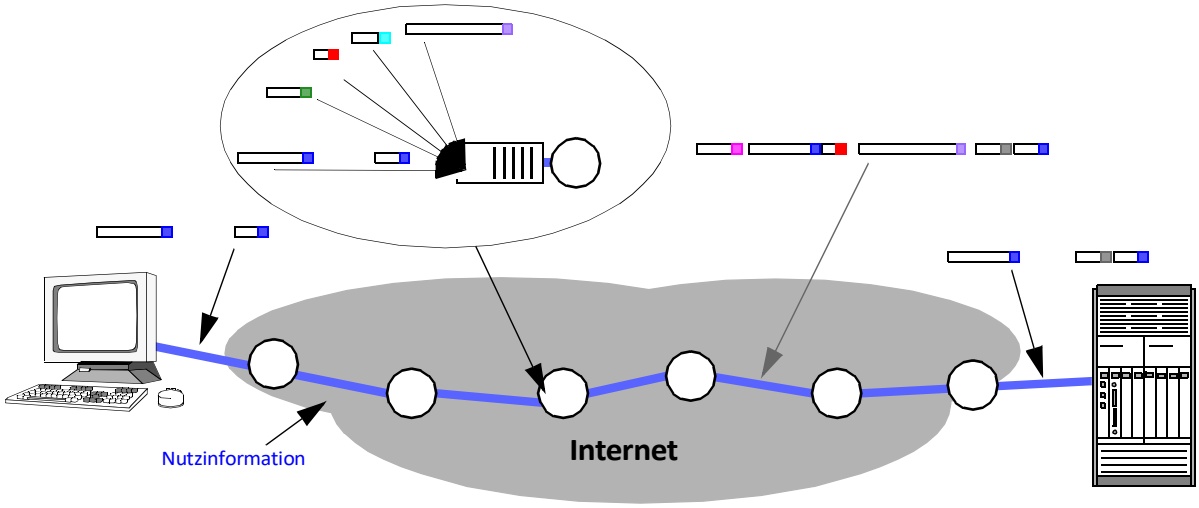


Abbildung 7: Zwischenspeicher in jedem Netzelement

Abhängigkeit von der Systemauslastung

Die Wartezeiten sind nicht konstant, sie hängen von der augenblicklichen Systemauslastung der einzelnen Systeme ab. An jeder Warteschlange kommen die Anfragen von vielen Schnittstellen oder Terminals zusammen, die jeweils zufällig und unkoordiniert Pakete senden.

Die Systemreaktionszeit oder Paketlaufzeit stellt sich für viele Benutzer als eine zufällige Größe dar. Für den Benutzer ist es oft leicht einsichtig, dass die Wartezeit von der Geschwindigkeit und der Leistungsfähigkeit der verwendeten Systeme und der Leitungen abhängt. Offensichtlich ist für alle Anwender, dass die Reaktionszeit von der Systemauslastung abhängig ist, unbekannt ist dagegen oft, dass diese Zeit nicht linear ansteigt. Tatsächlich steigt die Wartezeit exponentiell an, wenn sich die Systemauslastung der theoretischen Leistungsfähigkeit des Systems nähert (Abb. 8).

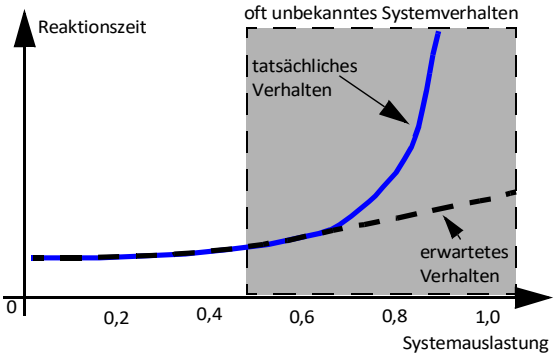


Abbildung 8: Erwartetes und tatsächliches Verhalten von Wartezeitsystemen

3.1 Auslastung des Systems

Für die weiteren Betrachtungen ist die Systemauslastung von besonderer Bedeutung. Die Systemauslastung bestimmt die mittlere Wartezeit in einem System. Das Verhältnis λ/μ wird als Auslastungsfaktor ρ (manchmal auch als Nutzungsgrad, engl. utilization) bezeichnet. Für einfache Systeme mit einer Bedieneinheit gilt:

$$\rho = \frac{\lambda}{\mu} =$$

$$\frac{\text{mittlere Ankunftsrate (Last)}}{\text{mittlere Bedienrate (Leistung des Systems)}} = \text{mittlere Auslastung.}$$

Stabilität des Systems, wenn $\rho < 1$ -> $\lambda < \mu$.

Aus dem *Satz von Little* lassen sich die Kenndaten bestimmen, beispielsweise auch die mittlere Anzahl von Anfragen im System:

$$N = \frac{\rho}{1-\rho}.$$

Dieser Zusammenhang ist in der Auslastungskurve in der Abbildung 9 dargestellt. Ab einer Auslastung von ca. 80 % steigt die Anzahl der Anfragen im System drastisch an. Die durchschnittliche Wartedauer in der Warteschlange (W) ist

$$W = \frac{N_q}{\lambda}.$$

Zwischen der Anzahl der Wartenden im System (N_q) und der Wartezeit besteht ein linearer Zusammenhang. Dieses Verhalten kann auch unter Laborbedingungen gemessen werden. In der folgenden Abbildung 9 wurde die mittlere Antwortzeit eines Routers in Abhängigkeit von der Auslastung aufgenommen.

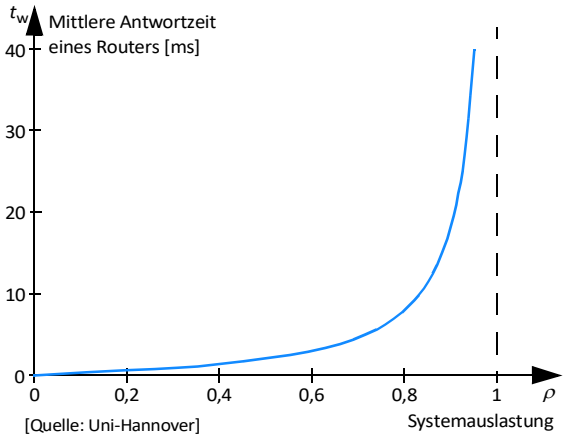


Abbildung 9: Mittlere, gemessene Antwortzeit eines Routers

Die Kurve entspricht genau dem erwarteten Verhalten für einfache Systeme in denen die Ankunftsrate der Pakete und die Bedienrate durch die Netzelemente klare Mittelwerte und eine begrenzte Varianz aufweisen. Reale Systeme in den lokalen Netze sind viel komplexer. Sie verfügen über keine klaren Mittelwerte und die Varianz strebt gegen unendlich. Das bedeutet, dass Messungen der Paketlaufzeiten in diesen Netzen immer wieder völlig andere Werte. Die Kurve aus der Abbildung 9 spiegelt also nicht das Verhalten eines Routers in einem typischen LAN wider. Untersuchungen an realen Systemen zeigten, dass hier bereits bei deutlich geringeren Auslastungen relativ große Wartezeiten entstehen.

3.2 Laufzeitvarianzen

In Abhängigkeit von der Auslastung steigt die Wartezeit und damit die Durchlaufzeit für ein Paket durch das System an. In der Abbildung 10 ist der nichtlineare Zusammenhang zwischen der Last, der Paketankunftsrate und der Durchlaufzeit (Delay) durch ein Netzelement dargestellt. Auslastungen oberhalb von 80 % (ρ < 0,8) verursachen bereits in diesen einfachen Systemen deutlich größere Paketlaufzeiten, bei geringeren Auslastungen bleiben die Laufzeiten dagegen sehr klein. Werden mehrere solcher Systeme hintereinander geschaltet, d. h. nacheinander von den Paketen durchlaufen, verstärken sich diese Einflüsse, d. h. die Verkehrsspitzen werden im Vergleich zu den geringeren Auslastungen noch stärker hervortreten.

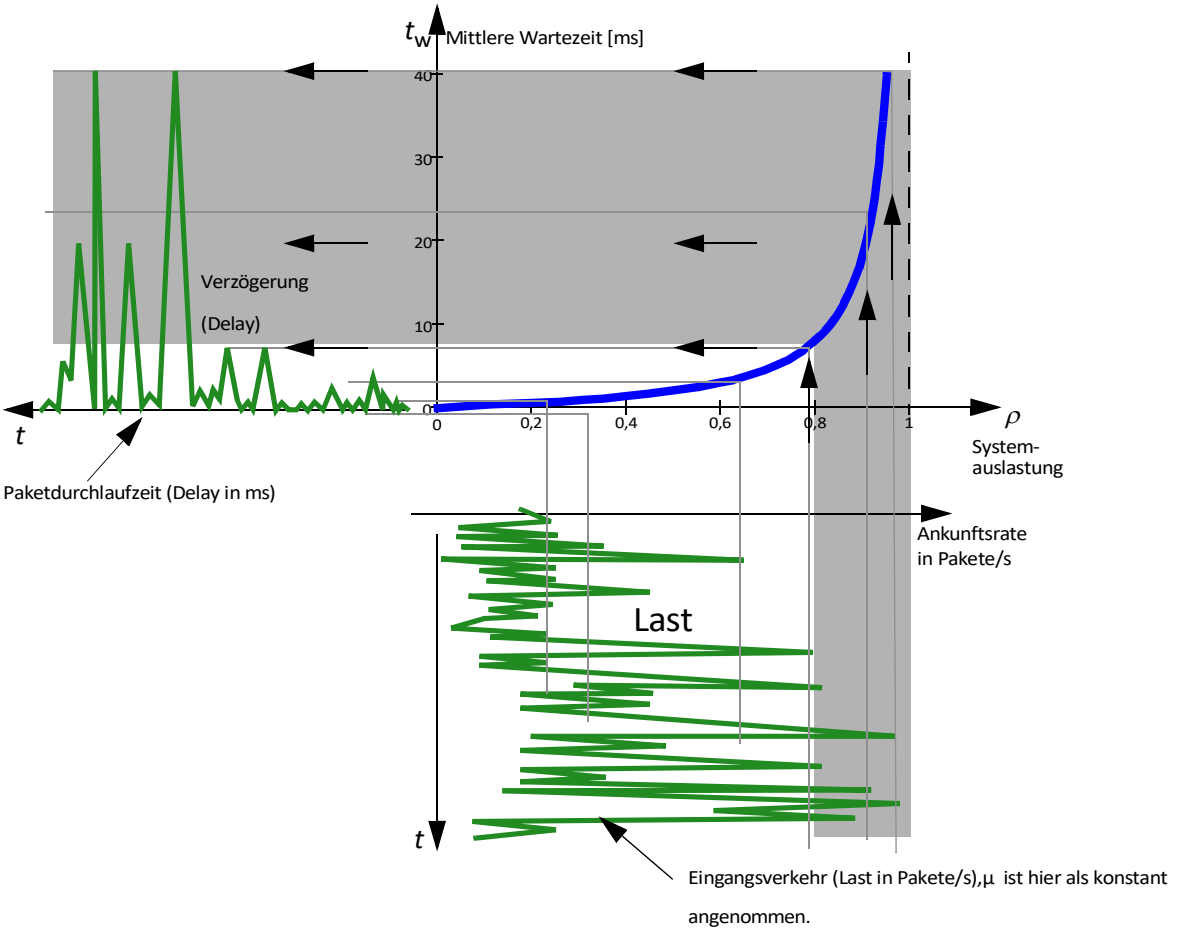


Abbildung 10: Paketlaufzeiten sind Lastabhängig

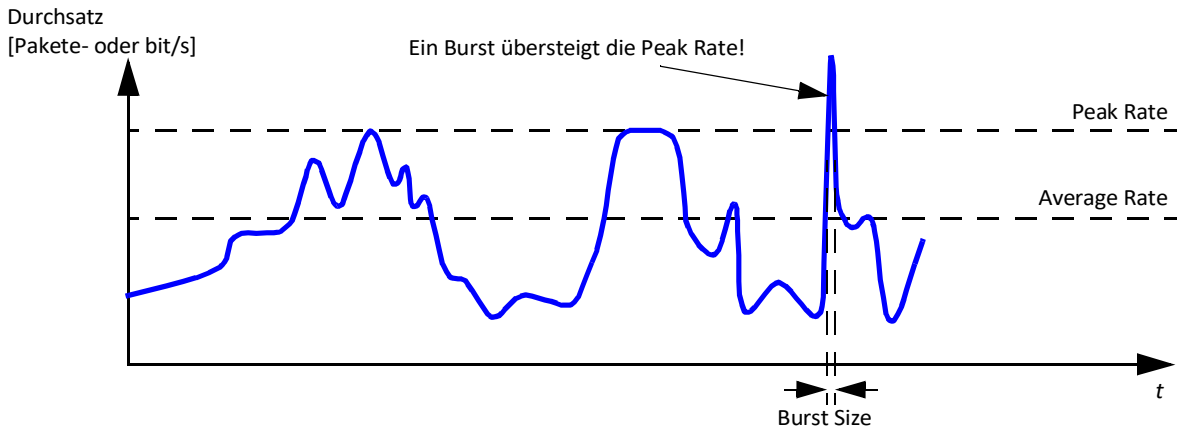


Abbildung 11: Durchsatz im Internet

3.3 Das Internet ist anders

Während einer Kommunikation im Internet werden von einer Anwendung nicht immer die gleiche Anzahl an Paketen oder der gleiche Mittelwert der Verteilung der Pakte erzeugt. Die Beanspruchung, der Durchsatz, des Systems wechselt über die Beobachtungszeit hinweg sehr stark. Je nach Beobachtungszeitraum lassen sich die folgenden Kenndaten unterscheiden (Abb. 11):

- Average Rate (mittlere Paket- oder Bitrate) – Beobachtung der Rate über längeren Zeitraum (typisch Minuten, z. B. 6000 Pakete/min).
- Peak Rate (maximale Paket- oder Bitrate) – Beobachtung der Rate über kurzen Zeitraum (typisch Sekunden, z. B. 1500 Pakete/s).
- Burst Size – Beobachtung der Rate in sehr kurzen Zeiträumen (typisch ms, z. B. 100 Pakete/10 ms, häufig wird der Beobachtungszeitraum noch kürzer gewählt).

Selbstähnlichkeit und Hurst-Parameter

Bei der Beobachtung des IP-Verkehrs einer großen Anzahl von Nutzern und über eine große Zeitspanne fällt es schwer, Strukturen oder wiederkehrende Muster zu erkennen. Die Verkehrskurven weisen immer sehr starke Nutzungsunterschiede auf, die Kurven sind sehr spitzig, haben viele sehr steile *Peaks* und ändern sich rasch von intensiver zu einer geringen Nutzung. In der Aktivitätsphase wird oft gleich eine Anzahl von Paketen zwischen beiden Endpunkten der Kommunikation übertragen. Diese zeitlich begrenzten Schübe der Aktivität werden als „Bursts“ bezeichnet (eine schubweise Paketübermittlung). Dieses Burst-artige Verhalten findet man sowohl in Verkehrskurven von sehr kleinen Zeitabschnitten als auch in den Kurven von sehr langen Abschnitten. Man sagt, der Verkehr besitzt eine *Selbstähnlichkeit* (sieht immer gleich aus, im Großen wie im Kleinen).

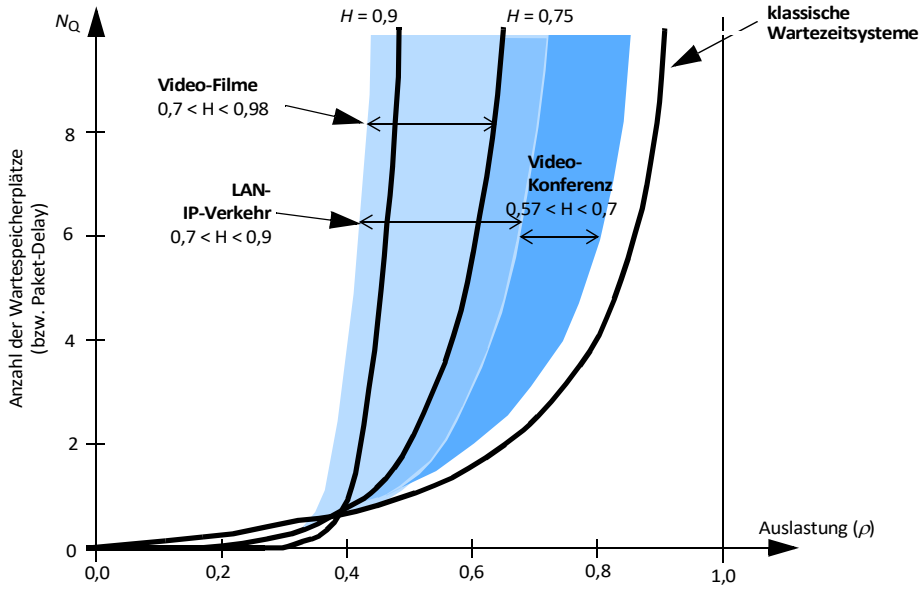


Abbildung 12: Dienste mit unterschiedlichen Verkehrseigenschaften und Hurst-Parameter

Hurst-Parameter

Der Grad der Selbstähnlichkeit wird durch den sog. *Hurst-Parameter (H)* beschrieben. In der Abbildung 12 ist dieser Zusammenhang was unter Selbstähnlichkeit in der Datenkommunikation zu verstehen ist dargestellt. Die dargestellte Belastung eines Netzes ist für unterschiedliche Zeiträume immer gleich. Eine Messung des Datenverkehrs über eine Stunde ergibt beispielsweise die obere Kurve. Legt man hier die Lupe an und betrachtet einen kleinen Ausschnitt dieser Messung (z. B. für 5 min) ergibt sich ein ähnliches Bild, das gleiche passiert, wenn man diesen Abschnitt wiederum vergrößert und nur 500 ms betrachtet. Der Hurst-Parameter gibt den Grad der Selbstähnlichkeit an, er kann zwischen 0,5 und 1 liegen. Eine bekannte Verteilung ist die Gauß'schen Verteilung, hier kann man H mit 0,5 ansetzen. Mit H = 0,5 ergeben sich klare Mittelwerte mit einer begrenzten Varianz. Bei $H=1$ ist der Verkehr in jedem Zeitraum nicht mehr unterscheidbar von dem Verkehr zu einem beliebigen anderen Zeitraum. Mit einem Hurst-Parameter von 0,7 oder 1 sind Mittelwerte kaum bestimmbar und die Varianz der gemessenen Werte strebt gegen unendlich. In der Praxis bedeutet das, dass jede Messung einen anderen Wert ergibt. Selbst bei sehr langen Zeiträumen für die Messungen werden die Werte nicht besser oder genauer – sind immer wieder anders. Der Grad der Selbstähnlichkeit hat daher Auswirkungen auf die Systemauslastung und damit auf die entstehenden Paketlaufzeiten.

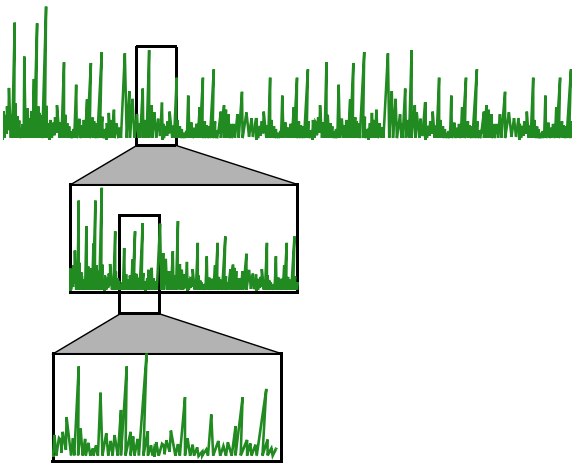


Abbildung 12: Selbstähnlichkeit

Praktische Werte

Mehrere Untersuchungen zeigten, dass für einen typischen Datenverkehr in einem LAN (Ethernet-basierend) der Hurst-Parameter mit ca. $H=0{,}8$ bis 0,9 angenommen werden kann. Das hat Auswirkungen auf die Auslastungskurve und damit auf die Paketlaufzeiten (s. Abbildung 13).

Achtung: In den Kurven werden immer nur Mittelwerte dargestellt, in Spitzen der Belastung können augenblicklich alle Ressourcen (Buffer) belegt sein, auch wenn die Belastung innerhalb der Beobachtungsdauer deutlich kleiner als die maximal mögliche Belastung ist. Daher kann es für einzelne Pakete zu deutlich größeren Wartezeiten kommen, als nach der mittleren Auslastung zu erwarten war.

Zusammenfassend kann man den typischen LAN-Verkehr in einem weiten Bereich von $0{,}6 < H < 0{,}9$ annehmen, und $H = 0{,}83$ wäre ein guter Ansatz für einen gemischten LAN-Verkehr.

3.4 Echtzeitkommunikation in IP-Netzen

Mischung IP und Echtzeitkommunikation

In der Mischung von Daten- und Echtzeitkommunikation wird häufig als Beispiel für die Echtzeitkommunikation die Sprache gewählt. Zum einen, weil speziell die Sprachkommunikation relativ hohe Forderungen an die Übertragung stellt, und zum anderen, weil das menschliche Ohr eine hohe Empfindlichkeit gegenüber schwankenden Verzögerungszeiten und Paketverlust hat. Gleichzeitig stellt die Sprachübertragung eine vergleichsweise geringe Last für das Gesamtsystem dar. Eine erste überschlägige Betrachtung lässt zunächst kaum Probleme bei der Übertragung erkennen – warum sollte VoIP mit 64 kbit/s an einem Netz mit mehreren Mbit/s ein Problem darstellen?

Das klassische Kommunikationsnetz

Klassische Kommunikationssysteme arbeiten häufig als Verlustsysteme. Das alte Telefonnetz ist ein typisches Beispiel hierfür, das Gleiche gilt aber auch für die Mobilkommunikation, in der den Benutzern feste Kanäle für die Kommunikation zugeordnet werden. In Verlustsystemen werden freie Ressourcen mit festgelegten Eigenschaften den anfragenden Kommunikationsquellen für die Dauer der Kommunikation fest zugeordnet.

Diese Eigenschaften ändern sich auch nicht mit zunehmender Auslastung. Sind alle Ressourcen in einem Verlustsystem vergeben, kommt es zu Verlust, d. h., weitere Anfragen werden abgelehnt. Verlustsysteme mit sehr großen Belastungen bedienen die zugeordneten Ressourcen mit den unveränderten Eigenschaften, nur die Anfragen, die über die Leistungsfähigkeit der Systeme gehen, werden nicht bedient.

Bei der Auslegung von Verlustsystemen ist es das Ziel des jeweiligen Netzbetreibers, einen möglichst geringen Verlust für die anfragenden Kunden zu haben. Die Systeme sollen aber auch aus wirtschaftlichen Gründen nicht zu stark überdimensioniert sein.

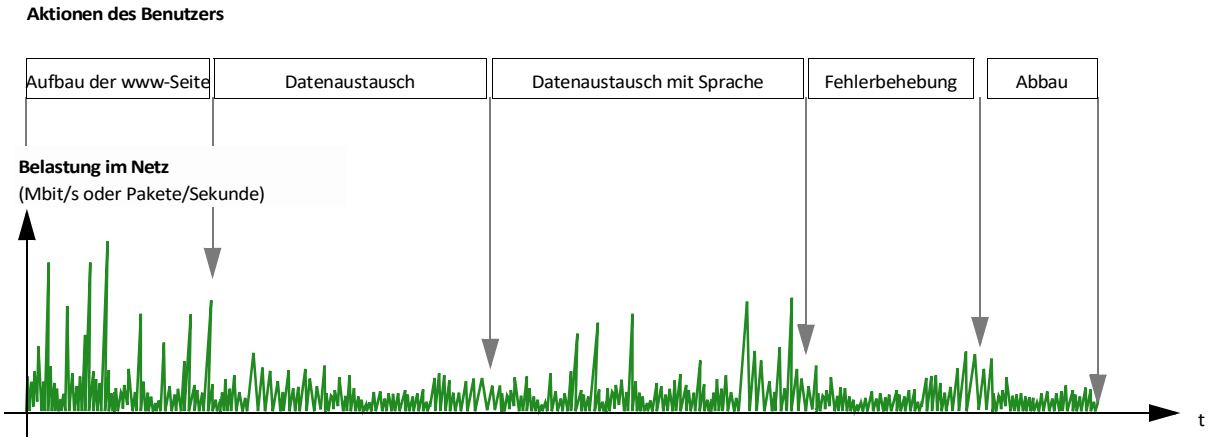


Abbildung 13: Bandbreitenanforderungen/Paketrate im zeitlichen Verkauf

Die Internet-basierte Übertragung

Systeme, in denen die Sprache (oder andere Echtzeitverkehre) zusammen mit Nutzinformationen über das Internet transportiert werden, reagieren anders. Diese Systeme arbeiten als Wartezeitsysteme. Die Eigenschaften der Nutz- und Steuerinformationen für die Sprachübertragung haben sich nicht geändert. Die Datenkommunikation lässt sich, wie oben gezeigt, nicht so einfach beschreiben. Die Pakete der Datenkommunikation beeinflussen aber die Pakete der Nutz- und Steuerdaten für die Sprachkommunikation. Da es in diesen Systemen keine fest zugeordneten Kanäle gibt, können die Datenpakete auch die bereits existierenden Sprachkommunikationen beeinflussen.

Paketverlust

Ein Paketverlust kann auch durch die Überschreitung der Paketlaufzeit oder durch Fehler beim Pakettransport bzw. innerhalb der Netzelemente verursacht werden. Die meisten Netze sind so ausgelegt, dass Paketverlust trotzdem relativ selten ist (< 5 %, oftmals < 2 %). Der Paketverlust kann aber auch bei sehr guter Netzauslegung nicht völlig vermieden werden. Wie oben beschrieben, lässt sich ein IP-basiertes Netz nicht vollständig vorausberechnen. Paketverlust kommt vor, ist aber im Normalfall relativ selten. sind für den Benutzer kaum spürbar. Die Auswirkungen sind größere Paketlaufzeiten und ggf. Paketverlust in der Kommunikation. Beim Surfen im Internet verwendet der Benutzer meist das Transport Control Protocol (TCP), bei dem jeder übertragene Informationsblock gesichert wird. Bei Paketverlust wird der Block einfach noch einmal wiederholt. Der Paketverlust ist also für den Benutzer nicht merklich, die Systembelastung macht sich also nur durch lange Reaktionszeiten, die selten auftreten, bemerkbar.

Belastung des Netzes

Die Paketlaufzeit hängt von der augenblicklichen Belastung des Netzes ab. Normalerweise arbeitet das Netz mit einer sehr geringen Belastung, nur die kurzzeitigen Verkehrsspitzen bringen eine nennenswerte Systemauslastung.

In der Abbildung 13 wurde der reine Datenverkehr eines LAN aus der Abbildung 13 auf die Systemauslastung abgebildet (Systemauslastung steigt mit der Ankunftsrate, die Bedienrate ist für das System konstant). In vorhandenen Systemen werden diese großen Auslastungsfälle eher selten sein (sonst würden sie bereits beim vorhandenen Datenverkehr stören). Die meisten LAN sind kräftig überdimensioniert. Diese Reserven in der LAN-Leistungsfähigkeit und die eher geringen Bandbreiten-Anforderungen des VoIP-Verkehrs verleiten dazu, den Sprachverkehr als eine eher geringere Belastung einzuordnen, die problemlos vom Netz übertragen werden kann.

4 Die Eigenschaften des Internet-Verkehrs

Der klassische IP-Verkehr schwankt von Augenblick zu Augenblick sehr stark. Für diese Kommunikation sind die kurzzeitigen, großen Belastungen gefolgt von relativ langen Ruhephasen typisch – man nennt dieses Verhalten auch einen „Burst-artigen Verkehr“. Diese kurzzeitigen Auslastungen des Systems

4.1 VoIP ist eine große Belastung

Der VoIP-Verkehr bedeutet für das Netz eine sehr große Belastung. Der Codec erzeugt sehr viele, kleine Pakete, die in kurzen Abständen mit RTP und UDP übertragen werden. Die Anzahl der IP-Pakete pro Zeit (Paketankunftsrate) ist dabei für die Netzbelastung bedeutsam, nicht die erforderliche Bandbreite.

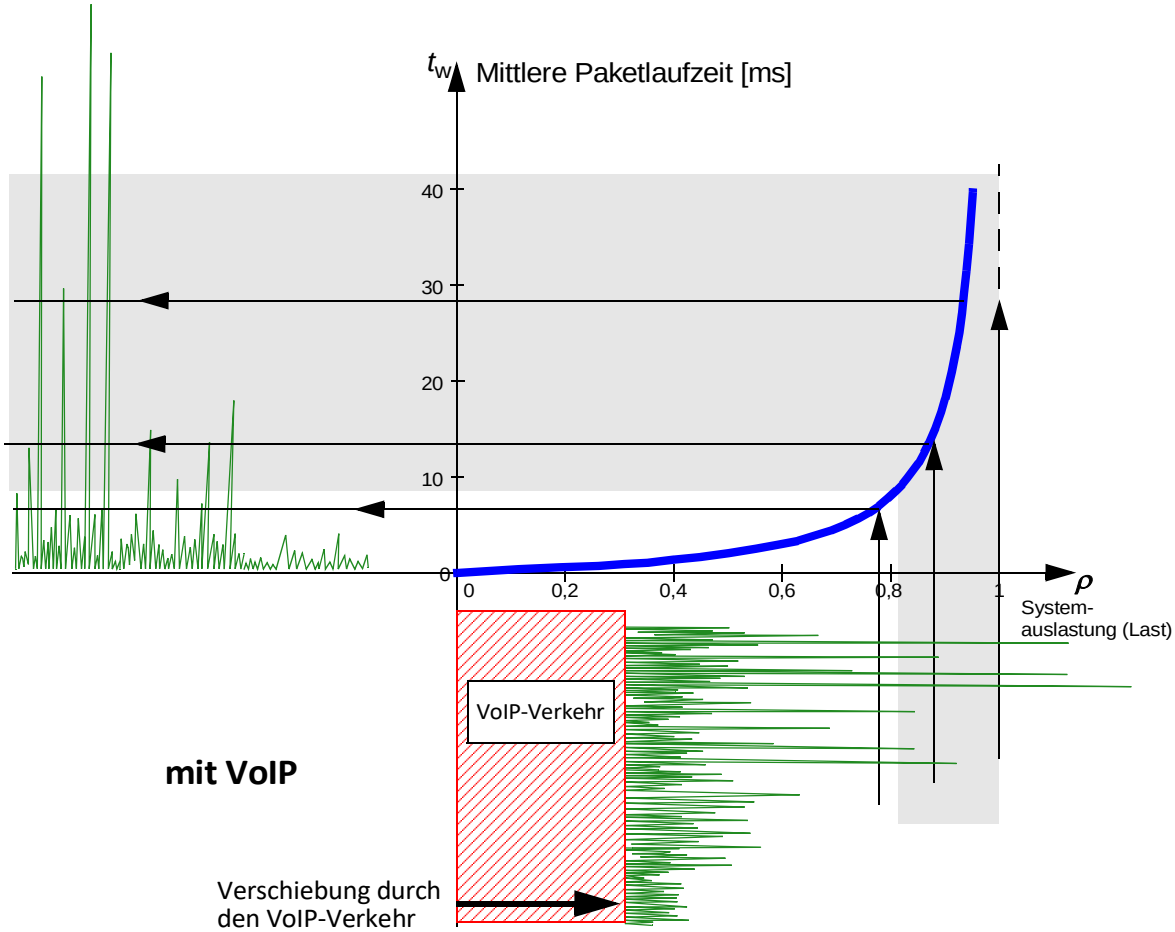


Abbildung 14: Paketlaufzeit in Abhängigkeit von der Belastung des gemischten Verkehrs

Unabhängig von der Paketlänge

Für den Router ist das Paket die Belastung, nicht so stark die Länge des Pakets. Das Routing erfolgt anhand der im Overhead übertragenen Adressen, unabhängig von der Paketlänge. Danach muss das Paket allerdings noch über die gewählte Leitung „abfließen“, dieser Vorgang hängt von der Übermittlungsgeschwindigkeit und der Paketgröße ab. Die VoIP-Belastung schwankt mit dem Verkehrsaufkommen für die Sprachkommunikation, dafür sind die Verbindungen pro Zeit wichtig. Wenn eine Verbindung aufgebaut ist, hat sie allerdings immer die gleiche Paketrate (z. B. alle 20ms ein RTP-Paket mit 230 Byte), der VoIP-Verkehr hat nicht das Burst-artige Verhalten des klassischen Datenverkehrs. Dieser Verkehr kommt zusätzlich in das vorhandene Datennetz, d. h., er ist eine Art „Offset“ zum vorhandenen Verkehr. In der Auslastungskurve wird der Verkehr dadurch nach rechts in den Bereich höherer Auslastung verschoben.

Belastung für alle Anwendungen

Durch die Verschiebung des Verkehrs geraten immer mehr Verkehrsspitzen in den Bereich der großen Belastung und damit der großen Wartezeiten. Für den gesamten Verkehr aller IP-Pakete (Daten wie Sprache) treten längere Verzögerungen und vermehrter Paketverlust auf. Die Zeiten größer Belastung sind jetzt im Vergleich zu vorher öfter und länger. Die Leistungsfähigkeit des Netzes wird für alle Anwendungen geringer. Der Datenverkehr wird durch den VoIP-Verkehr beeinträchtigt, stärker, als die Bandbreitenanforderungen vermuten lassen.

Ideales Netzverhalten

Mit zunehmender Belastung steigt die Durchlaufzeit (Delay) in einem Netz, das lässt sich nicht vermeiden. Ideal wäre es, wenn die Paketlaufzeit bis zum Erreichen der zur Verfügung stehenden Bandbreite oder Leistungsfähigkeit linear ansteigt und dann keine weiteren Pakete transportiert werden.

Dieses Verhalten hatten das klassische digitale Kommunikationsnetz, das ISDN. In den klassischen Kommunikationssystemen ging man immer von einer einheitlichen Kommunikationsart mit einer konstan-

ten Übermittlungsgeschwindigkeit von 64 kbit/s aus. Für höhere Anforderungen konnten eine Anzahl dieser Grundeinheiten zusammengeschaltet werden (N · 64 kbit/s), geringere Anforderungen nutzten diesen Kanal nur teilweise aus. Eine Belegung eines Kanals erfolgte immer für die gesamte Kommunikationsdauer (Dauer des Gesprächs, Dauer der Session). Die Anzahl der Kommunikationskanäle ist in in diesen Systemen begrenzt. Mit steigenden Anforderungen werden immer mehr Kanäle den Verbindungen zugeordnet. Wenn alle Kanäle vergeben sind, erhalten weitere Anfragen eine Absage (z. B. den Besetztton). Die bestehenden Verbindungen werden nicht beeinflusst. Das Netz strebt im Falle großer Belastung gegen die Grenze der Belastung, gegeben durch die Anzahl der zur Verfügung stehenden Kanäle. Die erforderlichen Kanäle werden vorausberechnet, und diese Auslegung wird regelmäßig überprüft. In der Abbildung 15 ist der normalisierte Durchsatz, bezogen auf die normalisierte Belastung, dargestellt. Der Durchsatz und die Belastung werden auf die Leistung, die das System leisten kann, bezogen (gegeben durch die verfügbare Anzahl von Kanälen). Der normierte Durchsatz kann nicht größer als 1 werden, dann sind alle verfügbaren Kanäle vergeben. Weitere Steigerungen der Belastung werden abgewiesen.

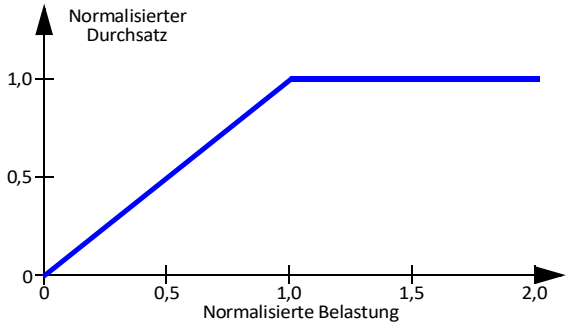


Abbildung 15: Steigende Belastung der klassischen Systeme

Reales Netzverhalten

Die Eigenschaften des IP-basierten-Verkehrs unterscheiden sich in vielen Punkten von denen der klassischen Kommunikation. Hergebrachte Auslegungsrichtlinien können nicht mehr verwendet werden. Die Auslastungkurve in der Abbildung 16 zeigt, dass dieses Netz auf eine zunehmende Belastung mit wachsenden Wartezeiten und damit sich rasch füllenden Warteschlangen reagiert. In der Praxis versucht man natürlich immer, ein System möglichst leistungsfähig zu betreiben. Dies bedeutet aber, dass man versucht, den Zustand der Überlast (Congestion) möglichst frühzeitig zu vermeiden. Reale Systeme reagieren auf die sinkende Leistungsfähigkeit. Bis zu dem Punkt A in der Abbildung steigt der Durchsatz im System linear an (no Congestion). Ab dem Punkt A nehmen die Laufzeiten durch das nun belastete System stark zu, innerhalb eines Zeitfensters steigt damit der normalisierte Durchsatz nicht mehr linear mit der Last an (moderate Congestion). Ab dem Punkt B ist das System in einem Überlastungszustand (severe Congestion).

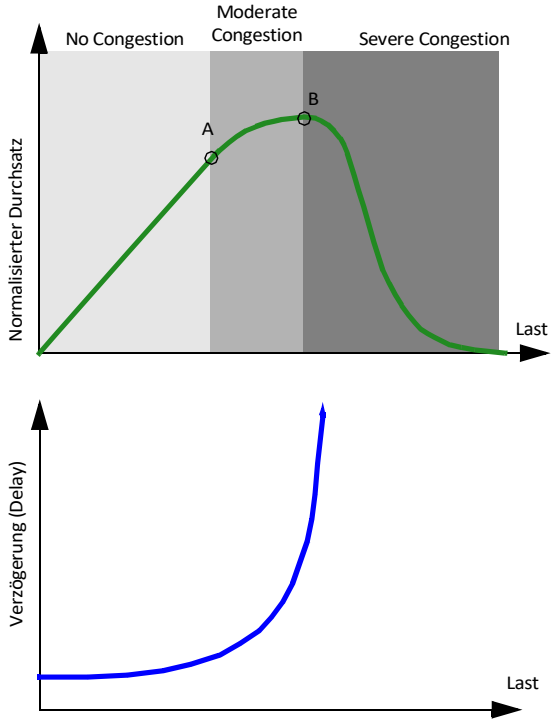


Abbildung 16: Leistungsfähigkeit realer Systeme

Ab dem Punkt B kann die angebotene Last nicht mehr verarbeitet werden. In diesem Zustand:

- nehmen die Paketlaufzeiten drastisch zu,
- sind die Buffer in den Netzelementen gefüllt,
- werden Pakete verworfen.

Aufgrund von verschiedenen Mechanismen in den Netzkomponenten steigt die Last aber noch weiter. Was passiert nach Punkt B:

- Die Router erzeugen mehr Verkehr, Update der Routen (Pfade).
- Die Quellen wiederholen ihre Sendungen, manche Anwendungen senden bei höheren Paketverlusten die Pakete gleich doppelt.

All dies führt zu einer weiteren Verschärfung der Situation. Anders ausgedrückt: Mit steigender Belastung strebt das System bezüglich des Durchsatzes gegen 0. – Nicht, wie die klassischen

Kommunikationsnetze, gegen eine vorgegebene maximale Belastung! Der Grund hierfür ist die Auslastungskurve, die nicht linear ansteigt. Die Paketlaufzeit (Delay) steigt exponentiell, die Buffer laufen über und in der Protokollbearbeitung laufen die Timer ab. In der Praxis sind extrem lange Paketlaufzeiten aus diesem Grund nicht realistisch. Mit zunehmender Auslastung streben reale Systeme daher an einen Grenzwert der Paketverzögerungszeit für alle tatsächlich transportierten Pakete. Gleichzeitig kommen bei zunehmender Belastung immer weniger Pakete durch das Netz durch – d. h. der Paketverlust steigt. Der Grenzwert für die maximale reale Paketlaufzeit hängt von viele netzspezifischen Faktoren, wie der Übermittlungsgeschwindigkeit, Buffergröße usw.) ab.

4.2 Beurteilung der Netzbelastung

Für die Beurteilung der Netzauslastung wird oft die genutzte Übertragungskapazität im Vergleich zu der möglichen Übertragungskapazität verwendet. Für die Übertragung der eigentlichen Nutzdaten müssen aber Pakete mit einer begrenzten Länge gebildet werden, die noch mit den Paketköpfen (den Header) der verschiedenen OSI-Sichten versehen werden müssen. Die Paketgröße ist dabei sehr variabel und hängt von verschiedenen Parametern ab. Im Standardformat des Ethernet-Frames (IEEE 802.3 und auch Ethernet II) können im Payload-Feld minimal 46 Byte und maximal 1500 Byte übertragen werden [Rec08], daneben gibt es für die schnellen Netze (z. B. Gigabit-Ethernet) sog. Jumbo-Frames mit 9000, 16 000 oder gar 64 000 Byte eine genaue Größe ist hier nicht standardisiert. Eingeführt wurden die großen Pakete von verschiedenen Herstellern um insbesondere die Übertragung von großen Daten effizienter zu gestalten. Das Verhältnis von Nutzdaten zur Länge des gesamten Pakets wird mit zunehmender Paketlänge immer günstiger, weil der sog. Overhead für jedes Paket gleich ist. Der Overhead ist die Summe der verschiedenen Kopffelder bei der Übertragung von Datenpaketen. Der Overhead ist beispielsweise für den Transport der Pakete, die Sicherung der Übertragung oder zur Regelung des Zugriffs auf ein Übertragungsmedium notwendig – es sind aber keine Nutzdaten. Der Overhead benötigt eine Übertragungskapazität, die dann für die eigentliche Nutzdatenübertragung nicht mehr zur Verfügung steht. Standardmäßig sind dies die Kopffelder von TCP mit 20 Byte, IP ebenfalls 20 Byte und Ethernet mit 14 Byte für den Header, 4 Byte für die CRC-Informationen und 8 Byte für die Preambel und zwei weitere Byte für die Typ-Kennzeichnung: zusammen sind das 26 Byte für den Ethernet-Rahmen + 20 Byte für IP und 20 Byte für TCP.

4.3 Warteschlangen

Es gibt Quellen, bei denen ist die Ankunftsrate konstant, d. h. die Pakete werden in einem festen, immer gleichen zeitlichen Raster gesendet. Der klassische Codec G.711 für die Sprachübertragung erzeugt durch die Abtastung eines analogen Signals beispielsweise alle 125 µs ein Byte. Mehrere Bytes werden zu einem Paket zusammengefasst, da die Anzahl der Byte pro Paket immer gleich ist, ist auch die Paketrate *l* immer gleich. In der Praxis werden häufig 160 Byte zu einem Paket gebündelt, dies entspricht dann einer Paketankunftsrate von 50 Paketen/s.

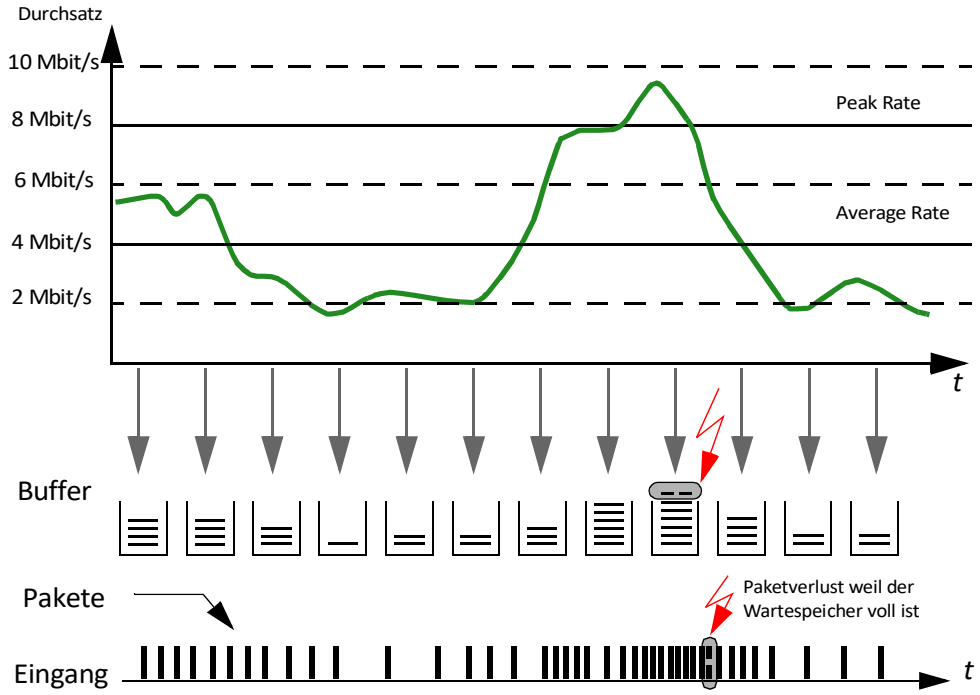
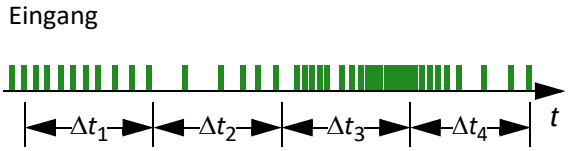


Abbildung 17: Schwankungen der Paketankunftsrate

In der Datenkommunikation schwanken die Paketankunftsraten von Augenblick zu Augenblick sehr stark. Die eintreffenden Pakete transportieren Nutzinformationen, die in eine effektive Übermittlungsrate (Nutzinformationen in bit/s) umgerechnet werden können. Wie für die mittlere Paketankunftsrate, lässt sich auch für die Übermittlungsrate rechnerisch immer ein Mittelwert bestimmen – dieser wird dann oft für eine Abschätzung der Systemauslastung angenommen. Die augenblickliche Auslastung kann aber um ein vielfaches höher sein.

In der Abbildung 17 wird eine Datenübertragung mit einer schwankenden Übermittlungsrate dargestellt. Die Daten werden immer in Form von Paketen übertragen, die in einem Netzelement vor der weiteren Verarbeitung (z. B. dem Routing) in einen Zwischenspeicher (Buffer) aufgenommen werden. Wenn zu viele Pakete innerhalb einer bestimmten Zeit von der Quelle produziert werden, kann die Kapazität des Speichers überschritten werden, d. h. weitere, eintreffende Pakete finden keinen freien Speicherplatz in der Warteschlange, sie gehen verloren. Im unteren Teil der Abbildung 17 sind die eintreffenden Pakete dargestellt. Zur Ermittlung der Ankunftsrate dividiert man die Anzahl der eingetroffenen Pakete innerhalb eines Zeitabschnitts durch die Dauer des Abschnitts.

Die Pakete aus dem Beispiel in der Abbildung 18 werden etwas vereinfacht in gleiche Zeitabschnitte mit jeweils 1 ms Dauer unterteilt. Die Division der Anzahl der Pakete je Zeitintervall liefert völlig unterschiedliche Werte für die jeweilige Ankunftsrate (*l*). In diesem vereinfachten Beispiel konnten in der Darstellung aus Übersichtsgründen nur sehr wenige Pakete je Zeitabschnitt übertragen werden. Reale Werte für *l* würden daher noch viel stärker schwanken. Die jeweilige Paketlänge spielt für die Ermittlung der Ankunftsrate keine Rolle. Entsprechend der sehr unterschiedlichen Ankunftsraten füllen sich die Buffer an den Eingängen der Netzelemente mit den eintreffenden Paketen. Jeder Buffer hat eine begrenzte Größe, d. h. es kann immer nur eine begrenzte Anzahl von Paketen gespeichert werden. Wie viele Pakete ein Buffer aufnehmen kann, hängt auch von der Paketlänge ab. Lange Pakete benötigen einen größeren Speicherbereich als kurze Pakete.



Ankünfte:

$$\lambda_1 = \frac{10 \text{ Pakete}}{1 \text{ ms}} = 10\,000 \text{ Pakete/s}$$

$$\lambda_2 = \frac{5 \text{ Pakete}}{1 \text{ ms}} = 5\,000 \text{ Pakete/s}$$

$$\lambda_3 = \frac{16 \text{ Pakete}}{1 \text{ ms}} = 16\,000 \text{ Pakete/s}$$

$$\lambda_4 = \frac{9 \text{ Pakete}}{1 \text{ ms}} = 9\,000 \text{ Pakete/s}$$

Abbildung 18: Berechnung der Paketankunftsrate λ

4.4 Traffic Shaping

Die Datenpakete in einem Netz werden spontan von den Endsystemen und den dort laufenden Anwendungen unabhängig und zufällig erzeugt. Eine kleine Anfrage an einen Server kann dabei eine Vielzahl von Paketen verursachen. Die Paket-Ankunftsrate ist sehr unregelmäßig und schwankt sehr stark. Eine große Anzahl von Paketen wird *büschelartig* (Burstartig), gefolgt von langen Ruhephasen übertragen. Dieser Verkehr ist kaum im Voraus zu berechnen. Es lassen sich zwar immer Mittelwerte berechnen, aber die Varianz dieser Werte strebt gegen Unendlich, was die Mittelwerte unbrauchbar macht.

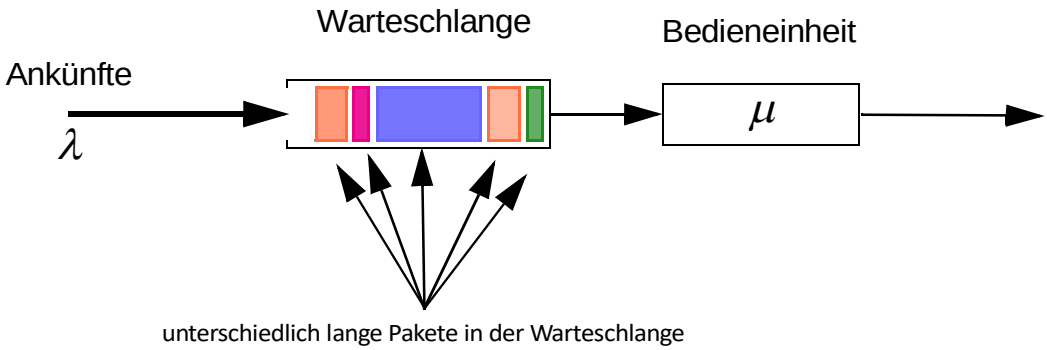


Abbildung 19: Pakete in der Warteschlange

Verkehrsformung

Mit einem relativ einfachen Mechanismus, dem *Traffic Shaping*, kann man daraus einen Verkehr mit einem vorgegebenen Grenzwert formen. Die eintreffenden Pakete werden in einem Speicher mit einer begrenzten Größe zwischengespeichert. Ein fest vorgegebener Takt ermöglicht in regelmäßigen, immer gleichen Abständen die Entnahme eines Pakets. Dieses Verfahren der Paketentnahme wird auch als *Leaky Bucket* bezeichnet. Der Takt der Paketentnahme entspricht der Vorgabe für einen vereinbarten, maximalen Durchsatz. Treffen viele Pakete in einem sehr kurzen Abstand am Eingang ein, müssen einige Pakete warten, bis eine Freigabe für die Entnahme eines Paketes eingetroffen ist. Dadurch werden die Verkehrsspitzen auf eine etwas längere Zeit verteilt. Der Eingangsspeicher ist in seiner Größe begrenzt, d. h. wenn zu viele Pakete in zu kurzer Zeit eintreffen, kann es sein, dass kein Speicherplatz mehr frei ist, dann geht ein Paket verloren. Am Ausgang kann die Paketrate, die vorgegebene maximale Rate nicht überschreiten. Wird dieser Mechanismus für jeden Eingang angewendet, verhält sich das gesamte Netz deterministisch und vorhersagbar. Überlastungen einzelner Systeme im Netz werden dadurch vermieden, die Paketlaufzeiten sind begrenzt und schwanken nicht so stark und der Paketverlust wird geringer. Zu Paketverlust kommt es nur bei deutlichen Überschreitungen vereinbarter Datenraten am Eingang des Netzes. Diese Überwachung des Verkehrs und die Garantie der Einhaltung von vorgegebenen Parametern wird auch als *Policing* bezeichnet.

Beispiel für die Arbeitsweise von Traffic Shaping

In dem folgenden Beispiel werden von drei unterschiedlichen Datenquellen unabhängig voneinander Pakete gesendet (Quelle 1 bis 3). Parallel werden von einer VoIP-Quelle (Quelle 4) in regelmäßigen Abständen relativ kleine Pakete über die gleiche Schnittstelle gesendet.

Wird der Verkehr der vier Quellen ohne Regulierungsmaßnahmen zusammengefasst, setzen sich vor allem die größeren Datenpakete durch. Die Verzögerungszeiten für die kleinen Pakete sind sehr groß und variieren sehr stark (s. Abb. 20).

Mit der Anwendung des Traffic-Shaping-Verfahrens bleiben die Verkehrsmuster nicht so, wie sie von den Quellen erzeugt werden. Um für alle Quellen eine gewisse Qualität der Übertragung mit begrenzten Paketlaufzeiten zu ermöglichen, müssen die Paket-Bursts auf eine maximale Paketrate angepasst werden. Für jede Quelle kann dabei eine eigene maximale Paketrate vorgegeben werden. Die Vorgaben werden entweder durch eine Signalisierung (Session-Aufbau) am Anfang der Verbindung oder (meistens) durch Management-Einträge von den Netzbetreibern definiert. In der Abbildung 21 wird zunächst der Verkehr der Quelle 2 angepasst. Das Freigaberaster entspricht der vereinbarten Paketrate (und damit einer bestimmten maximalen Übertragungsgeschwindigkeit), die Quelle kann also den vereinbarten Durchsatz auch tatsächlich übertragen. Durch dieses Verfahren gehen keine Pakete der Verkehrsquellen verloren. Das Traffic-Shaping-Verfahren ändert nur den zeitlichen Abstand der Pakete.

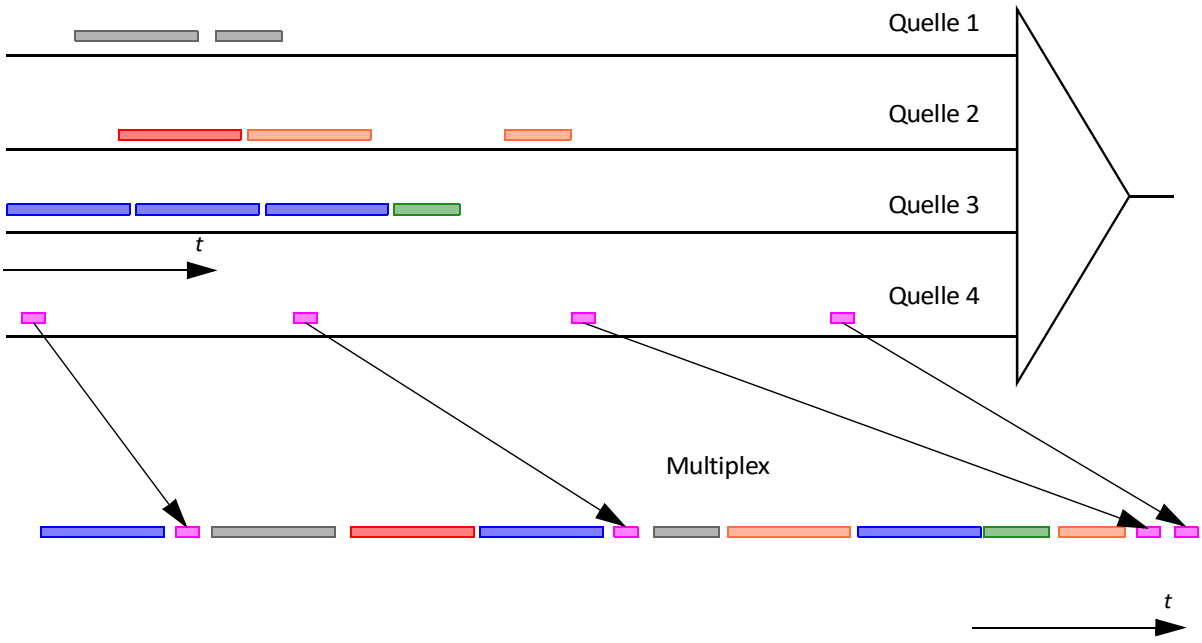


Abbildung 20: Ohne Policing

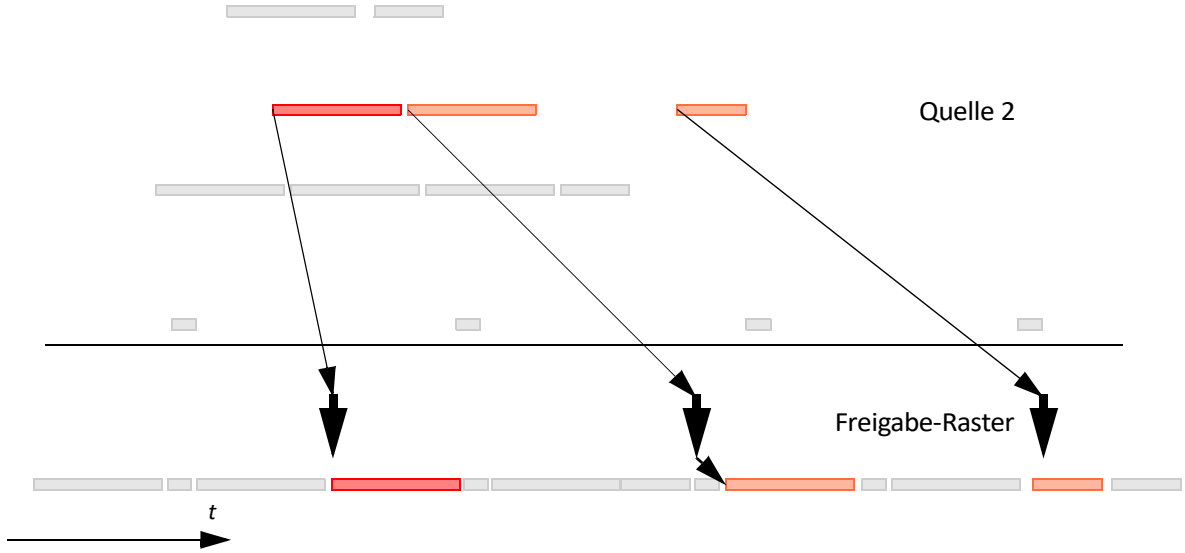


Abbildung 21: Anpassung des Verkehrs der Quelle 2

Verkehrsmischung

Werden nun die angepassten Verkehre zusammengefasst, entspannt sich die Situation beispielsweise für die Echtzeitkommunikation der Quelle 4. Die Verzögerungszeiten für die Pakete der Quelle 4 sind deutlich kleiner und sie schwanken nicht mehr so stark (s. Pfeile in der Abb. 20 und Abb. 21). Die Effekte der Verkehrsformung mit dem Traffic-Shaping-Verfahren werden bereits an diesem einfachen Beispiel deutlich. In der Realität sind die Datenpakete im Verhältnis zu den Paketen der Echtzeitkommunikation viel größer, d. h., die Behinderung der Echtzeit-Pakete ist noch viel stärker als hier darstellbar.

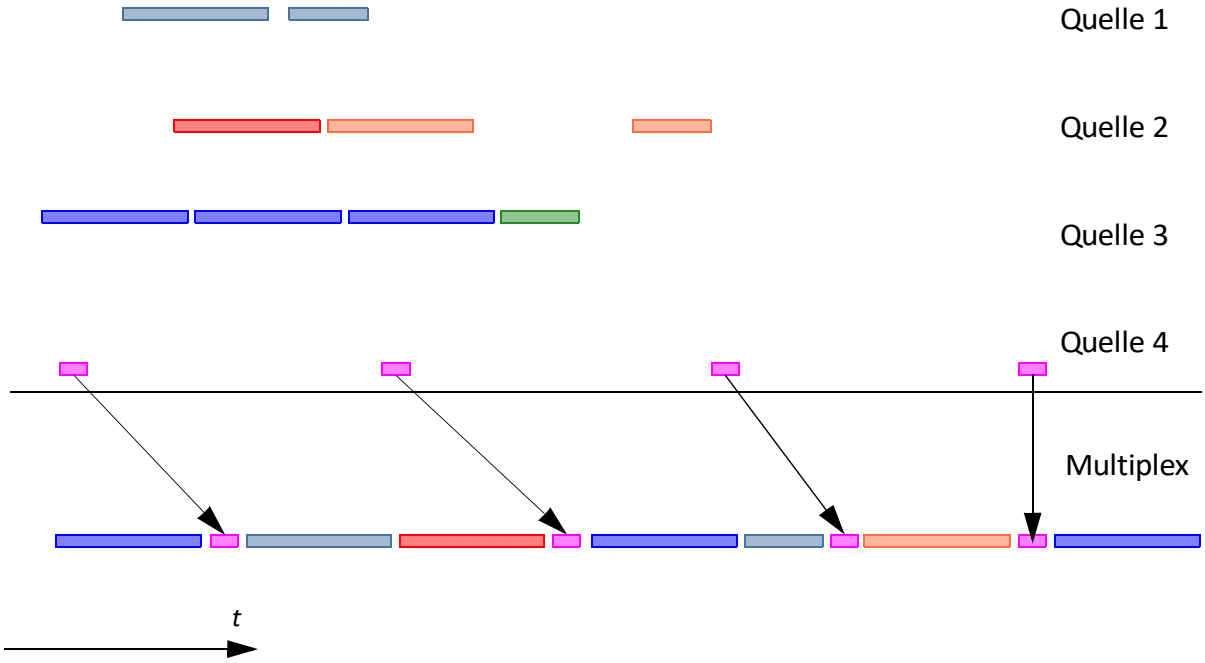


Abbildung 22: Mit Policing (angepasste Verkehre)

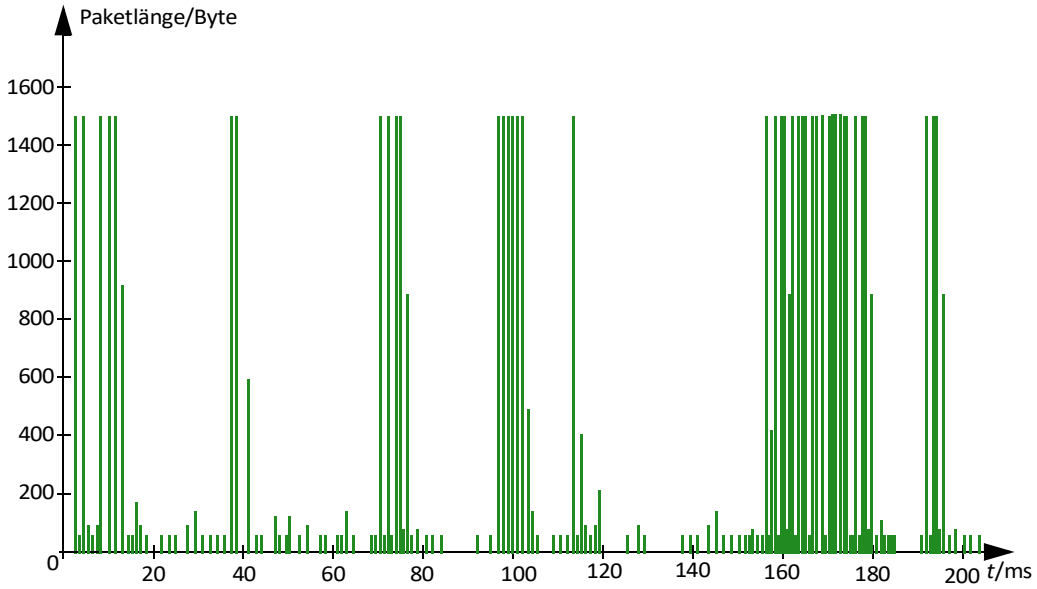


Abbildung 23: Aufgezeichneter Datenverkehr

4.5 Verkehrsmischungen

In den folgenden Abbildungen ist ein Datenverkehr über 200 ms dargestellt. Für ein Intervall von jeweils 20 ms werden die Mittelwerte für die Ankunftsrate (λ) der eintreffenden Pakete, die Bedienrate (μ) für den Transport mit 10 Mbit/s (alles andere, wie Eingangswartespeicher, Paketdurchlaufzeit durch einen Switch oder Router usw. sind hier noch nicht berücksichtigt), die Systemauslastung ($\rho = \lambda/\mu$) und die Durchlaufzeit für ein durchschnittliches Paket (T) dargestellt. Mit den Darstellungen und Berechnungen sollen die Auswirkungen von Verkehrsmischungen betrachtet werden. Oft wird die augenblickliche Belastung und die erwartete Belastung nur anhand der Übertragungsgeschwindigkeiten beurteilt. Dabei werden die Mittelwerte der Kommunikation berechnet und im Vergleich zu der möglichen Geschwindigkeit der Datenübertragung ins Verhältnis gesetzt. Dieses Ergebnis wird dann als ein Maß für die Systemauslastung genommen. Tatsächlich ist das aber keine praktikable Methode, um die Auswirkungen zusätzlichen Echtzeitverkehrs abschätzen zu können. Die Mittelwertbildung verschleiert kurzzeitige Belastungen, deren Auswirkungen (lange Paketlaufzeiten und ggf. Paketverlust) dann aber im praktischen Betrieb doch deutlich spürbar werden. Die Probleme in der Übertragung werden etwas deutlicher, wenn man die gesamte Übertragung in kleinere Zeitintervalle aufteilt und für jedes Intervall einzeln die Auslastung beurteilt. Im folgenden Beispiel wird dann aufgezeigt, wie ggf. auftretende kurzzeitige Auslastungen und die damit verbundenen langen Paketlaufzeiten durch Traffic Shaping vermieden werden können.

In diesem Beispiel wurden innerhalb von 200 ms insgesamt 136 Pakete mit (zusammen) 64 330 Byte aufgezeichnet (s. Abb. 23), die Datenübertragung erfolgte mit einer Geschwindigkeit von 10 Mbit/s. In der Abbildung sind auf der Zeitachse alle eintreffenden Pakete mit ihrer jeweiligen Paketlänge (y-Achse) dargestellt. Im weiteren wird zunächst nur dieser reine Datenverkehr ohne eine Mischung mit Paketen aus Echtzeitübertragungen betrachtet.

Die vereinfachte Beurteilung

Im ersten Ansatz einer sehr vereinfachten Beurteilung der Verkehrsauslastung nimmt man einfach die übertragene Verkehrsmenge, teilt diese durch die Beobachtungszeit und setzt dies in Verhältnis zur möglichen Übertragungsrate. In dem Beispiel bedeutet dies:

In dem Beispiel werden 136 Pakete mit zusammen 64 330 Byte übertragen.
Diese Übertragung dauert 200 ms, die Übertragungsrate ist dann:

$$\frac{64\,330 \cdot 8\ \text{bit}}{0{,}2\ \text{s}} = 2\,573\,200\ \text{bit/s} = 2{,}6\ \text{Mbit/s}$$

Bei einer Übertragung mit 10 Mbit/s hat dieses System eine Auslastung von:

$$\frac{2{,}6\ \text{Mbit/s}}{10\ \text{Mbit/s}} = 0{,}26 = \underline{\underline{26\ \%}}$$

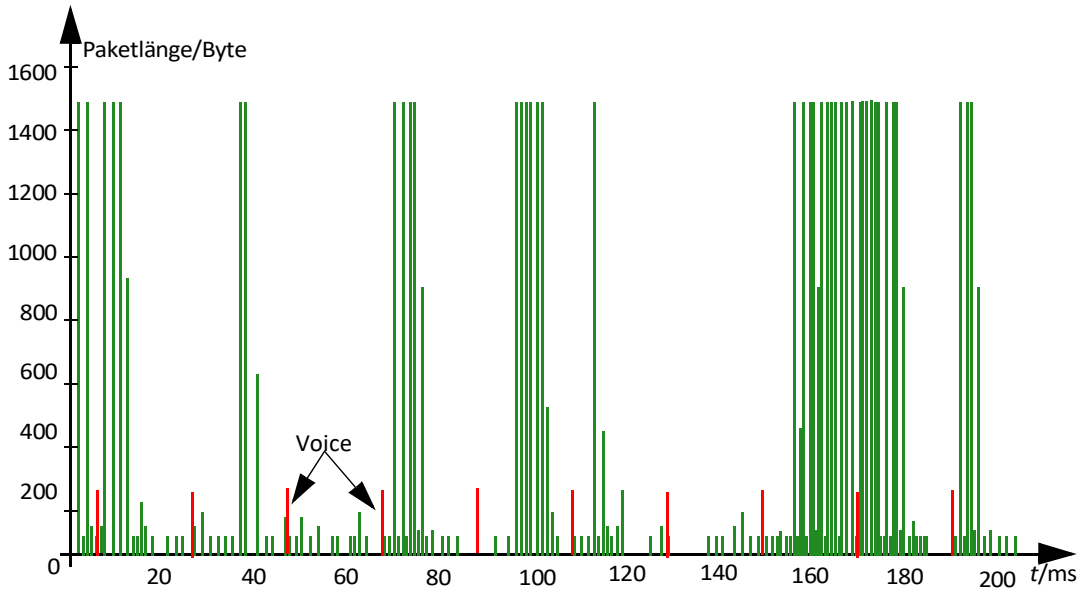


Abbildung 24: Datenverkehr plus ein VoIP-Gespräch

Die 136 Pakete in den 200 ms stellen also eine Belastung von 26 % dar. Kommt jetzt ein VoIP-Gespräch mit dem Codec G.711 dazu (Abb. 24), wird alle 20 ms ein Paket mit jeweils 200 Byte zusätzlich übertragen. Addiert man jetzt einfach die 80 000 bit/s (200 Byte alle 20 ms) zu den bereits vorhandenen 2,57 Mbit/s ergibt das 2,65 Mbit/s, das sind ca. 27 % Auslastung der Übertragungsstrecke.

Beurteilung der Belastung je 20-ms-Intervall

Betrachtet man die Datenübertragung (noch ohne die VoIP-Übertragung) in jedem 20-ms-Intervall einzeln ergibt sich ein anderes Bild. Hier zeigt sich, dass jeder einzelne Abschnitt ganz unterschiedlich belastet ist. Für jeden einzeln 20-ms-Abschnitt wird die Ankunftsrate und die Bedienrate ermittelt:

Ankunftsrate:

$$\lambda = \frac{\text{Anzahl der Pakete}}{0{,}02\ \text{s}} = X\ \text{Pakete/s}$$

Zur Ermittlung der Leistungsfähigkeit des Systems (die Bedienrate) je 20-ms-Abschnitt werden die Längen der jeweiligen Pakete innerhalb eines Abschnitts aufaddiert und anschließend durch die Anzahl der Pakete geteilt, das ergibt die mittlere Paketlänge. Die Übertragungsgeschwindigkeit C = 10 Mbit/s wird dann durch die mittlere Paketlänge in bit geteilt:

mittlere Paketlänge:

$$\ell_m = \frac{\text{Summe aller Paketlängen in einem 20-ms-Intervall}}{\text{Anzahl der Pakete}}$$

Bedienrate:

$$\mu = \frac{C}{\ell_m} = Y\ \text{Pakete/s}$$

Die Systemauslastung kann als $\rho = \lambda/\mu$ berechnet werden. Diese Auslastung wechselt in den einzelnen Abschnitten zwischen 1 % und 96 %. Im vorletzten Abschnitt ist die Belastung allein durch die laufende Datenkommunikation bereits 96 %, die Durchlaufzeit für ein Paket liegt in diesem Fall bei 28,44 ms (Abb. 25). In den anderen Abschnitten bleibt die Durchlaufzeit unter einer Millisekunde. Mit einem zusätzlichen VoIP-Gespräch steigt die Belastung im vorletzten Abschnitt auf 97 % und 36,35 ms, also ca. 8 ms mehr als ohne VoIP. Mit vier VoIP-Gesprächen (Abb. 25 rechts) könnte man nun 4 · 8 ms = 32 ms erwarten, es sind aber mehr als zwei Sekunden! Bei einer Belastung mit vier VoIP-Gesprächen (Abb. 25 rechts) steigt die Auslastung auf fast 100 % und die Paketdurchlaufzeit dadurch auf über zwei Sekunden! Man kann gut erkennen, dass mit vielen Paketen in einem 20-ms-Intervall die Ankunftsrate steigt und mit vielen kleinen Paketen μ größer wird. Die Systemauslastung ρ ist λ/μ und wird mit großen λ und kleinem μ größer.

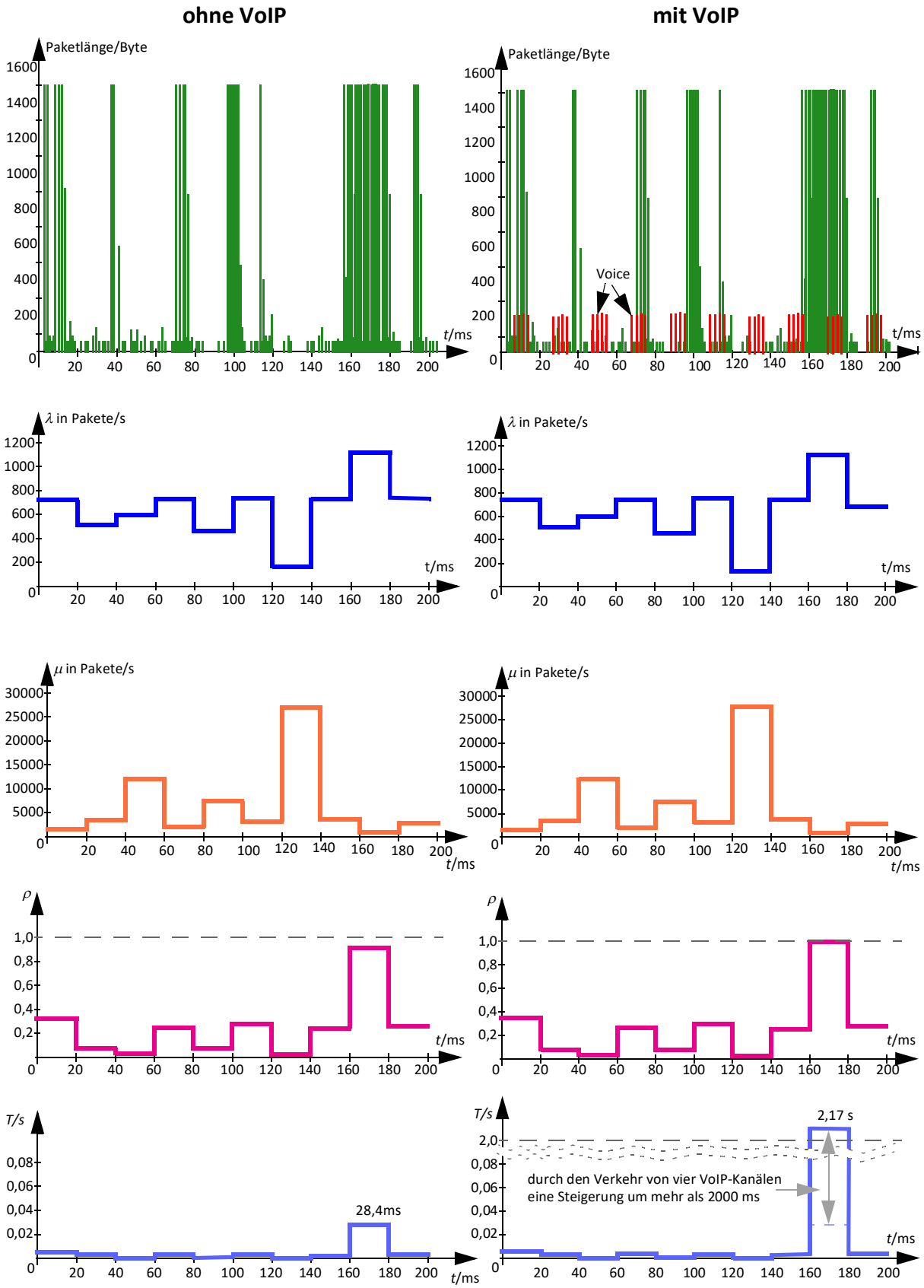


Abbildung 25: Analyse des Datenverkehrs (links ohne, rechts mit VoIP)

Mit der Einführung von Traffic Shaping kann sich diese Situation insgesamt deutlich entspannen. Wenn die Datenkommunikation auf 3 Mbit/s beschränkt wird, entspricht das, bei einer mittleren Paketlänge von 400 Byte, ca. 937,5 Pakete je Sekunde oder 18,75 Pakete je 20-ms-Intervall. In den meisten 20-ms-Intervallen ändert sich nichts. Nur im vorletzten 20-ms-Intervall entspannt sich die Situation sehr deutlich. Die mittlere Wartezeit sinkt von 2,17 s auf 2,41 ms und die Paketrate fällt von 1150 Pakete/s auf 950 Pakete/s, dafür steigt die Paketrate im letzten Intervall von 750 Pakete/s auf ebenfalls 950 Pakete/s. Der Verkehr wurde also auf das folgende 20-ms-Intervall verschoben. Für die Datenkommunikation bedeutet dies, dass einige wenige Pakete maximal 20 ms später ankommen. Dies ist insgesamt aber eine deutliche Verbesserung im Vergleich zur Ausgangssituation ohne VoIP-Verbindungen (Abb.26).

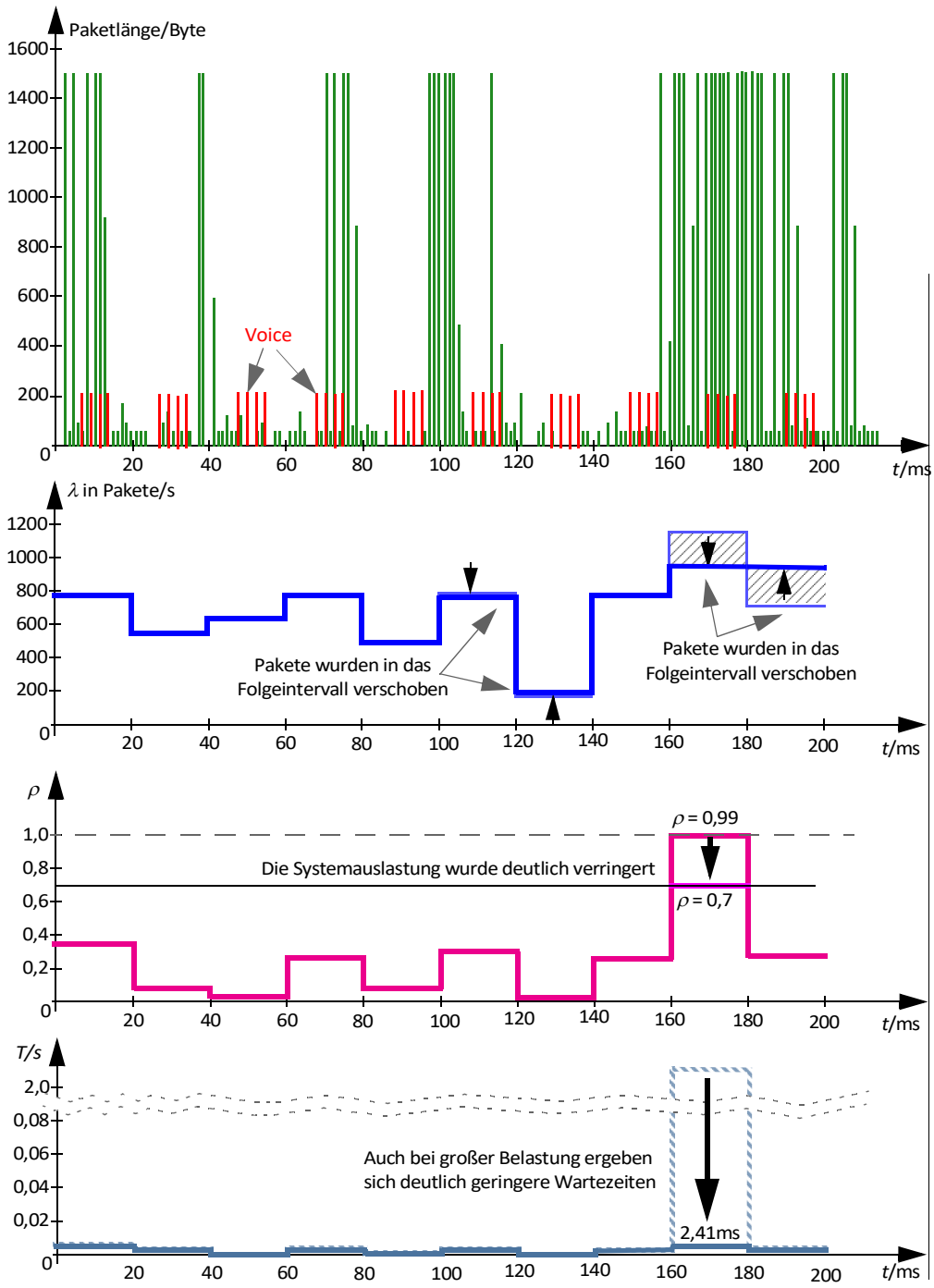


Abbildung 26: Datenverkehr plus vier VoIP-Gespräche und Traffic Shaping

5 Auslegung in VoIP-Systemen

Die Anforderungen, d. h. die Verkehrswerte der Teilnehmer und das von diesen erzeugte Verkehrsangebot bleibt! Für die externe Kommunikation ist ein Angebot von 20 Erl zu verarbeiten, die Frage in VoIP-Systemen ist: Wie viel Bandbreite ist hierfür erforderlich?

5.1 Berechnung der erforderlichen Bandbreite

5.1.1 Der Codec

Unterschiedliche Codecs verwenden verschiedene Übertragungsraten (ISDN, G.711: 64 kbit/s, GSM: 13 kbit/s, AMR-Codec: 4,75 bis 12,2 kbit/s) und arbeiten ggf. mit einer unterschiedlichen Abtastraten (Sampling-Rate z.B. 8 kHz, 16 kHz ...). In TDM-Systemen ist die Basis immer der 64-kbit/s-Kanal mit dem Codec G.711, in VoIP gibt es keine „eingeprägte“ Bandbreite oder festgelegten Codec. G-711 wird jedoch auch in den allermeisten VoIP-TK-Systemen als voreingestellte Option angeboten. Um die nachfolgenden Berechnungen vergleichbar zu halten, wird dafür der Codec G.711 exemplarisch zugrunde gelegt.

5.1.2 RTP-Parameter

Die codierten Sprachinformationen werden mit dem Real-Time-Transport Protocol (RTP) übertragen. Hierbei werden mehrere Sprachproben zu einem Paket zusammengefasst (s. Abb. 27)

Wie viele Sprachproben ein RTP-Paket bilden, ist nicht festgelegt und kann im System konfiguriert werden. In der Beispielausschreibung gibt es hierzu keinen Hinweis.

Welcher Wert wird für die Größe des RTP-Pakets gewählt?

- Mit 80 Sprachproben je RTP-Paket und mit dem Codec G.711 wird alle 10 ms ein RTP-Paket von der Quelle gesendet.
- Werden 160 Sprachproben mit dem Codec G.711 übertragen, wird alle 20 ms ein RTP-Paket gesendet.
- Mit 240 Sprachproben und G.711 wird alle 30 ms ein RTP-Paket erzeugt.
- Mit 320 Sprachproben und G.711 wird alle 40 ms ein RTP-Paket gesendet.

Ist das für die Auslegung der Bandbreite wirklich wichtig? Hat diese Festlegung Auswirkungen auf die Leistungsfähigkeit des Systems?

5.1.3 Gewählte Annahmen

Für die weitere Berechnung wurde der Codec G.711 gewählt, und die RTP werden alle 20 ms gesendet. Eine weitere Annahme: Um die Berechnungen vergleichbar zu machen, werden zur Übertragung die 2,048-Mbit/s-Systeme (aber ohne Kanaleinteilungen) verwendet.

5.1.4 Erste Näherung

In der ersten Näherung kann man die Belastung berechnen, indem man die Codec-Daten und die zusätzlichen Overheads berücksichtigt: 64 kbit/s + Overhead (RTP, UDP, IP, Layer 2, Layer 1), bei G.711 und alle 20 ms ein RTP-Paket macht das: 160 Byte (RTP) + 70 Byte (RTP 12, UDP 8, IP 20 und Ethernet VLAN 30 Byte) = 230 Byte alle 20 ms.

Übermittlungsrate ist dann (230 Byte · 8 bit)/0,02s = 92 kbit/s.

Also trägt unsere 2,048-Mbit/s-Leitung jetzt nur noch 2048 kbit/s/92 kbit/s = 22,3 also 22 Kanäle. Für die oben geforderten 30 Kanäle braucht man also mit VoIP 1,4 Systeme mit jeweils 2,048-Mbit/s.

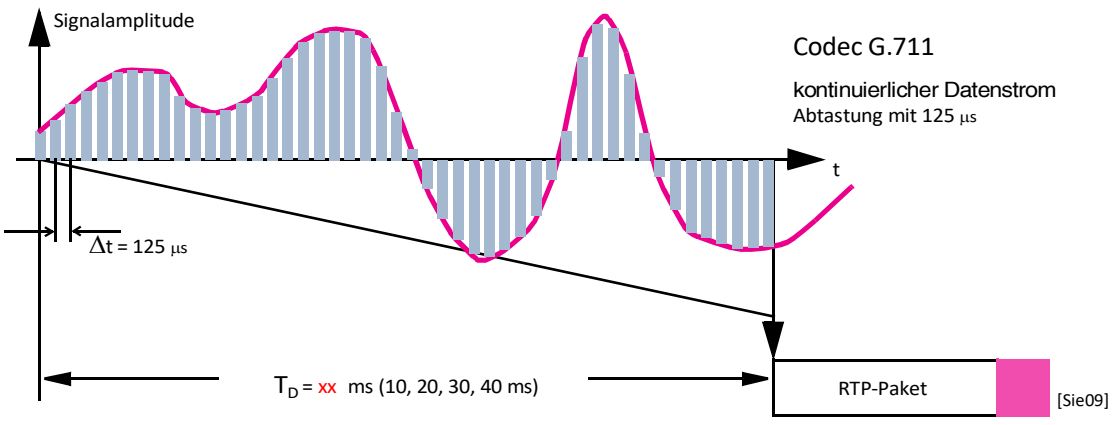


Abbildung 27: Viele Sprachproben bilden ein Paket

Diese einfache Berechnung ist für den ersten Ansatz ganz brauchbar, aber Vorsicht! Eigentlich kann man so nicht rechnen! Diese Berechnung kann man nur anstellen wenn keine weiteren Datenpakete mit der gleichen Leitung transportiert werden oder durchgängig exklusive, virtuelle Kanäle mit VLAN oder MPLS auf der Leitung zur Verfügung gestellt werden.

Grundsätzlich kann man nicht einfach die Bandbreiten mit der Anzahl der VoIP-Kanäle multiplizieren, um die erforderliche Gesamtbandbreite zu berechnen, denn: Die einzelnen Übertragungen mischen sich nicht auf Bit-, sondern auf Paketebene!

Bit-Multiplex:

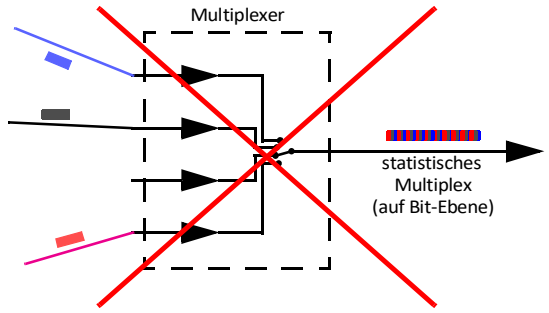


Abbildung 28: Kein Multiplex auf Bitebene

Bei der paketorientierten Übertragung wird ein RTP-Paket nach dem anderen übertragen. Für die Übertragung eines Pakets wird eine bestimmte Zeit benötigt (abhängig von der Übertragungsrate). In dieser Zeit ist die Leitung durch den Pakettransport belegt. Einige Pakete müssen daher warten, bis die Leitung zur Übertragung wieder frei ist, d. h., es kommen Wartespeicher hinzu:

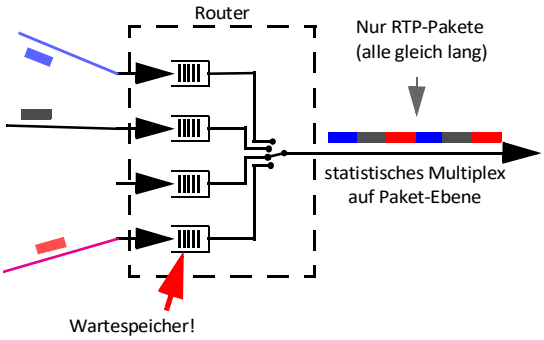


Abbildung 29: Paketmultiplex

5.2 Mischungen zwischen VoIP und der Datenkommunikation

Die Übertragung von RTP-Paketen in einer Mischung mit klassischen Datenpaketen ist etwas komplexer. Der erzeugte Verkehr der VoIP-Systeme kann immer noch nach Erlang ausgewiesen werden. Aber schon bei der Zusammenfassung muss man aufpassen: Wird über die Leistungen nur VoIP-Verkehr ohne einen Datenanteil übertragen, kann man die Verkehrswerte (die Erlang-Werte) der einzelnen Teilnehmer zusammenfassen. Bei einer Verkehrsmischung mit klassischem Datenverkehr ist das nicht möglich.

Die Belastung einer Leitung hängt von sehr vielen Faktoren ab: Welcher Codec wird verwendet, wie viele Sprachproben bilden ein RTP-Paket, wie wird QoS in dem Netz realisiert (Priorisierung, virtuelle Kanäle mit MPLS oder Überdimensionierung)? Ohne diese detaillierten Angaben kann man keine verlässlichen Aussagen zu der erzielbaren Qualität der Übertragung machen.

Woran liegt das?

Die Formeln von Erlang stellen gewisse Bedingungen, um sie anwenden zu können. Unter anderem muss der Verkehr aus einer großen Anzahl von Verkehrsquellen erzeugt werden, die voneinander unabhängig sein müssen. Des Weiteren muss die Wahrscheinlichkeit für das Auftreten einer neuen Belegung konstant sein, und die Wahrscheinlichkeit für das Auslösen einer Verbindung muss ebenso konstant sein.

Unter diesen Voraussetzungen zeigt der beobachtete Verkehr klare Mittelwerte für die mittlere Paketankunftsrate und die mittlere Belegungsdauer. Diese Mittelwerte verfügen über eine begrenzte Varianz. Für den Datenverkehr, wie er beim Surfen im Internet vorkommt, gilt das nicht, weil die Verkehrsaktivitäten nicht voneinander unabhängig sind. Ein „Klick“ auf eine komplexe Homepage verursacht eine Anzahl von Verbindungen (nicht nur eine).

Formell darf man also bei einer Mischung von klassischem Datenverkehr mit VoIP-Verkehr nicht mit den Erlang-Formeln rechnen. Die Voraussetzungen zur Anwendung der Formeln sind nicht gegeben. Dies gilt auch für priorisierten VoIP-Verkehr.

VoIP und Datenkommunikation gemeinsam übertragen

Werden mit einer Leitung sowohl Datenpakete als auch RTP-Pakete übertragen, können weniger VoIP-Kanäle realisiert werden. Ist beispielsweise eine 2,048 Mbit/s-Verbindungsleitung im Durchschnitt mit einem Datenverkehr von 15 % belastet, verbleiben 2,048 Mbit/s – 307,2 kbit/s = 1,7408 Mbit/s für Sprache. Mit ca. 92 kbit/s macht das dann 18 Sprachkanäle -> diese vereinfachte Rechnung ist falsch!

Der Datenverkehr wird den Durchschnittswert nur bei einer sehr langen Beobachtungsdauer erreichen. Durch die bei diesem Verkehr häufig vorkommenden Spitzen oder Bursts ist die Augenblicksbelastung deutlich höher als die mittlere Belastung.

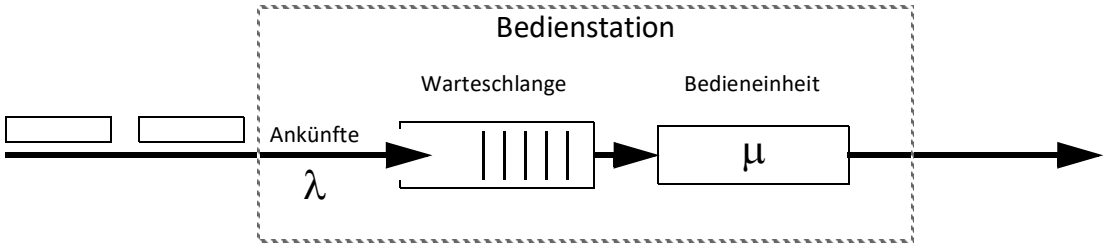


Abbildung 30: Modell eines Wartezeitsystems

Zudem erfolgt die Verkehrsmischung auf der Ebene der IP-Pakete. Für die Bewertung der Belastung einer Leitung sind zwei Werte sehr wichtig. Zum einen die Anzahl der Pakete pro Sekunde, die über diese Leitung transportiert werden sollen (die sog. Ankunftsrate), und zum anderen die Anzahl der Pakete, die von dieser Leitung transportiert werden können (die Bedienrate; dieser Wert hängt von der Paketgröße und der Übertragungsgeschwindigkeit der Leitung ab).

5.3 Wartezeitsystem

Die Übertragung von IP-Paketen kann durch ein sog. „Wartezeitsystem" modelliert werden. In Wartezeitsystemen wird die Systemauslastung als Verhältnis zwischen der mittleren Ankunftsrate und der mittleren Bedienrate definiert. Diese kann man aber nicht zu 100 % auslasten, weil sowohl die Ankunftsrate als auch die Bedienrate jeweils Durchschnittswerte sind, die vom Datenverkehr nur sehr schlecht erreicht werden, bzw. bei einer sehr langen Beobachtungsdauer erreicht werden.

Mathematisch können solche Wartezeitsysteme beschrieben werden. Der „Auslastungsfaktor" (ρ) ist dabei das Verhältnis von Ankunftsrate (λ) zu Bedienrate (μ):

mittlere Auslastung (Durchsatz)

$$\rho = \frac{\lambda}{\mu} = \frac{\text{mittlere Ankunftsrate (Last)}}{\text{mittlere Bedienrate (Leistung des Systems)}}$$

Die Wartezeit, bis ein Paket auf der Leitung übertragen wird, hängt von der augenblicklichen Auslastung, dem ρ, ab. Mit zunehmender Auslastung nimmt die Wartezeit durch eine kleine Auslastungserhöhung überproportional zu [Kle75].

Die Berechnung der Wartezeit hängt von sehr vielen Faktoren ab:

- der augenblicklichen Systemauslastung (nicht der mittleren Systemauslastung),
- der Größe der Pakete (sowohl die Größe der RTP-Pakete als auch die der IP-Datenpakete),
- der Art, wie QoS in dem System realisiert wird (ein System mit Priorisierung, z. B. DiffServ, verhält sich anders als ein System mit VLAN).

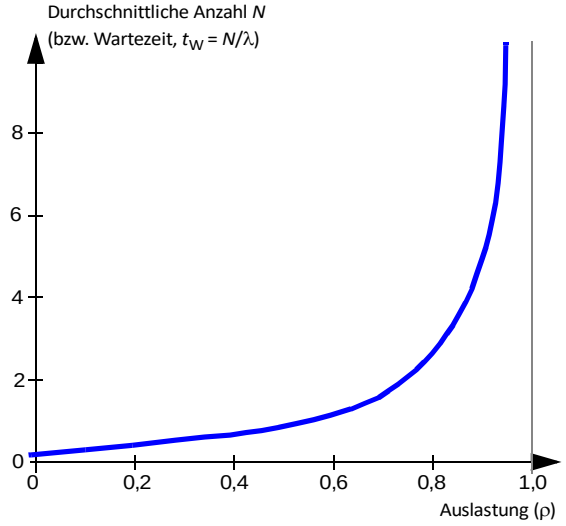


Abbildung 31: Die Wartezeit steigt mit der Systemauslastung

Noch einmal das TDM-System

Die 2,048-Mbit/s-Strecke ist das Primärmultiplexsystem der TDM-Technik. Von den 32 Kanälen mit jeweils 64 kbit/s können 30 Nutzkanäle verwendet werden. Solange ein Kanal frei ist, wird dieser auf Anfrage einer bestimmten Kommunikation zugeordnet. Wenn alle Kanäle vergeben sind (gassenbesetzt), erhalten die weiteren Anfragen das Besetztzeichen. Die Anfragen gehen zu Verlust, das Ganze wird auch als Verlustsystem bezeichnet.

Die Übertragung ist kanalorientiert, d. h., jeder Verbindung ist eine bestimmte Kanalnummer zugeordnet. Das System transportiert mit dem Codec G.711 alle 125 µs 8 Bit vom Sender, dies entspricht genau der Übertragungsrate des Codecs, 64 kbit/s.

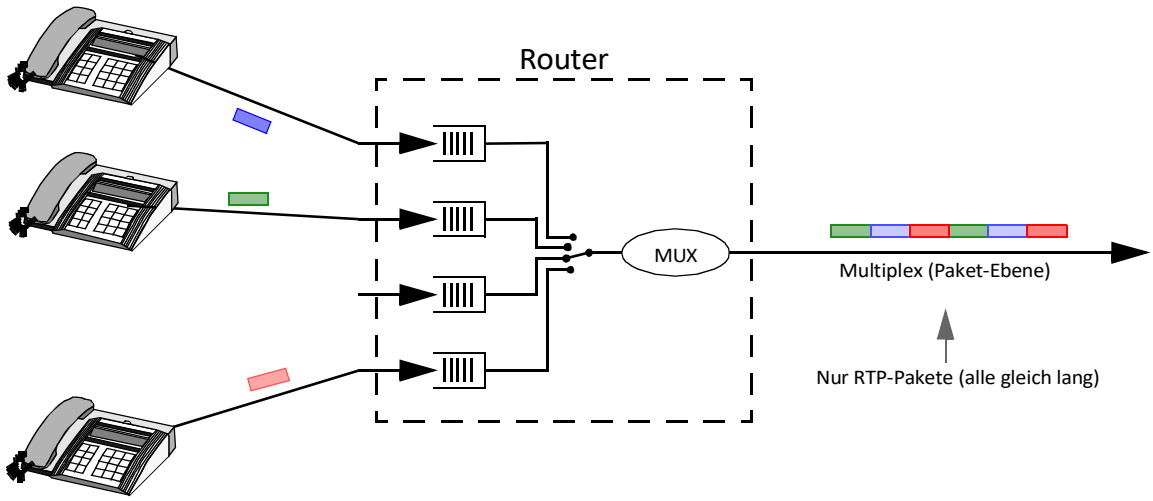


Abbildung 32: Nur Pakete mit Sprachproben (RTP)

Für VoIP-Systeme werden im Folgenden verschiedene Ansätze genauer betrachtet. Im ersten Ansatz werden nur RTP-Pakete und keine anderen Datenpakete übertragen. Im zweiten Ansatz werden die Echtzeitdaten der RTP-Pakete gegenüber den Datenpaketen bevorzugt übertragen (Priorisierung). Im dritten Ansatz wird einfach davon ausgegangen, dass eine 2,048-Mbit/s-Strecke nur gering mit Datenpaketen belastet ist – der Ansatz der Überdimensionierung.

Für alle drei Ansätze wird jeweils betrachtet, wie viele VoIP-Kanäle übertragen werden können und welche durchschnittliche Verzögerungszeit bei der Übertragung entsteht.

5.4 Die Strecke transportiert nur RTP-Pakete

5.4.1 Exklusive Nutzung

Die 2,048-Mbit/s-Strecke wird im ersten VoIP-Beispiel für die Übertragung von RTP-Paketen verwendet. Hier werden nur RTP-Pakete übertragen, dies entspricht einer exklusiven Nutzung, wie sie für eigene Leitungen oder die Verwendung von VLAN bzw. mit einem MPLS-System zur Verfügung gestellt wird.

In diesem Beispiel werden die Nutzdaten (64 kbit/s) vom Codec noch zusätzlich mit den sog. Overhead-Informationen der Transportprotokolle RTP, UDP, IP und der verwendeten Schicht 2 (z. B. Ethernet, VLAN) versehen. Wie oben bereits betrachtet, erhöht sich die erforderliche Übertragungsrate für einen Sprachkanal dadurch auf ca. 92 kbit/s.

Die Rechnung:

$$N = \frac{2{,}048\ \text{Mbit/s}}{92\ \text{kbit/s}} = 22{,}3\ \text{Kanäle}$$

Teilt man die 2,048 Mbit/s durch die 92 kbit/s des einzelnen VoIP-Kanals, kann man die maximale Anzahl von VoIP-Kanälen berechnen. N = 22 Kanäle je Übertragungsrichtung ist die maximale Last, die eine 2,048-Mbit/s-Strecke übertragen könnte. Diese Annahme gilt auch nur, wenn die RTP-Pakete alle 20 ms übertragen werden.

Wie immer bei der RTP-Übertragung werden erst ein paar Sprachproben gesammelt, bevor ein RTP-Paket mit UDP und IP transportiert wird. Wie viele Sprachproben je RTP-Paket übertragen werden, kann in den Kommunikationssystemen konfiguriert werden (z. B. 10 ms, 20 ms oder 30 ms). Diese Zeit sollte nicht zu groß gewählt werden, da die erste Sprachprobe diese Zeit warten muss, bis das RTP-Paket abgesendet wird. Auf der anderen Seite erhöhen sehr kleine RTP-Pakete die Paketrate im System (Ankunftsrate) und erhöhen damit die Systemauslastung insgesamt. Die RTP-Pakete sind dann aber alle gleich groß. (Wichtig für die Bedienrate: Wie viele RTP-Pakete können von der Leitung übertragen werden?).

Da die Gespräche völlig unabhängig voneinander beginnen, müssen unter Umständen eine Anzahl von RTP-Paketen gleichzeitig übertragen werden – dies bedeutet Wartezeit. Da hier die Mischung der verschiedenen Kommunikationen nicht kanalbezogen ist (alle 8 Bit wird der Kanal gewechselt), sondern erst eine Anzahl von Sprachproben gesammelt wird, bis ein RTP-Paket erzeugt wird, können deutlich größere Wartezeiten entstehen. Da (theoretisch) alle Verbindungen zur selben Zeit begonnen wurden, müssen die RTP-Pakete vor dem Transport warten (Paketmultiplex statt Bitmultiplex). Die RTP-Paketgröße ist allerdings immer gleich groß, daher muss ein RTP-Paket ein, zwei, drei usw. andere RTP-Pakete abwarten, bis es transportiert werden kann.

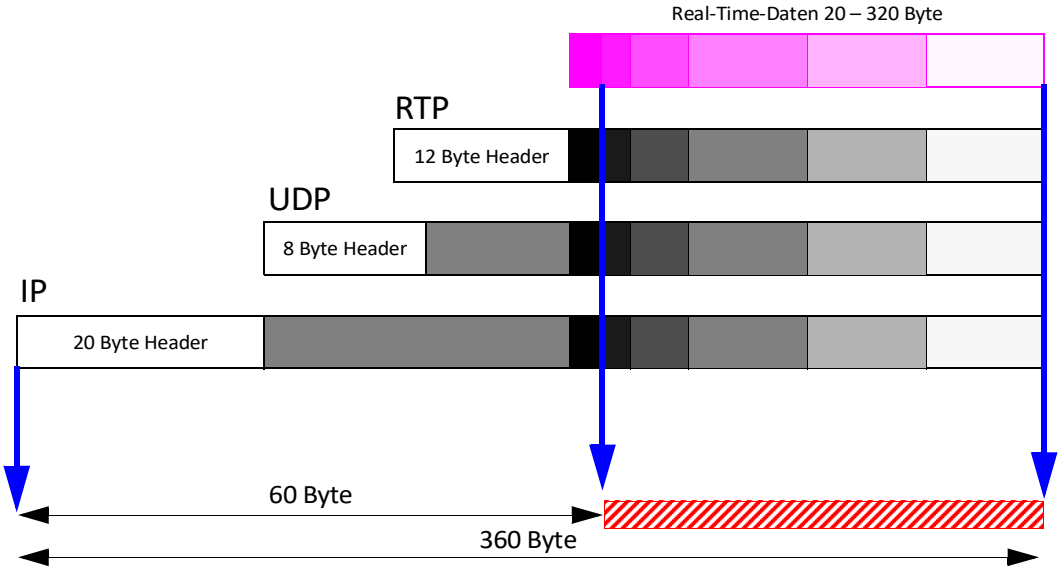


Abbildung 33: Nutzinformationen und Overheads

5.4.2 Berechnung der mittleren Wartezeit für ein RTP-Paket

Die Paketgröße ist für alle VoIP-Kommunikationen immer gleich. Ein RTP-Paket besteht aus den Real-Time-Daten (den digitalisierten Sprachinformationen) plus den Overheads von RTP, UDP, IP und den unterliegenden Schichten.

Die Größe der Sprachinformationen hängt auch vom verwendeten Codec ab. Mit dem Codec G.711 (der Standard-ISDN-Codec) werden 80 Byte je RTP-Paket alle 10 ms übertragen. Wird alle 20 ms ein RTP-Paket erzeugt, werden 160 Byte je RTP-Paket übertragen. Bei 30 ms Paketabstand sind es 240 Byte und bei 40 ms 320 Byte je RTP-Paket. Kleinere und größere RTP-Pakete sind für diese Betrachtungen nicht sinnvoll. Zu diesen Größen der RTP-Nutzlast kommen jeweils noch die Overheads für die verschiedenen unterliegenden Protokolle hinzu.

Die Wartezeit für ein RTP-Paket hängt damit von der gewählten RTP-Größe, der Übertragungsgeschwindigkeit und von der Anzahl der laufenden Verbindungen ab. Werden die RTP-Pakete alle 20 ms übertragen, ist die Zeit für den Transport eines RTP-Pakets 230 · 8 bit/2,048 Mbit/s = 0,898 ms. Für zwei RTP-Pakete ist die Zeit 2 · 0,898 ms = 1,796 ms, für drei RTP-Pakete ist die Zeit 3 · 0,898 ms = 2,694 ms usw. Die maximale Wartezeit wird erreicht, wenn zufällig alle laufenden Verbindungen genau gleich begonnen wurden. Mit dem Start der Verbindung werden die RTP-Pakete in einem festen Raster erzeugt. Die mittlere Wartezeit bei N gleichzeitigen Verbindungen liegt dann zwischen 0 und N mal die jeweilige Paketlaufzeit. Mehr dazu und zu den weiteren Berechnungen ist im Anhang zu finden.

Dieser Ansatz wird auch für die logische Trennung mit virtuellen Kanälen wie MPLS- und VLAN-Systemen verwendet, auch wenn es für diese ein idealisierter Ansatz ist. In der Realität sind die Verhältnisse etwas komplexer. Die vereinfachte Übernahme der Berechnungen mit exklusiven Leitungen auf die MPLS- bzw. VLAN-Verhältnisse ist für diese Betrachtungen aber ausreichend genau, da ja nur die grundsätzlichen Verhaltensweisen verdeutlicht werden sollen – exakte Berechnungen in diesen Systemen sind komplizierter.

5.4.3 Die Ergebnisse

In Abhängigkeit von der RTP-Paketgröße ergeben sich damit die folgenden maximalen Übertragungskapazitäten:

- Wird alle 10 ms ein RTP-Paket erzeugt, können 17 Sprachkanäle übertragen werden. Setzt man diese 17 Kanäle dem N in der Erlang'schen Verlustformel gleich, entsprechen die 17 Sprachkanäle, bei einem Verlust von B < 0,1 %, 9,65 Erl. Die mittlere Wartezeit t_W für ein RTP-Paket beträgt im Durchschnitt 2,3 ms.

- Wenn alle 20 ms ein RTP-Paket erzeugt wird, können 22 Sprachkanäle oder 13,7 Erl mit einer mittleren Wartedauer von 4,7 ms, unterstützt werden.

- Beträgt der Abstand zwischen zwei RTP-Paketen 40 ms, können sogar 26 Sprachkanäle oder 17,0 Erl, mit einem t_w von 9,5 ms, unterstützt werden.

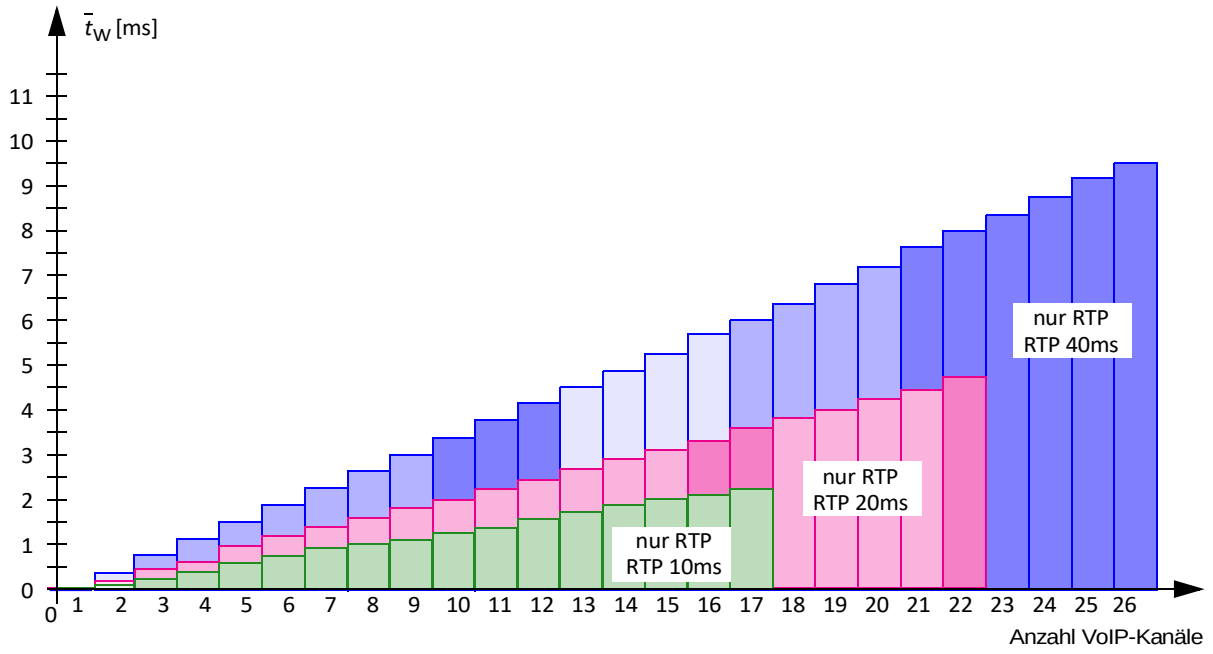


Abbildung 34: Delay in Abhängigkeit von der Auslastung (nur Sprache)

Je größer die RTP-Pakete gewählt werden, umso mehr Sprachkanäle können mit der gleichen Übertragungskapazität realisiert werden. Automatisch steigt aber auch die Gesamtlaufzeit, weil die erste Sprachprobe beispielsweise fast 40 ms (genau 40 ms – 125 µs = 39,875 ms) warten muss, bis sie mit RTP transportiert wird. Je größer das RTP-Paket gewählt wird, umso weniger Overhead-Informationen durch die zusätzlichen Protokolle (RTP, UDP, IP und Ethernet) müssen anteilig zur Nutzinformation übertragen werden. Kleinere Pakete erhöhen zudem die Ankunftsrate des Systems, damit steigt automatisch die Belastung – insgesamt können weniger Verbindungen realisiert werden.

Die Werte gelten aber nur, wenn ausschließlich RTP-Pakete über die Strecke transportiert werden (keine Datenpakete, auch keine Priorisierung wie DiffServ)! Dies ist (näherungsweise) auch der Fall für durchgängig (auch in der Anschaltung an den externen Netzbetreiber) getrennte und überwachte virtuelle Kanäle, wie sie bei MPLS, VLAN oder den Software-defined Networking (SDN) mit QoS verwendet werden.

5.5 RTP wird priorisiert übertragen

5.5.1 Der Ansatz

In vielen Netzen werden neben den Echtzeitinformationen auch Datenpakte übertragen. Die folgenden Betrachtungen nehmen eine Bevorrechtigung der RTP-Pakete mit DiffServ an. Auf der Übertragungsstrecke werden dann sowohl RTP-Pakete mit einer festen Paketlänge als auch Datenpakete mit einer variablen Paketlänge übertragen. Wie bei einem DiffServ-Router wird bei der folgenden Betrachtung angenommen, dass ein eigener Wartespeicher für die bevorrechtigten Pakete besteht. Für die weiteren Betrachtungen wird davon ausgegangen, dass der DiffServ-Router nach der vollständigen Bearbeitung eines Pakets immer in die bevorrechtigte Warteschlange zuerst wieder adressiert wird (Strict Priority Routing). Um einen Vergleich zu dem exklusiven RTP-Transport zu erhalten, wird auch hier die Anzahl der möglichen VoIP-Kanäle berechnet, die unter der Annahme einer Belastung von 15 % erzielt werden können.

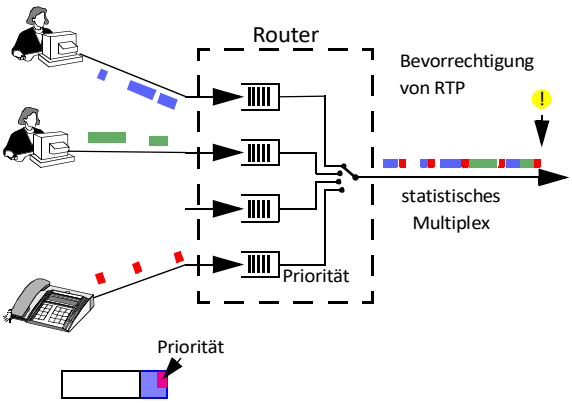


Abbildung 35: Gemischte Übertragung von Sprache und Daten

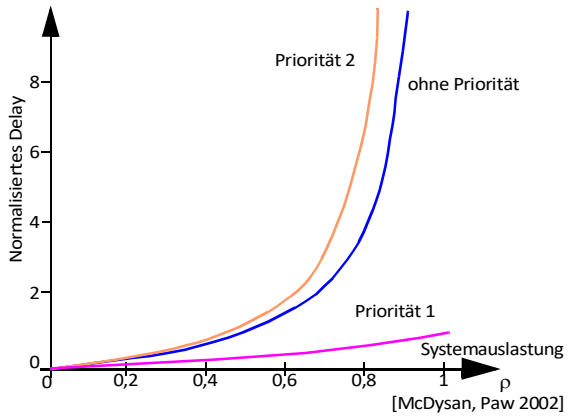


Abbildung 36: Geringes Delay durch Priorisierung

Das Verfahren DiffServ geht allerdings von einem geringen Anteil von vorberechtigten Paketen aus (ca. 5 – 15 %). Dieses Grundprinzip wird in dieser Betrachtung nicht eingehalten. Pakete mit höherer Priorität werden gegenüber anderen Paketen bevorzugt, das bedeutet aber auch, dass die Datenpakete mit geringerer Priorität benachteiligt werden. Da die Warteschlange mit der höheren Priorität bevorzugt geleert wird, müssen die Pakete in den anderen Warteschlangen länger als ohne eigene Warteschlange für die bevorrechtigten Pakete warten [Tra05], [Her05].

5.5.2 Ergebnisse

Für die Berechnung wird vereinfacht angenommen, dass die Ankunftsrate der Pakete der RTP-Rate entspricht (weil hierfür ein eigener Wartespeicher vorhanden ist). Für die Bedienrate ist die durchschnittliche Paketlänge wichtig, diese wird durch die transportierten Datenpakete gegenüber der ersten Betrachtung verändert. Die Datenpakete sind deutlich größer und werden, mit geringerer Priorität, auch auf der Strecke übertragen. Die Zeit, bis die Leitung dann wieder frei ist, um RTP-Pakete zu transportieren, hängt von der Paketgröße der Datenpakete ab und ist ggf. immer unterschiedlich. In der Berechnung wurde die mittlere Paketlänge anteilig ermittelt (5 % bzw. 15 % Datenpaket mit bis zu 1518 Byte je Paket und der Rest mit der jeweiligen RTP-Paketgröße, einschließlich der Overheads).

Für den bevorrechtigten Transport bei einer Last von 15 % für die klassische Datenkommunikation ergeben sich die folgenden Werte:

- Alle 10 ms ein RTP-Paket: 6 Sprachkanäle und eine mittlere Wartezeit von 8,14 ms
- Alle 20 ms ein RTP-Paket: 9 Sprachkanäle mit t_w = 13,9 ms
- Alle 40 ms ein RTP-Paket: 13 Sprachkanäle mit t_w = 21,6 ms

Die Auslastung (RTP alle 20 ms) ist in der folgenden Darstellung zusammengefasst (s. Abb.: 37)

Bei den Berechnungen wurde das DiffServ-Verfahren nach dem Strict-Priority-Mechanismus angenommen. Bei diesem Verfahren wird nach der Bearbeitung eines Pakets wieder bei der Warteschlange mit der höchsten Priorität begonnen. Bei anderen Verfahren wie dem Weighted Fair Queueing sind die Ergebnisse noch schlechter, weil, obwohl ein RTP-Paket in der Warteschlange mit der größten Priorität vorhanden ist, ggf. ein anderes Paket einer geringeren Priorität transportiert wird, d. h., das RTP-Paket muss länger warten. Damit können zwei oder mehrere IP-Pakete übertragen werden, bevor das RTP-Paket transportiert wird.

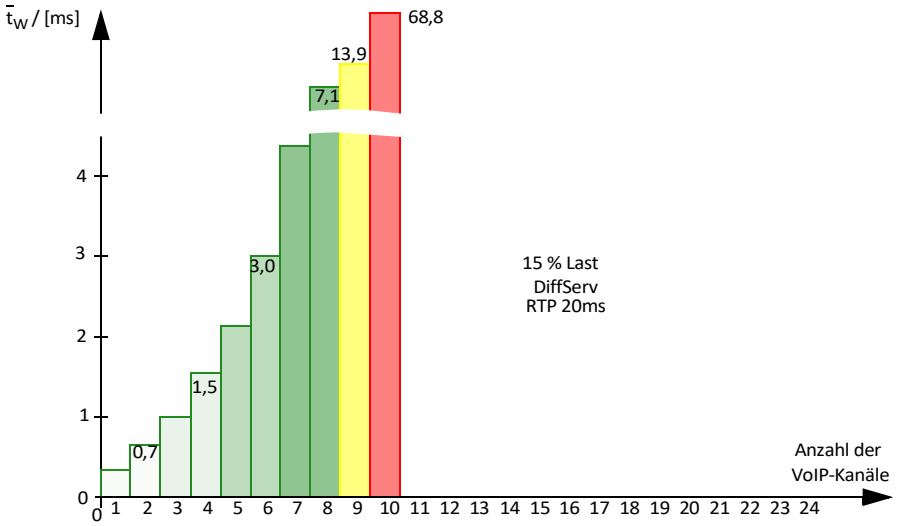


Abbildung 37: Delay der VoIP-Kanäle

5.6 QoS wird durch Überdimensionierung des Systems realisiert

5.6.1 Der Ansatz

In manchen Fällen wird angenommen, dass eine Belastung von 15 % durch vorhandenen Datenverkehr die Übertragung von Sprachinformationen mit RTP kaum beeinflusst. Das ist nicht der Fall. Eine einfache Überdimensionierung ist keine QoS-Maßnahme.

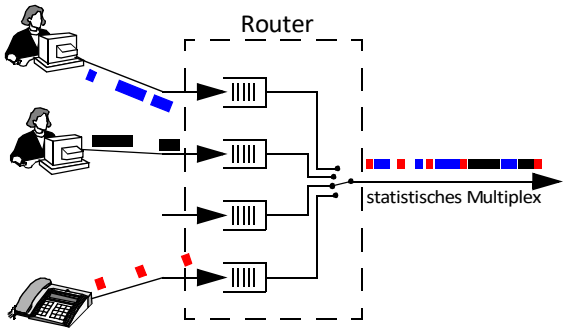


Abbildung 38: Statistisches Multiplex

Als Grundlage für die folgende Berechnung wird von einer Verkehrsmischung von einer 15%-prozentigen Belastung der 2,048-Mbit/s-Strecke mit Datenverkehr und RTP-Paketen für den Transport der Sprachinformationen ausgegangen. Die Pakete werden in der Reihenfolge ihres Eintreffens bearbeitet.
Nach den ersten überschlägigen Überlegungen sollte eine 2,048-Mbit/s-Übertragungsstrecke, die zu 15 % mit Datenverkehr ausgelastet ist, nur mit 307,2 kbit/s belastet sein. Der Rest könnte dann RTP-Pakete übertragen, das sind dann 2,048 Mbit/s – 307,2 kbit/s = 1,7408 Mbit/s für Sprache. Mit der Annahme oben, dass jeder VoIP-Kanal 92 kbit/s benötigt, macht das dann 18 Sprachkanäle -> dieser Ansatz ist falsch!

Die viel größeren Datenpakete (typisch: 1500 Byte) mischen sich mit den relativ kleinen RTP-Paketen (bei 20 ms Raster sind RTP-Pakete 230 Byte groß) der Echtzeitübertragung. Die Datenpakete kommen häufig in Paketgruppen (Burst-artig) mit längeren Ruhephasen dazwischen. Da die RTP-Pakete hier nicht bevorzugt übertragen werden, können sich zwischen den RTP-Paketen immer wieder ein oder mehrere große Datenpakete mischen.

Diese Systeme gehen viel früher in die Auslastungsphase, d. h., bereits bei geringen Belastungen kommt es hier zu langen Wartezeiten für den Transport der Pakete. Dies ist eine Eigenschaft des Internetverkehrs, völlig unabhängig von der Übertragung von Echtzeitdaten. Für die Sprachübertragung ist dieses Verhalten sehr problematisch, solche Strecken weisen auch bei relativ geringen Auslastungen große Verzögerungen auf.

Die Eigenschaften des Datenverkehrs sind völlig andere als die der Echtzeitkommunikation. Diese Systeme weisen einen selbstähnlichen Verkehr auf (sie sind nicht mehr gedächtnislos). Die Ankünfte der Pakete (oder Paket-Bursts) sind nicht mehr unabhängig voneinander. Man spricht in diesem Zusammenhang von einer Long-Range-Dependence (LRD). In diesen Systemen kann man kaum einen Mittelwert bestimmen, die Varianz -> ∞.

Wichtig: In Netzen mit einem gemischten Betrieb von VoIP und klassischer IP-Datenkommunikation (wie WWW) gelten der Erlang-Ansatz und seine Formeln nicht!

Der Erlang-Ansatz gilt nicht, weil die von der Datenkommunikation erzeugten Verbindungen nicht unabhängig voneinander sind (ein „Klick“ auf eine Seite erzeugt eine Anzahl von abhängigen Verbindungen)!

Die Verkehrskurven der klassischen Datenkommunikation weisen sehr starke Nutzungsunterschiede auf, die Kurven sind sehr „spitzig“, haben viele sehr steile Peaks und ändern sich rasch von intensiver zu einer geringen Nutzung. In der Aktivitätsphase wird oft gleich eine Anzahl von Paketen zwischen beiden Endpunkten der Kommunikation übertragen. Diese zeitlich begrenzten Schübe der Aktivität werden als „Bursts“ bezeichnet. Dieses Burst-artigen Verhalten findet man sowohl in Verkehrskurven von sehr kleinen Zeitabschnitten als auch in den Kurven von sehr langen Abschnitten. Man sagt, der Verkehr besitzt eine Selbstähnlichkeit. Ein Maß für die Selbstähnlichkeit ist der sog. Hurst-Parameter H: 0,5 < H < 1 (0,5 kaum selbstähnlich, 1 sehr große Selbstähnlichkeit, typische Werte liegen zwischen 0,75 und 0,9) [Gri04].

5.6.2 Ergebnisse bei 15 % Last

In manchen Fällen wird vor der VoIP-Installation eine Verkehrsmessung durchgeführt. In den meisten Fällen wird man feststellen, dass die Datennetze relativ gering belastet sind. Aus dieser Tatsache wird manchmal geschlossen, dass die geringe Belastung durch die VoIP-Kommunikation ohne Probleme vom Netz getragen werden kann. Das ist falsch: Wird die Übertragungsstrecke mit nur 15 % der Kapazität im Durchschnitt durch Datenverkehr belastet, können nur noch wenige VoIP-Kanäle mit sehr großen Verzögerungszeiten unterstützt werden:

- Alle 10 ms ein RTP-Paket: 3 Sprachkanäle mit einer mittleren Wartezeit von
  t_W = 16,7 ms

- Alle 20 ms ein RTP-Paket: 5 Sprachkanäle mit einer mittleren Wartezeit von
  t_W = 19,2 ms

- Alle 40 ms ein RTP-Paket: 8 Sprachkanäle mit einer mittleren Wartezeit von
  t_W = 33,7 ms

Statt der oben errechneten 18 Sprachkanäle sind es gerade einmal 5 mögliche VoIP-Kanäle mit akzeptablen Verzögerungszeiten.

Die (wie auch immer) gemessenen 15 % durchschnittliche Belastung durch Datenkommunikation ist keine kontinuierliche Belastung. IP-Pakete treten häufig Burst-artig in Paketschüben auf. Die kurzen RTP-Pakete müssen dann sehr lange warten, bis sie transportiert werden können.

Durch eine Anpassung der Paketlänge (Maximum Transmission Unit – MTU) kann man die Auswirkungen auf die RTP-Pakete beeinflussen. Die MTU bezeichnet die Summe aller Byte, die mittels IP (inklusive aller Protokollanteile der höheren Schichten) in einem Schicht-2-Block übertragen können. Für Ethernet sind dies im Allgemeinen 1500 Byte für IP und alle höheren Protokolle plus 18 Byte für die Schicht 2, zusammen sind dies 1518 Byte. Die MTU-Size ist in vielen Systemen konfigurierbar. Eine kleinere MTU-Size ist für eine Verkehrsmischung mit RTP-Paketen vorteilhaft, weil dann keine langen, durchgängigen Belegungszeiten für den Transport großer IP-Pakete auftreten – oder anders ausgedrückt: Die RTP-Pakete können sich immer wieder zwischen zwei kleinere IP-Pakete schieben. Größere IP-Pakete (Jumbo Frames, bis 9000 Byte), wie sie in Gigabit-Ethernet verwendet werden, bewirken genau das Gegenteil, d. h., die Verzögerungszeiten nehmen stark zu. Das gleiche trifft auch auf die Laufzeitschwankungen, den Jitter, zu: Je größer die MTU, umso größer werden die Laufzeitschwankungen.

Auf der anderen Seite bringen VoIP-Anwendungen eine sehr große zusätzliche Belastung in das vorhandene Netz. Die relativ kleinen, aber sehr häufig übertragenen RTP-Pakete erhöhen die Paketlast sehr stark. Als Folge davon treten zusätzliche Verzögerungen für alle Pakete im Netz auf. Die zusätzlichen Verzögerungen führen zu Paketverlust, weil die Pakete für die VoIP-Anwendungen zu spät das Ziel erreichen.

5.6.3 Ergebnisse bei 5 % Last

Auch bei einer Reduzierung der durchschnittlichen Datenbelastung auf nur 5 % der Kapazität der Übertragungsstrecke sind die erzielten Ergebnisse (s. Bild) ernüchternd:

- Alle 10 ms ein RTP-Paket: 7 Sprachkanäle mit einer mittleren Wartezeit von
  t_W = 28,6 ms.

- Alle 20 ms ein RTP-Paket: 10 Sprachkanäle mit einer mittleren Wartezeit von
  t_W = 28,7 ms.

Alle 40 ms ein RTP-Paket: 12 Sprachkanäle mit einer mittleren Wartezeit von
t_W = 22,8 ms.

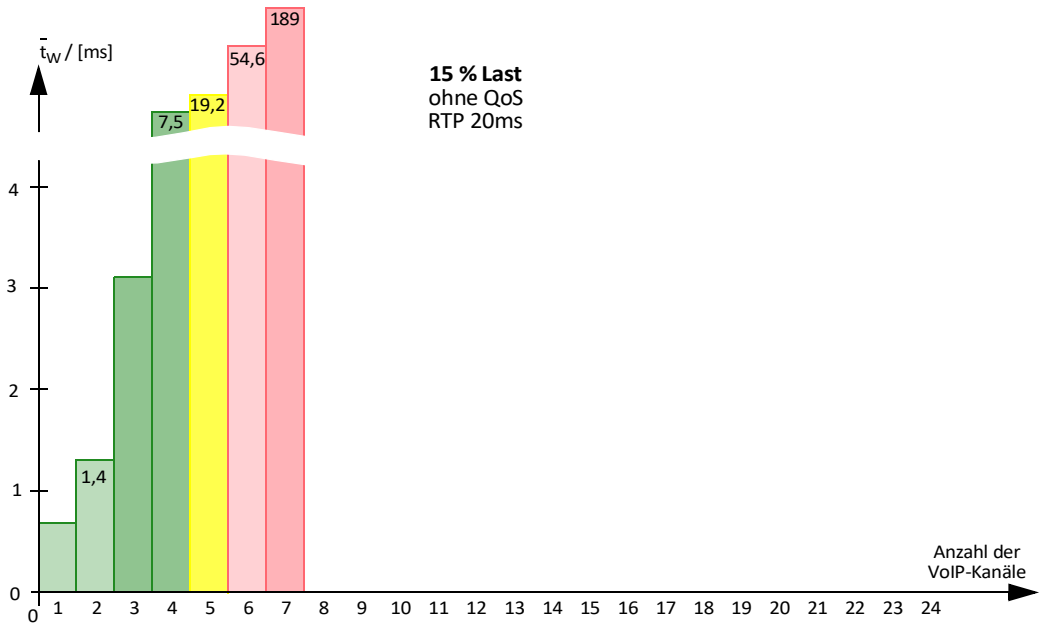


Abbildung 39: Wartezeiten für 15 % Datenlast

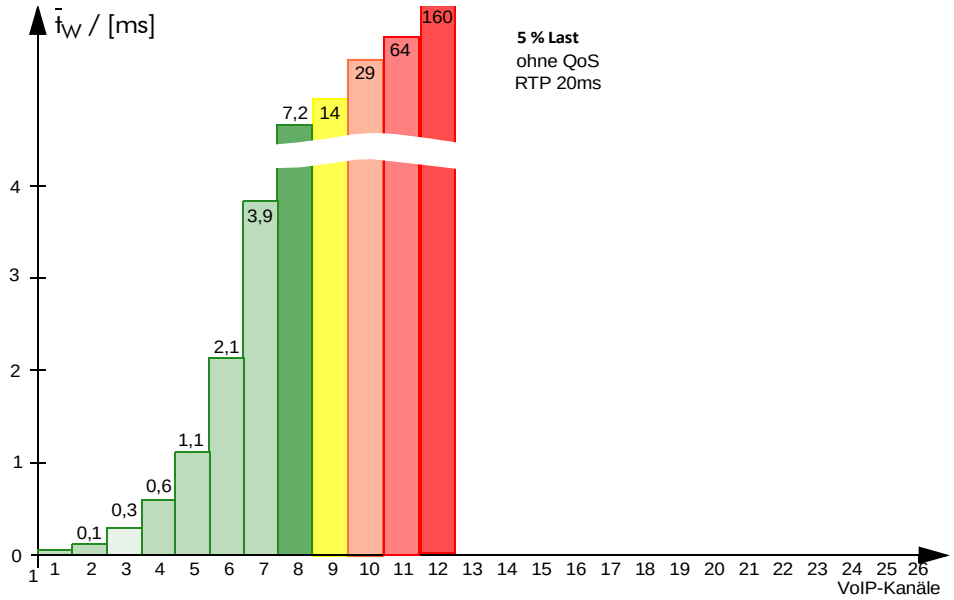


Abbildung 40: Wartezeiten für 5 % Datenlast

Alle errechneten Werte sind Durchschnittswerte, die in realen Netzen um ein Vielfaches überschritten werden können. Die Maßnahme der Überdimensionierung ist als Basis für eine Übertragung von Echtzeitinformationen, wie Sprache mit RTP-Paketen, ungeeignet.

5.6.4 Ein neuer Ansatz mit VLAN und 5 % Last

Werden für die Übertragungsstrecke zwei virtuelle Kanäle (mit VLAN oder MPLS) eingerichtet und die Daten- (5 % der Gesamtkapazität) von der Echtzeitkommunikation so getrennt, können deutlich mehr VoIP-Kanäle mit viel geringeren Verzögerungszeiten übertragen werden.

- Wird alle 10 ms ein RTP-Paket erzeugt, können 16 Sprachkanäle übertragen werden mit einer mittleren Wartezeit $\bar{t}_w$ = 2,3 ms für jedes RTP-Paket.
- Wenn alle 20 ms ein RTP-Paket erzeugt wird, können 21 Sprachkanäle mit einer mittleren Wartezeit von 4,7 ms übertragen werden.
- Beträgt der Abstand zwischen zwei RTP-Paketen 40 ms, können sogar 24 Sprachkanäle mit einem $\bar{t}_w$ von 9,2 ms unterstützt werden.

Das folgende Bild zeigt die Auslastung des virtuellen Kanals für die Echtzeitkommunikation, wenn für 5 % der Übermittlungsrate ein eigener virtueller Kanal für die Datenkommunikation angelegt wurde (hier für RTP alle 20 ms):

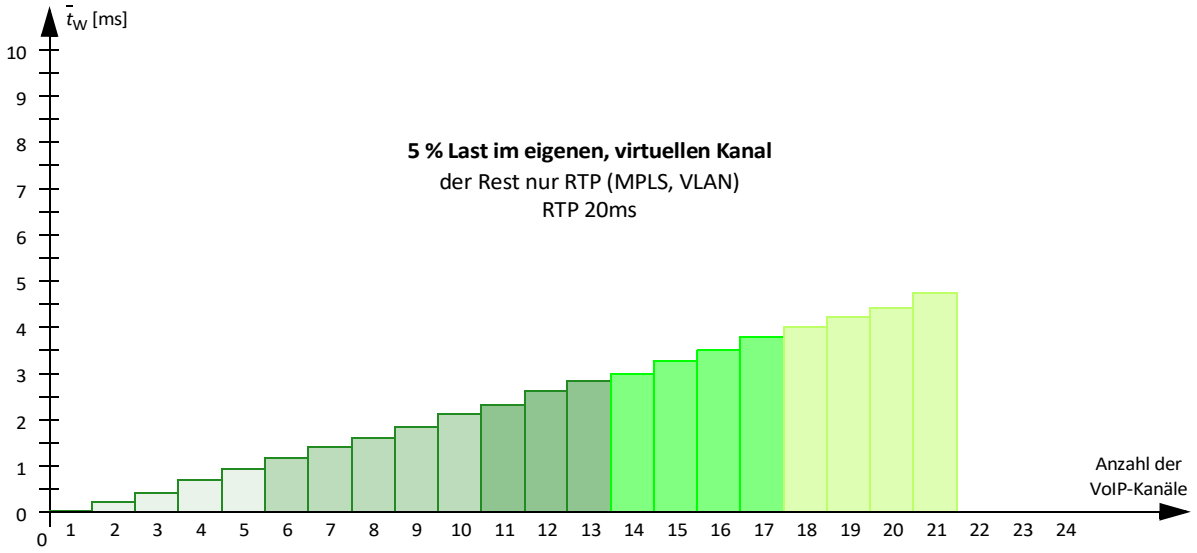


Abbildung 41: Zeiten für 5 % Datenlast mit VLAN

Die Trennung der beiden Verkehrsarten durch unterschiedliche virtuelle Kanäle ermöglicht die 3- bis 4-fache Übertragungskapazität für die Echtzeitkommunikation. Die Eigenschaften der Datenkommunikation werden dabei nicht wesentlich beschnitten. Neben der Zuordnung zu einem virtuellen Kanal ist die Überwachung und Regulierung der verwendeten Übermittlungsraten ein wichtiges Element, um dies zu ermöglichen.

5.7 Geschwindigkeitswechsel

Ein kritischer Punkt für die Leistungsfähigkeit eines Systems und die Entstehung größerer Verzögerungszeiten sind Stoßstellen, an denen Netze und Leitungen mit unterschiedlichen Geschwindigkeiten aufeinanderstoßen. Dies ist beispielsweise bei einer Kopplung zweier Standorte über eine gemietete Festverbindung der Fall. Während die Pakete aus dem lokalen Netz mit einer hohen Geschwindigkeit (Dauer der Paketübertragung mit der lokalen Geschwindigkeit: t_1) und großer Paketrate transportiert werden, bietet die gemietete Festverbindung meist nur eine deutlich geringere Geschwindigkeit. Die Pakete werden über diese Schnittstelle deutlich langsamer transportiert (t_2) und es ist nur eine deutlich geringere Paketrate erzielbar. An der Stoßstelle, wo beide Systeme zusammentreffen, müssen Zwischenspeicher (Buffer) vorgesehen werden, die Pakete in schneller Folge vom lokalen Netz aufnehmen und sie mit einer geringeren Geschwindigkeit über die Mietleitung transportieren. Dieser Vorgang führt für den Pakettransport zu einer Verzögerungszeit, die in Abhängigkeit von der Verkehrsdichte stark schwanken kann (Jitter). Da die Zwischenspeicher immer auch eine begrenzte Größe haben, kann es auch zu einem zusätzlichen Paketverlust kommen.

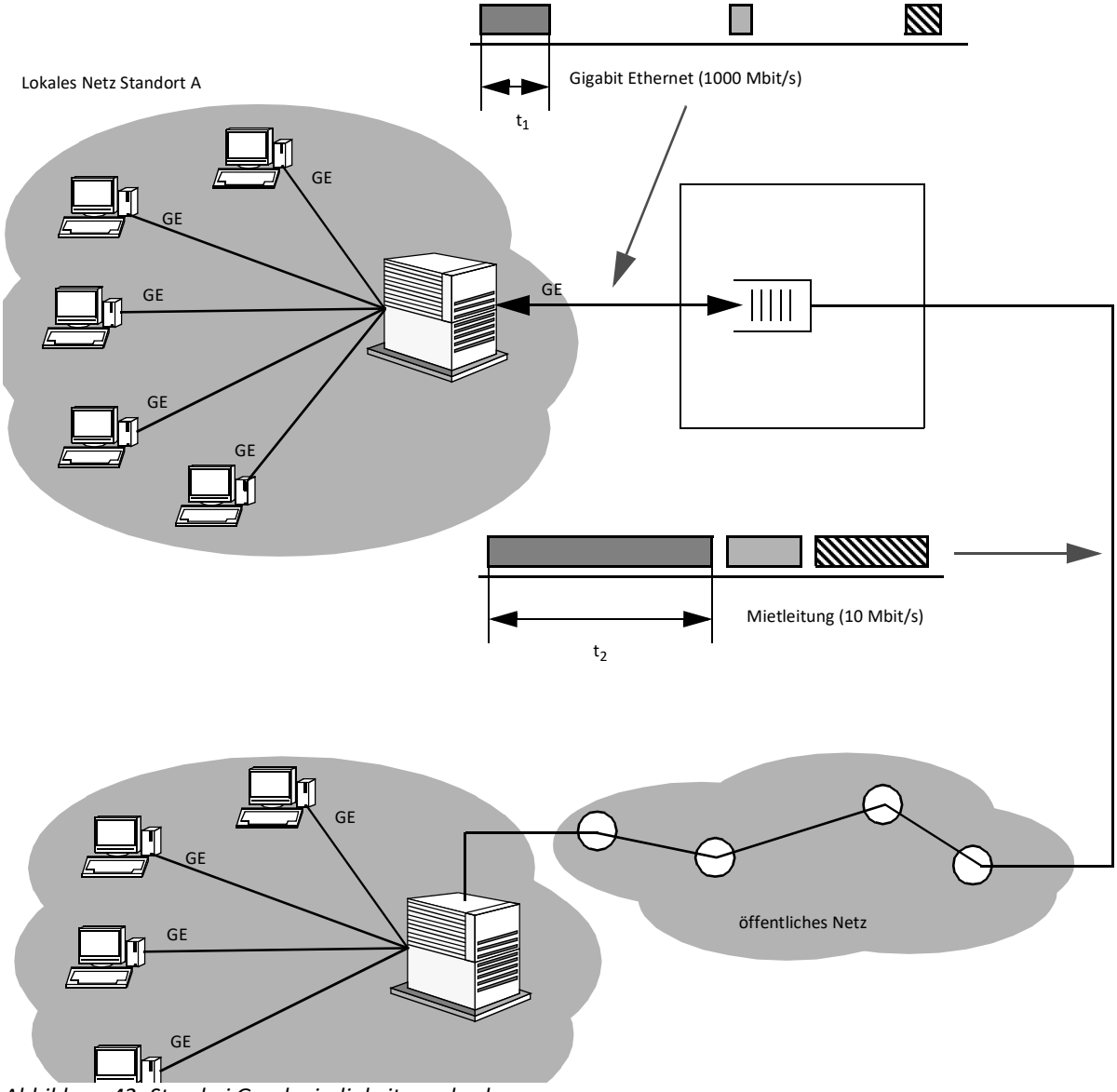


Abbildung 42: Stau bei Geschwindigkeitswechsel

5.8 Zusammenfassung

Verschiedene QoS-Maßnahmen für die Übertragung von Sprachinformationen in einem VoIP-System wurden durch die vereinfachte Berechnung verglichen. In der folgenden Tabelle sind die Ergebnisse für die verschiedenen QoS-Ansätze zusammenfassend dargestellt. Bei den Werten der Tabelle wurde in allen Fällen eine durchschnittliche Grundlast von 5 % für die klassische Datenkommunikation (H = 0,8) angenommen. Je nach gewählten QoS-Maßnahmen und gewählten RTP-Parametern können 7 bis 24 VoIP-Kanäle je 2,048-Mbit/s-Strecke unterstützt werden, ohne QoS-Maßnahmen sind es nur 7 bis 12.

Die Trennung von VoIP-Paketen (RTP-Pakete) und der klassischen Datenkommunikation ist auf jeden Fall die bevorzugte Lösung. Mit der Trennung können am meisten VoIP-Kanäle über eine gegebene Strecke übertragen werden. Die Trennung kann physikalisch durch eigene Leitungen oder logisch durch eigene virtuelle Kanäle wie bei MPLS oder VLAN mit QoS-Maßnahmen erfolgen. Diese Trennung muss dann durchgängig in der externen Anschaltung durchgehalten werden. VLAN und MPLS im eigenen Netz ist nur ein Teil, wenn dann doch wieder verschiedene Verkehrsarten mit einer Leitung an das externe Netz gegeben werden, hebt sich die saubere Trennung auf und die guten Werte für die Paketlaufzeiten und den Jitter können für die externe Kommunikation nicht erreicht werden.

Die Priorisierung der RTP-Pakete (DiffServ, Strict-Priority-Mechanismus) ist deutlich besser, als keine QoS-Maßnahmen zu ergreifen. DiffServ ist jedoch nicht mit der Trennung von Sprache und Daten vergleichbar, es können deutlich weniger VoIP-Kanäle übertragen werden.

Völlig ohne QoS-Maßnahmen werden die kurzen RTP-Pakete immer wieder durch längere IP-Pakete verzögert. Die durchschnittlichen Laufzeiten und die zu erwartenden Jitter werden deutlich größer.

6 Fazit

6.1 Was zeigen die Ergebnisse?

Gehen wir zurück zur Anfangsfrage: Wie viel Bandbreite ist zur externen Anschaltung des Systems erforderlich? Die eingangs als Beispiel angeführte Ausschreibung war hier nicht sehr präzise, allein die Aussage „Das Netz ist VoIP-Ready“ ist sehr dehnbar! Angaben zum Codec und zur RTP-Größe fehlten in der Angebotsaufforderung. Um 20 Erl zu bedienen, braucht man von 1,5 (MPLS oder VLAN) bis 5 Systeme (ohne QoS, mit RTP alle 10 ms) mit jeweils 2,048-Mbit/s. In TDM war genau eine 2,048-Mbit/s-Strecke erforderlich.

Die Aussage „Das Netz ist VoIP-Ready“ sagt nichts! Hier muss man viel genauer nachfragen, ob und wenn ja wie QoS-Maßnahmen auf welcher Basis ergriffen werden. Die Auswirkungen der verschiedenen Maßnahmen schlagen drastisch auf die Leistungsfähigkeit des Systems durch, VoIP-Kanäle mit geringen Verzögerungszeiten transportieren zu können.

Ohne QoS-Maßnahmen wird VoIP auch das vorhandene Netz beeinflussen. Die neue, sehr große Last durch die vielen, sehr kurzen RTP-Pakete führt zu einer deutlichen Erhöhung der Netzbelastung. Die Laufzeiten werden für alle Anwendungen deutlich größer werden.

Datenlast: 5 % Last

max. Anzahl der Kanäle	MPLS oder VLAN	DiffServ	ohne QoS
RTP 10ms	16	11	7
RTP 20ms	21	16	10
RTP 40ms	24	21	12

Tabelle 2: Anzahl von VoIP-Kanälen bei 5 % Datenlast

6.1.1 Bewertung der QoS-Maßnahmen

Wie können die verschiedenen Maßnahmen bewertet werden?

- Exklusive Nutzung für VoIP: Durch (durchgängig) MPLS, VLAN (hardwarebasiert) oder durch eigene Leitungen kann man eine Mischung zwischen Paketen der Echtzeitkommunikation und des klassischen Datenverkehrs vermeiden. Unabhängig von dem Codec und der RTP-Rate zeigen diese Leitungen die größte Leistungsfähigkeit, sie transportieren bei vergleichbaren Werten am meisten VoIP-Kanäle mit geringen Wartezeiten. Diese strikte Trennung von Sprache und Daten ist auf jeden Fall zu bevorzugen.

- Priorisierung: DiffServ (Strict Priority Queueing) ermöglicht einen bevorzugten Transport der Echtzeitinformation. Für kleinere Verhältnisse und einen geringen VoIP-Anteil kann dies ein möglicher Weg sein. Gegenüber der exklusiven Nutzung wird aber mehr Bandbreite für die gleiche Anzahl von VoIP-Kanälen benötigt.

Überdimensionierung: Das Argument, die Strecke sei ja nur gering mit Daten ausgelastet, trägt nicht. Ohne QoS-Maßnahmen können nur sehr wenige Kanäle mit sehr großen Verzögerungszeiten realisiert werden. Zudem kann man sich durch eine ungenügende Verkehrstrennung Sicherheitsprobleme einhandeln, weil die Sprach- und Datenkommunikation im gleichen Netz stattfindet. Mit MPLS oder VLAN kann man beide logisch voneinander trennen. Der Transport von VoIP-Paketen ohne QoS-Maßnahmen erscheint damit insgesamt nicht empfehlenswert. Wie das letzte Beispiel zeigt, bringt eine Trennung durch VLAN oder MPLS deutlich mehr VoIP-Verkehr mit geringeren Verzögerungszeiten durch das Netz, und die Datenkommunikation wird dabei kaum beeinträchtigt (ggf. geringe Verzögerungen durch Traffic Shaping).

Eine kleinere MTU-Size ist für eine Verkehrsmischung mit RTP-Paketen vorteilhaft. Je kleiner die IP-Pakete sind, umso kleiner sind die Belegungszeiten durch diese Pakete. Die kleinen RTP-Pakete können sich dann eher zwischen zwei kleinere IP-Pakete übertragen. In Netzen mit großen Jumbo-Frames ist mit sehr großen Paketlaufzeiten und einem großen Jitter zu rechnen.

6.1.2 Wie viele VoIP-Kanäle können je 2,048-Mbit/s-Strecke übertragen werden?

Wie drastisch sich die unterschiedlichen Maßnahmen auswirken, zeigt die folgende Darstellung. Im ersten Fall wird eine 2,048-Mbit/s-Strecke mit TDM-Technik betrieben. Durch die festen Zeitkanäle und die genaue Anpassung dieser Technik an die Übertragung von Sprachinformationen können hier 30 Kanäle übertragen werden. Ohne eine Belastung durch eine parallel laufende Datenkommunikation können bei der gleichen Übertragungsstrecke mit VLAN bzw. MPLS bis zu 26 Kanäle unterstützt werden (abhängig von der Wahl der RTP-Paketgröße).

Mit einer (über die gleiche Strecke übertragen) Last durch Datenkommunikation von 5 % sind es nur noch max. 24 Kanäle (bei 40 ms je RTP-Paket) bzw. 21 (20 ms je RTP-Paket) oder 16 (10 ms je RTP-Paket). Mit DiffServ sind es bei der gleichen Datenlast von 5 % nur noch 11 bis 21 VoIP-Kanäle, ohne QoS sind es nur 7 bis 12 VoIP-Kanäle (H = 0,8).

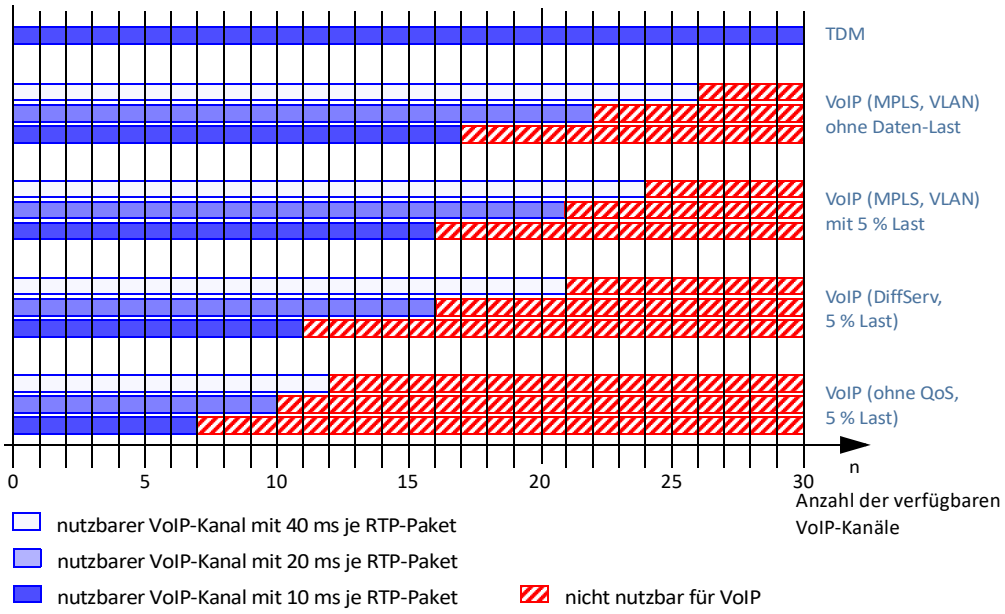


Abbildung 43: Vergleich der Maßnahmen

6.2 Was passiert, wenn ...?

6.2.1 Wie stellt sich ein Netz ohne QoS-Maßnahmen dar?

Was passiert, wenn man das spezifische Netzverhalten nicht beachtet und einfach die Bandbreitenanforderungen addiert und keine QoS-Maßnahmen ergreift?

- Meistens geht es gut. Die meisten LAN sind kräftig überdimensioniert, und die wirklichen Belastungsfälle treten selten auf.

- Manchmal ist dann die Verständigung in diesen Systemen sehr schlecht. Wenn die Belastungsfälle eintreten, wirken sie sich unmittelbar auf die Echtzeitkommunikation aus, viel früher als auf die klassische Datenkommunikation.

- Viele VoIP-Telefone gleichen einen Paketverlust durch einen eingebauten Mechanismus besser als andere Terminals aus. Obwohl die Pakete tatsächlich fehlen, ist es für die Benutzer kaum zu merken.

- Die klassische Datenkommunikation wird durch die vielen RTP-Pakete beeinträchtigt. Die Netzlast steigt durch VoIP signifikant an. Hierfür sind die vielen zusätzlichen kurzen RTP-Pakete verantwortlich, nicht die zusätzlich erforderliche Datenrate. Mit der Paketrate (Ankunftsrate λ) steigt die Netzauslastung ρ.

- Diese Fälle, in denen die Qualität der Sprachkommunikation sehr schlecht wird, lassen sich kaum reproduzieren. Um solche Belastungen zu rekonstruieren, müsste man eine große Belastung im Bereich der Sprachkommunikation und auch im Bereich der Datenkommunikation erzeugen.

- Netzmessungen sind hier nur dann aussagefähig, wenn gleichzeitig die Belastungsfälle durch spezielle Generatoren hervorgerufen werden.

- In Netzen, die ursprünglich über sauber getrennte Bereiche für die Sprach- und Datenkommunikation verfügten, können unsachgemäße „Umsteckaktionen“ diese Trennung aufheben und doch wieder zu einer Mischung von Sprach- und Datenkommunikation führen.

6.2.2 Messen und Monitoring

Um netzbedingten Problemen mit der VoIP-Sprachqualität, die oftmals auch nur sporadisch auftreten, entgegenzuwirken oder um diesen vorzubeugen, steht die Einrichtung von QoS-Maßnahmen im Netz zur Verfügung. Wie sollten diese im Einzelfall ausgestaltet werden? Wie geeignet ist das Netz für die VoIP-Anwendung? Liefert die eingerichtete Maßnahme die angestrebte Qualitätsverbesserung?

Um hierfür die benötigten Erkenntnisse zu gewinnen, können im Netz Messungen als Momentaufnahmen durchgeführt werden oder ein (VoIP-)Monitoring eingerichtet werden, mit dem sich auch eine laufende Qualitätssicherung unterstützen lässt. Dabei ist „Messung“ nicht gleich „Messung“, verschiedene Faktoren wie das gewählte Messverfahren, der Messaufbau sowie Zeitpunkt bzw. Zeitraum der Messungen bestimmen vorweg, welche Aussagekraft die Ergebnisse besitzen können.

Messung ohne/mit Last
Man kann das Netz durch einfache Messungen kennenlernen, dabei sind einfache Messungen immer nur wenig aussagefähige Augenblickswerte. Mehr Erkenntnisse liefern provozierende Tests mit bekannten Testverkehrsquellen, die einen vorher bestimmten Verkehr erzeugen. Mit dieser Testlast können dann Messungen in Grenzbereichen der Netzauslastung durchgeführt werden. Mit diesen Aussagen ist das Netzverhalten etwas besser vorauszusagen.

Messung mit Lastgeneratoren
Testlast kann mit Lastgeneratoren erzeugt werden. Diese speisen im Idealfall möglichst typische Daten, wie sie im (geplanten) Betrieb des VoIP-Systems vorkommen, als Lastproben in das vorhandene Netz ein. Die Durchführung von Lasttests zu verkehrsschwachen Zeiten begrenzt potenzielle Störungen im Netzbetrieb. Im Verlauf der Messungen können die netztypischen Parameter (mittlere und maximale Paketlaufzeit, mittlerer und maximal gemessener Jitter, Paketverlust ggf. MOS, R-Faktor oder PESQ, POLQA usw.) unter verschiedenen Bedingungen (Einzelmessung ohne Stresslast, mit einer durchschnittlichen Netzbelastung und einer Stressbelastung) gemessen werden. Messungen nach Inbetriebsetzung eines VoIP-Systems können als Momentaufnahme Auskunft über die Robustheit des Netzes als Basis für das VoIP-System geben.

Veränderungen im Netz
Allerdings ist auch mit einer Lastmessung nicht das Verhalten des Trägernetzes vollständig beschrieben und die Verhältnisse können sich zudem ändern. Ein sauber mit VLAN logisch getrenntes Netz für die Sprachübertragung kann beispielsweise durch ein unachtsames Umstecken einer Ethernetleitung dra-

stisch verschlechtert werden. Oft ist dies aber nicht sofort bemerkbar, weil die Netzauslastung der Datenkommunikation sehr stark schwankt. Hieraus resultieren dann Aussagen wie: „Letzte Woche Dienstag hatten wir eine sehr schlechte Verständlichkeit im Netz." In Zeiten mit einem geringen Datenverkehr bleibt der Fehler unentdeckt, in Zeiten mit großer Last kommt es zu sehr großen Paketlaufzeiten und einem erhöhten Paketverlust. Das gilt auch für die Datenkommunikation, VoIP-Anwendungen bringen mit ihren RTP-Paketen, die in kurzen Abständen gesendet werden, eine hohe Last in ein Datennetz. Werden beide Anwendungen nicht durch eigene, virtuelle Kanäle getrennt (Verkehrstrennung, VLAN), kommt es zu Übergriffen („Seitdem die TK-Anwendung im Netz ist, ist es viel schlechter geworden."). Die Priorisierung von Echtzeitkommunikation wirkt sich benachteiligend auf die Datenkommunikation aus. Solange der Anteil der bevorrechtigten Pakete klein bleibt, geht das ohne Probleme, bei einem größeren Anteil von bevorrechtigtem Verkehr „leidet" die klassische Datenkommunikation merklich.

Monitoring
Für eine laufende Sicherung der Qualität können in die Systeme eigene Mess-Clients eingebaut werden, die Probeverbindungen unterhalten. Diese Testverbindungen werden dann sehr genau gemessen und beobachtet, die Messergebnisse werden gesammelt und verdichtet. Änderungen an der Netzkonfiguration können so frühzeitig entdeckt und ggf. in einem Managementsystem angezeigt werden.

Alternativ können die Daten der durchgeführten Verbindungen erfasst und ausgewertet werden. Während eines VoIP-Gesprächs werden neben den RTP-Daten auch sog. RTCP-Daten ausgetauscht, die Informationen über die Eigenschaften der Übertragung enthalten. Mit RTCP werden während der Verbindung Rückkopplungen zur Übertragungsqualität an den Sender gegeben.

A Anhang

A.1 QoS-Maßnahmen bei VLAN und MPLS

Die Berechnungen haben gezeigt, dass eine strikte Trennung von Sprache und Daten (wie durch MPLS und VLAN) bei den Vergleichen immer am besten abgeschnitten hat. Die Bereitstellung eigener, virtueller Kanäle ermöglichte eine strikte Trennung zwischen den beiden Verkehrsarten und dadurch die meisten VoIP-Kanäle mit den geringsten Verzögerungen. Wie ist es möglich, dass so viel mehr VoIP-Kanäle in VLAN/MPLS-Systemen übertragen werden können?

A.1.1 Grundprinzip von VLAN und MPLS

Vereinfacht dargestellt, ordnen VLAN und MPLS jeder virtuellen Verbindung ein Label zu. Gleichzeitig werden für jede Verbindung die Verkehrseigenschaften vorgegeben. Die Eigenschaften der virtuellen Verbindungen werden laufend überwacht (Policing). Diese Überwachung kann auch für einen Port mit allen virtuellen Verbindungen angewendet werden. Des Weiteren kann der gesamte Verkehr von einem bestimmten Port einer bestimmten Verkehrsart zugeordnet werden (Port-based QoS). Überschreitungen der vereinbarten Eigenschaften werden nicht zugelassen. Die Netzelemente überwachen die tatsächlich verwendeten Verkehrsparameter (sog. Usage-Parameter) der aktiven Verbindungen. Die Missachtung der festgelegten Parameter durch eine bestimmte Quelle kann andere Verbindungen beispielsweise durch eine erhöhte Laufzeit dieser Zellen innerhalb des Netzelements beeinflussen. Unter Umständen können diese Laufzeiten so groß werden, dass die Qualitätsanforderungen dieser Verbindung nicht mehr eingehalten werden können. Die Verkehrsparameter werden je virtuelle Verbindung (je Label) getrennt überwacht.

Bei der Verletzung der Verkehrsparameter werden von einer Kommunikationsquelle mehr Pakete gesendet, als beim Verbindungsaufbau vereinbart wurden. Die zu viel produzierten Zellen können von Netzelementen

- verworfen werden – das bedeutet Paketverlust,
- verzögert werden, bis die Quelle weniger Pakete erzeugt (*Traffic Shaping*).

A.1.2 Beispielverkehr

In dem folgenden Beispiel werden von drei unterschiedlichen Datenquellen unabhängig voneinander Pakete gesendet (Quelle 1 bis 3). Parallel werden von einer VoIP-Quelle (Quelle 4) in regelmäßigen Abständen RTP-Pakete über die gleiche Schnittstelle gesendet.

Wird der Verkehr der vier Quellen ohne QoS-Maßnahmen zusammengefasst, setzen sich vor allem die größeren Datenpakete durch. Die Verzögerungszeiten für die kleinen RTP-Pakete sind sehr groß und variieren sehr stark (Bild 45a).

A.1.3 DiffServ

Werden die RTP-Pakete vorrangig transportiert, z. B. durch DiffServ (Strict Priority), erfolgt die Reihenfolge beim Pakettransport nicht in der Reihenfolge des Eintreffens der Pakete. Die kleinen RTP-Pakete können sich hier etwas „vordrängeln“, allerdings müssen sie warten, wenn gerade ein langes Datenpaket bearbeitet wird. Durch das Eintreffen eines bevorrechtigten Pakets wird der Bearbeitungsvorgang eines anderen Pakets nicht unterbrochen. Die Verzögerungszeiten sind hier bereits deutlich kürzer als ohne QoS-Maßnahmen (Bild 45b).

A.1.4 Übertragung mit VLAN bzw. MPLS

Bei VLAN und MPLS werden den Verkehrsquellen bestimmte Verkehrseigenschaften zugeordnet. Die Einhaltung dieser Eigenschaften wird ständig überwacht. Für den vorgegebenen Beispielverkehr bedeutet dies, dass die eintreffenden Datenpakete dem vereinbarten Verkehr (in Pakete/s) angepasst werden. Pakete, die in sehr kurzem Abstand hintereinander eintreffen, werden zeitlich auseinandergezogen (Bild 45c).

Mit einem Takt, der der vereinbarten Übermittlungsrate entspricht, werden die Pakete weitergeleitet. Die in unregelmäßigen (sehr kurzen) Abständen eintreffenden Datenpakete werden in regelmäßigen (der vereinbarten Datenrate entsprechenden) Abständen weitergeleitet, dies wird auch als Traffic Shaping (Verkehrsanpassung) bezeichnet. Dieser Mechanismus arbeitet in den hardwarebasierten MPLS- oder VLAN-Systemen sehr verlässlich. Diese strikte Trennung von VoIP- und Datenverkehr muss aber durchgängig gewährleistet werden. Werden für die externe Anschaltung doch wieder unterschiedliche Verkehre gemeinsam über eine Leitung bzw. einen logischen Kanal übertragen, können die guten Eigenschaften hierbei verloren gehen. Softwarebasierte Systeme sind von der Realisierung der jeweiligen Hersteller und der Last im System abhängig. Hier wurden nur die HW-basierten Systeme betrachtet. In dem ersten Bild wird der Verkehr der Quelle 2 angepasst.

Werden nun die angepassten Verkehre zusammengefasst, entspannt sich die Situation für die Echtzeitübertragung. Die Verzögerungszeiten sind bei diesem Verfahren deutlich geringer als bei DiffServ oder einer Übertragung ohne Maßnahmen. Durch die zeitliche „Dehnung“ des Eintreffens der Datenpakete sind die Paketabstände groß genug, dass die Pakete der Echtzeitkommunikation sich durchsetzen. Häufig blockieren hier auch keine langen Datenpakete die Bearbeitung der RTP-Pakete.

Die Effekte der verschiedenen QoS-Mechanismen machen sich bereits bei diesem einfachen Beispiel bemerkbar. In der Realität sind die Datenpakete im Verhältnis zu den Paketen der Echtzeitkommunikation viel größer, d. h., die Behinderung der RTP-Pakete ist noch viel stärker, als hier darstellbar. Dieses zeigt sich in den vereinfachten Berechnungen im folgenden Abschnitt.

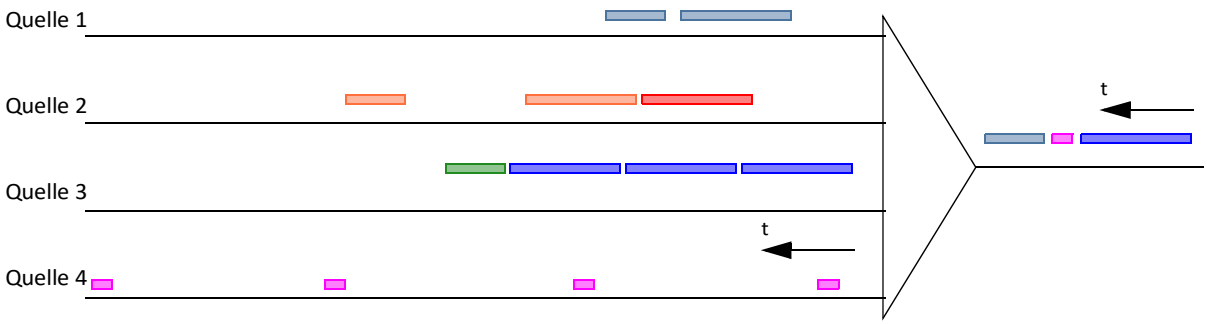


Abbildung 44: Musterverkehr

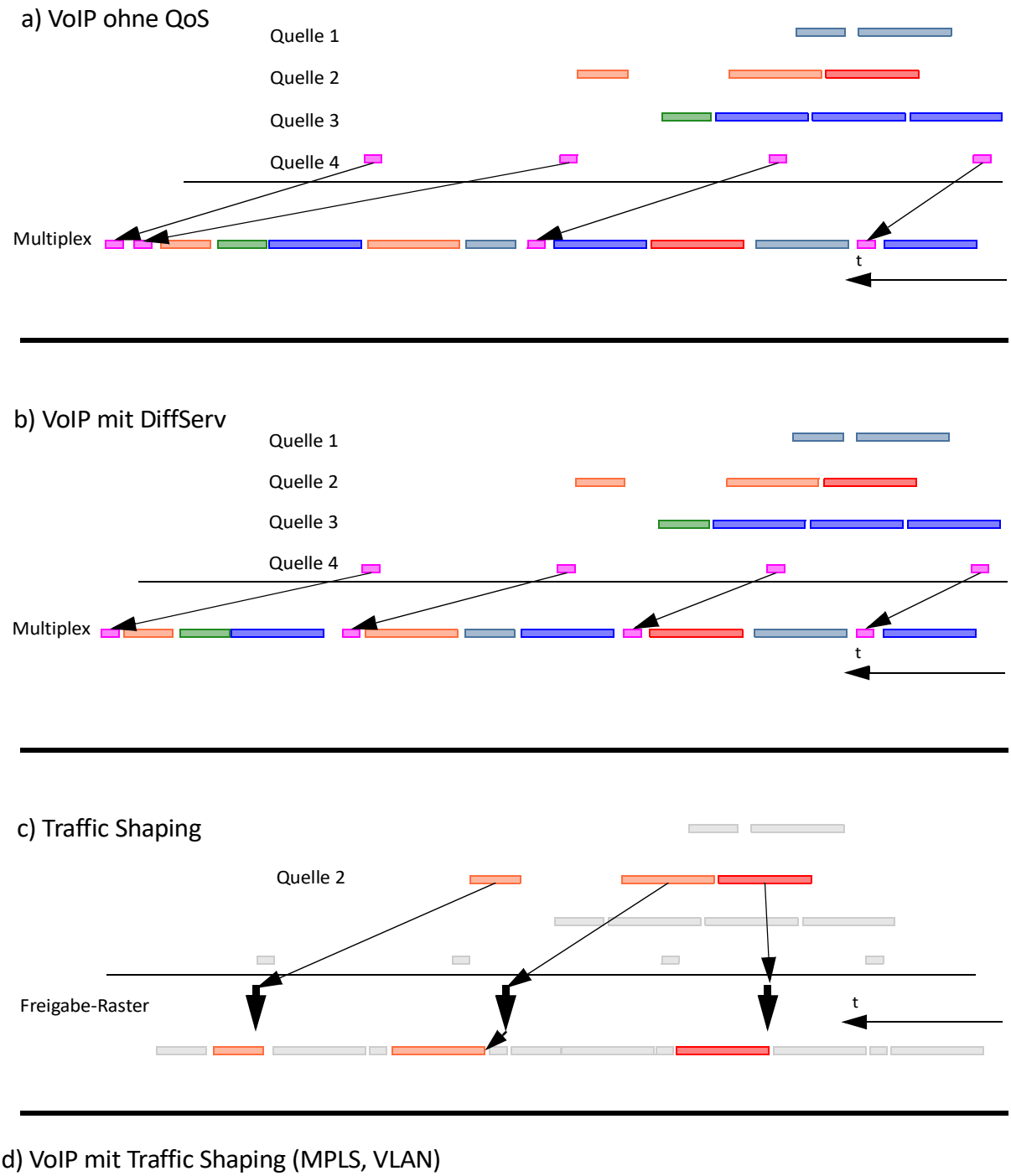


Abbildung 45: Wirkung von QoS-Maßnahmen

A.2 Jitter

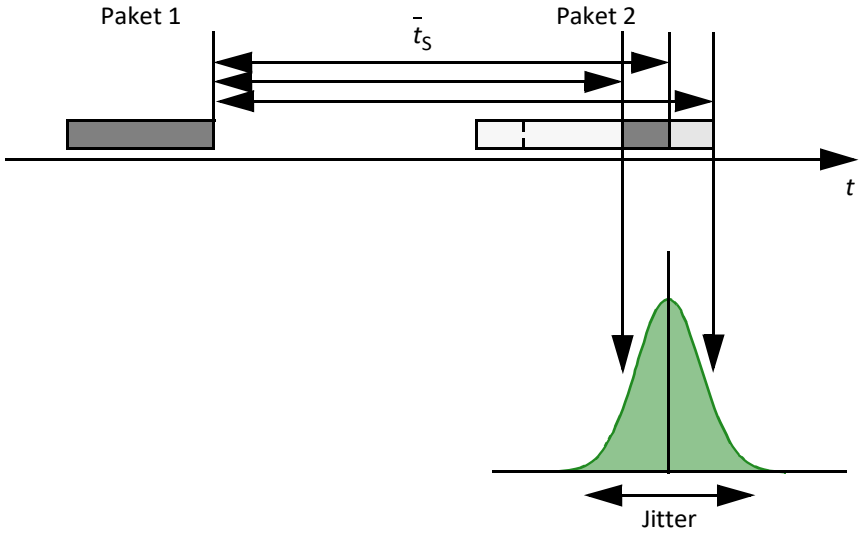


Abbildung 46: Jitter und Delay

Die Laufzeit eines RTP-Pakets schwankt von Paket zu Paket. Diese Varianz der Laufzeit eines einzelnen RTP-Pakets wird als Jitter bezeichnet.
Werden auf einer Übertragungsstrecke nur RTP-Pakete übertragen, bleibt der Jitter relativ klein. In einer Mischung mit den viel längeren IP-Paketen werden die mittlere Paketlaufzeit und der Jitter größer. Durchläuft das RTP-Paket mehrere Netzelemente, in denen eine Verkehrsmischung aus RTP-Paketen und Datenpaketen transportiert wird, wird der Jitter von Netzelement zu Netzelement immer größer.

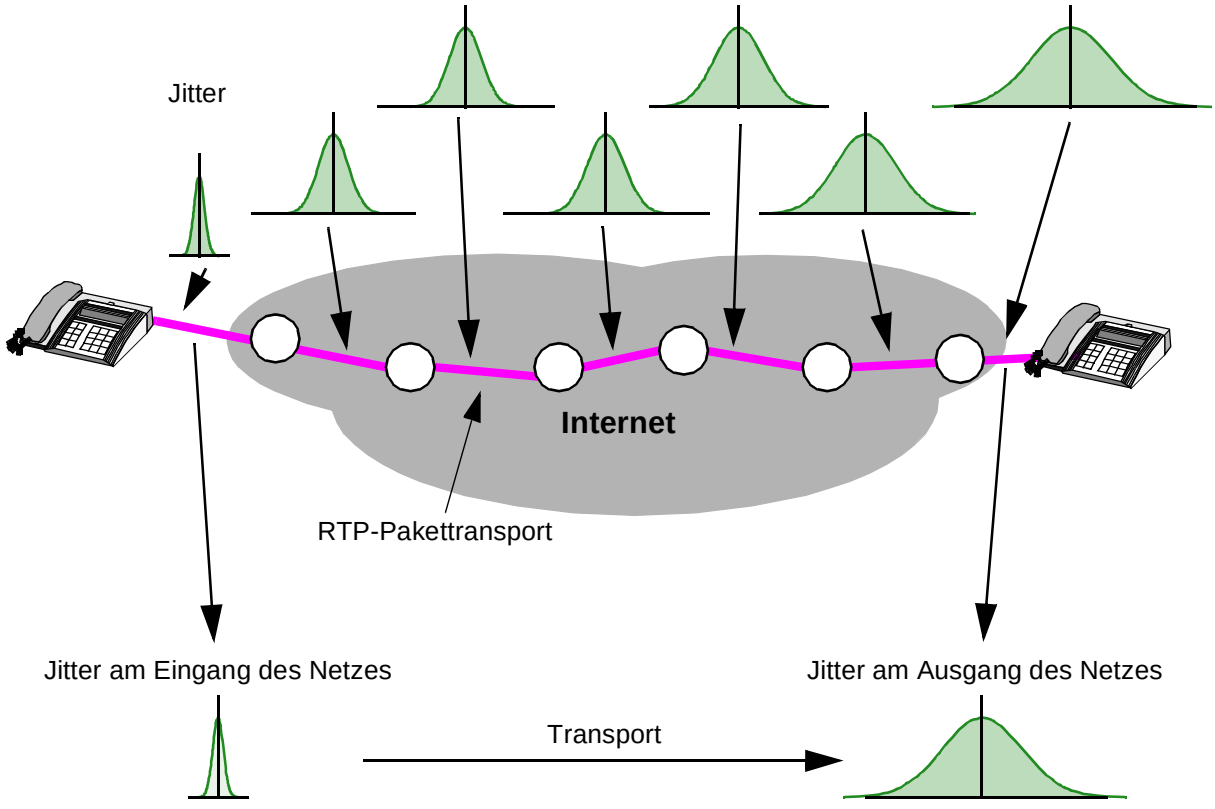


Abbildung 47: Mehr Netzelemente größerer Jitter

A.3 Die genaueren Berechnungen

Die oben beschriebenen Zusammenhänge liegen in den Eigenschaften der Netze und der eingebrachten Verkehre begründet. Diesem Verhalten versucht man sich mit mathematischen Modellen zu nähern, um diese besser verstehen und berechnen zu können. Die Berechnungsgrundlagen liegen in den Wartezeitmodellen begründet. Im Folgenden werden die verwendeten Formeln und Berechnungsansätze zusammenfassend dargestellt.

In den Darstellungen oben und in der Berechnung wurden einige Vereinfachungen gemacht, um die Zusammenhänge und Systemeigenschaften zu verdeutlichen. Genauere Berechnungen wären deutlich aufwendiger, deren Ergebnisse aber nicht unbedingt viel besser. Das liegt insbesondere in den speziellen Eigenschaften des Internetverkehrs begründet.

Der IP-Verkehr zeigt ein selbstähnliches Verhalten und eine Langzeitabhängigkeit. Mit diesen beiden Eigenschaften ergeben sich Mittelwerte im beobachteten Verkehr erst nach sehr langen Beobachtungszeiten. Diese Mittelwerte weisen zudem eine sehr große Varianz auf, die unter bestimmten Umständen gegen unendlich strebt. Damit werden die ermittelten Werte (ob gemessen oder errechnet) sehr ungenau und erlauben kaum Vorhersagen auf ein zukünftiges Verhalten.

A.3.1 Wichtige Anmerkungen zu den Berechnungen

Die Berechnungen wurden für diese Betrachtungen etwas vereinfacht:

- Für DiffServ wurde der Berechnungsweg etwas vereinfacht. Eigentlich ist das Verfahren für einen sehr geringen Anteil mit bevorrechtigtem Verkehr vorgesehen (ca. 5 %). Bei größeren Anteilen (die in diesen Berechnungen betrachtet wurden) wären die Berechnungen dann viel komplexer – hier sollte nur der Trend dargestellt werden, daher wurde auch bei größeren Anteilen vereinfacht gerechnet.
  Weiterhin wurde von einem „Strict Priority Queueing“ ausgegangen, d. h., dass die RTP-Pakete mit einem absoluten Vorrang bearbeitet werden. Für die RTP-Pakete existiert eine eigene Warteschlange. Ist ein Paket in dieser Warteschlange vorhanden, wird es als nächstes vom Router bearbeitet (DiffServ-Router). Ist kein Paket in dieser Warteschlange vorhanden, geht der Router zur Warteschlange mit der nächstniedrigeren Priorität über. Ist hier ein Paket vorhanden, wird es bearbeitet. Danach geht der Mechanismus wieder in die bevorrechtigte Warteschlange zurück. Ist hier inzwischen wieder ein Paket vorhanden, wird es bearbeitet.
  Andere Mechanismen arbeiten mit bestimmten Vorgaben für die Bearbeitung (Fair Queueing, Weighted Random Queueing usw.), in diesem Fall werden die Wartespeicher mit den verschiedenen Prioritäten nach den Vorgaben bearbeitet. Für den Transport von RTP-Paketen kann dies aber bedeuten, dass ein oder mehrere IP-Pakete mehr als bei dem Strict-Priority-Verfahren zwischen den RTP-Paketen liegen können. Die Anzahl der übertragbaren VoIP-Kanäle läge bei diesen Verfahren zwischen den Ergebnissen von DiffServ (Strict Priority) und der Übertragung ohne QoS-Maßnahmen (Überdimensionierung).

- Für den Transport mittels MPLS und VLAN wurde bei den Berechnungen von einer hardwarebasierten Lösung ausgegangen. Es sind auch softwarebasierte MPLS- oder VLAN-Lösungen auf dem Markt, die nicht über ein vergleichbares Verhalten verfügen. Softwarebasierte Lösungen sind in ihrer Leistungsfähigkeit von der Belastung und der jeweiligen Realisierung durch den Hersteller abhängig. Allgemeine Berechnungen sind hier nur schwer möglich. Die Ergebnisse wären Hersteller- und Lastabhängig.
  Die Berechnungen für MPLS- und VLAN-Systeme mit QoS und Policing-Funktionen wurden in die idealisierten Berechnungen für exklusive Leitungen übernommen. Dies ist eine Vereinfachung der wahren Verhältnisse, die deutlich komplexer berechnet werden müssen. Hier sollten nur die grundsätzlichen, unterschiedlichen Eigenschaften der verschiedenen QoS-Maßnahmen dargestellt werden. Es ist kein Ansatz für eine exakte Berechnung der Verhältnisse.

- Die Berechnungen ohne QoS-Maßnahmen wurden ebenfalls vereinfacht. Bei genaueren Betrachtungen müsste der Hurst-Parameter mit steigender RTP-Last verringert werden – dies wurde hier vernachlässigt. Insgesamt werden die Berechnungen bei genaueren Betrachtungen sehr viel komplexer, aber nicht genauer. Aus diesem Grund wurde auch hier der vereinfachte Ansatz gewählt.

Alle Ergebnisse sind Durchschnittswerte, gerade bei der Verkehrsmischung können diese um ein Vielfaches überschritten werden. Auf den folgenden Seiten sind die grundsätzlichen mathematischen Ansätze und die verwendeten Formeln ohne weitere Erläuterungen zusammenfassend dargestellt.

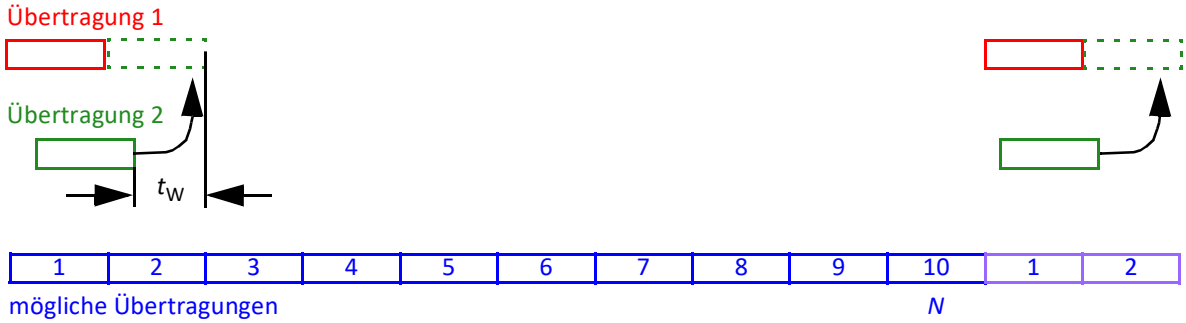


Abbildung 48: Entstehung von Wartezeiten

A.3.2 Berechnung der mittleren Wartezeit für den exklusiven RTP-Transport [Klot11]

Die mittlere Wartezeit für ein RTP-Paket hängt bei einer exklusiven Übertragung von RTP-Paketen nur von der gewählten RTP-Größe, der Übertragungsgeschwindigkeit und von der Anzahl der laufenden Verbindungen ab. Werden die RTP-Pakete alle 20 ms übertragen, ist die Zeit für den Transport t_S eines RTP-Pakets: (230 · 8 bit)/2,048 Mbit/s = 0,898 ms. Für 2 RTP-Pakete ist die Zeit 2 · 0,898 ms = 1,796 ms, für 3 RTP-Pakete ist die Zeit 3 · 0,898 ms = 2,694 ms usw. Allgemein ist t_s:

$$t_S = \frac{\text{Paketgröße in Byte} \cdot 8\ \text{bit}}{\text{Übermittlungsrate in bit/s}}$$

Für ein Paket mit einem RTP-Abstand von 20 ms ist ein RTP-Paket 160 Byte groß. Zusammen mit dem Overhead von RTP und den der unterliegenden Protokolle (in Summe 70 Byte) macht das 230 Byte. Die Zeit für den Pakettransport ist mit einer Übermittlungsrate von 2,048 Mbit/s:

$$t_S = \frac{230 \cdot 8\ \text{bit}}{2,048 \cdot 10^6\ \text{bit/s}}$$

In einem Zeitrahmen können N RTP-Pakete gesendet werden (hier N=10). Kommt ein zweites Gespräch dazu, passiert meistens nichts, da ja noch genügend freie Kapazität vorhanden ist. Allerdings könnte das erste Paket des zweiten Gesprächs während der Sendezeit eines Pakets des ersten ankommen, dann muss es warten.

Mit einer Wahrscheinlichkeit von 1/N muss das zweite Paket warten, die mittlere Wartezeit ist dann $t_S/2$. Das erste Paket wartet nicht (t_{W1}=0), das zweite wartet:

$$t_{W2} = \frac{1}{N} \cdot \frac{t_S}{2}$$

Die mittlere Wartezeit bei zwei Gesprächen ist dann:

$$\overline{t_W} = (t_{W1} + t_{W2})/2$$

$$\overline{t_W} = \left(0 + \frac{1}{N} \cdot \frac{t_S}{2}\right)/2 = \frac{1}{N} \cdot \frac{t_S}{4}$$

Bei 3 Gesprächen warten die Pakete des ersten nicht, die des zweiten t_{W2}=1/N* t_S/2 und die des dritten t_{W3}=2/N* t_S/2, da die Wahrscheinlichkeit des Wartens für die Pakete des dritten Gesprächs größer ist. Die mittlere Wartezeit ist jetzt

$$\overline{t_W} = (t_{W1} + t_{W2} + t_{W3})/3$$

$$\overline{t_W} = \left(0 + \frac{1}{N} \cdot \frac{t_S}{2} + \frac{1}{N} \cdot \frac{2 \cdot t_S}{2}\right)/3$$

Allgemein ist die mittlere Wartezeit für n von N möglichen Verbindungen:

$$\overline{t_W} = \frac{t_S}{2 \cdot N \cdot n} \sum_{i=1}^{(n-1)} i$$

$$\overline{t_W} = \frac{t_S}{2 \cdot N \cdot n} \cdot \frac{(n-1)n}{2} = \frac{t_S \cdot (n-1)}{4 \cdot N}$$

Werden die RTP-Pakete alle 10 ms erzeugt, ist die maximal mögliche Anzahl von VoIP-Kanälen (N) bei einer Übermittlungsrate von 2,048 Mbit/s:

$$N = \frac{2,048 \cdot 10^6\ \text{bit/s} \cdot 0,01s}{150\ \text{Byte} \cdot 8\ \text{bit}}$$

Werden die RTP-Pakete alle 20 ms bzw. alle 40 ms gesendet, ist N:

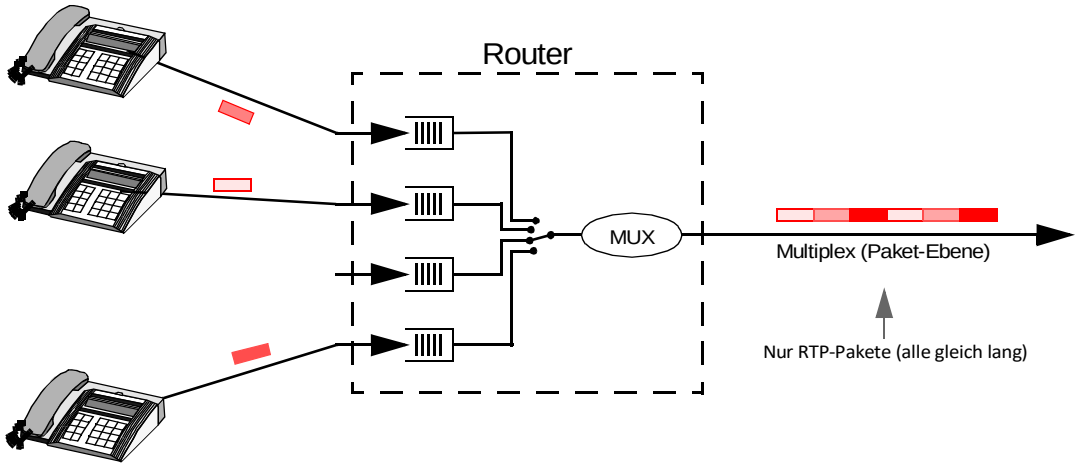
$$N = \frac{2,048 \cdot 10^6\ \text{bit/s} \cdot 0,02s}{230\ \text{Byte} \cdot 8\ \text{bit}}$$

$$N = \frac{2,048 \cdot 10^6\ \text{bit/s} \cdot 0,04s}{390\ \text{Byte} \cdot 8\ \text{bit}}$$

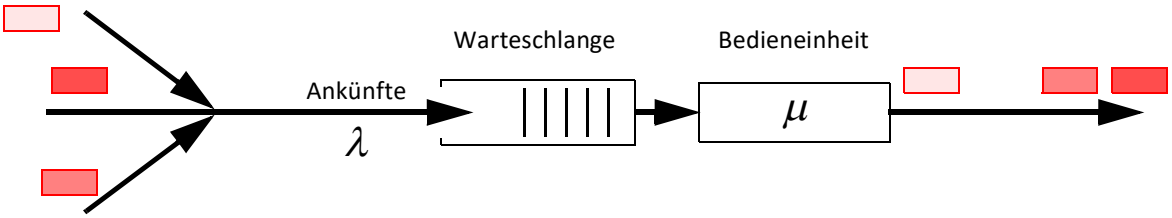
Die maximale Wartezeit wird erreicht, wenn zufällig alle laufenden Verbindungen genau gleich begonnen wurden.

Exklusiver Zugriff (MPLS und VLAN mit QoS)

Konfiguration



Modell



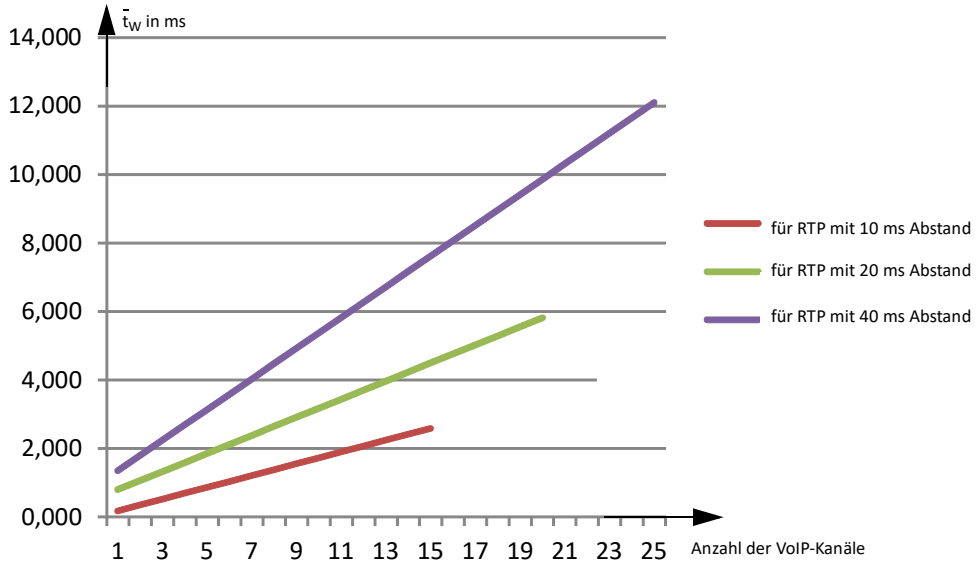
λ
nur RTP-Pakete
Feste Paket-Rate
Beginn aber zufällig

μ
alle RTP-Pakete gleich lang
Bearbeitungszeit ist immer gleich

D/D/1

Auslastung

(mittlere Wartezeit für ein Paket)

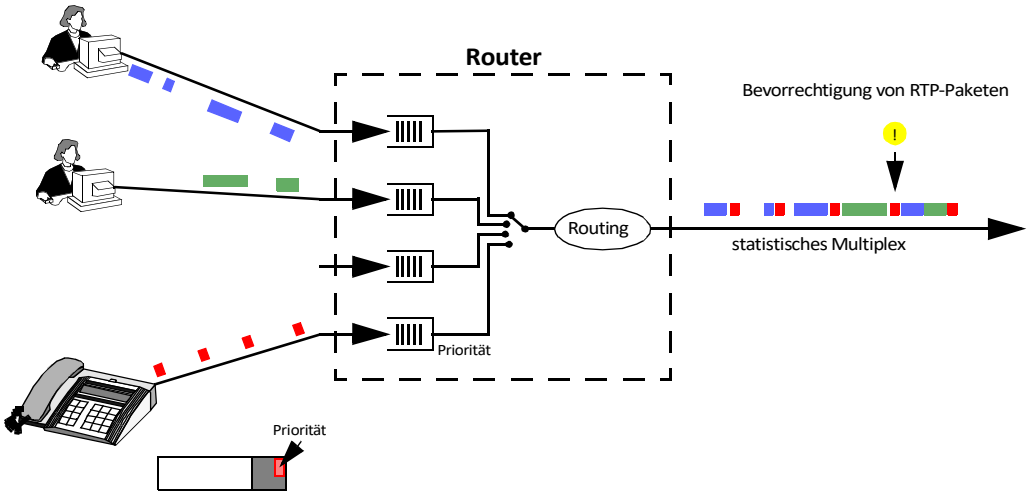


Formeln

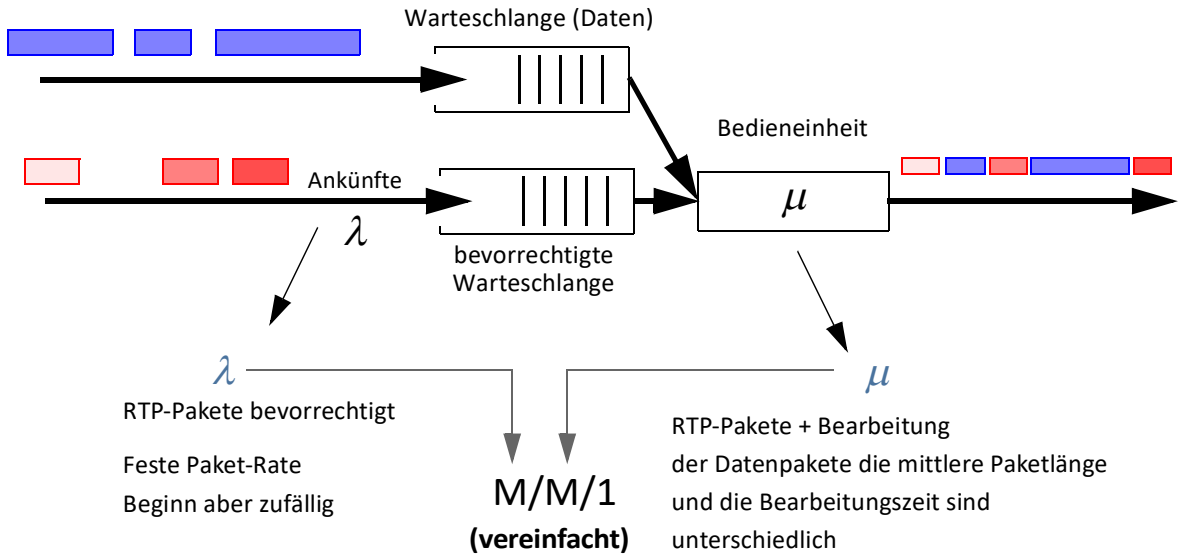
für 10 ms	für 20 ms	für 40 ms
$\bar{t}_W = \frac{150 \cdot 8\ \text{bit}}{2048 \cdot 10^3\ \text{bit/s}} \cdot \frac{(n-1)}{4 \cdot N}$	$\bar{t}_W = \frac{230 \cdot 8\ \text{bit}}{2048 \cdot 10^3\ \text{bit/s}} \cdot \frac{(n-1)}{4 \cdot N}$	$\bar{t}_W = \frac{390 \cdot 8\ \text{bit}}{2048 \cdot 10^3\ \text{bit/s}} \cdot \frac{(n-1)}{4 \cdot N}$

Priorisierung (DiffServ, strict priority queuing)

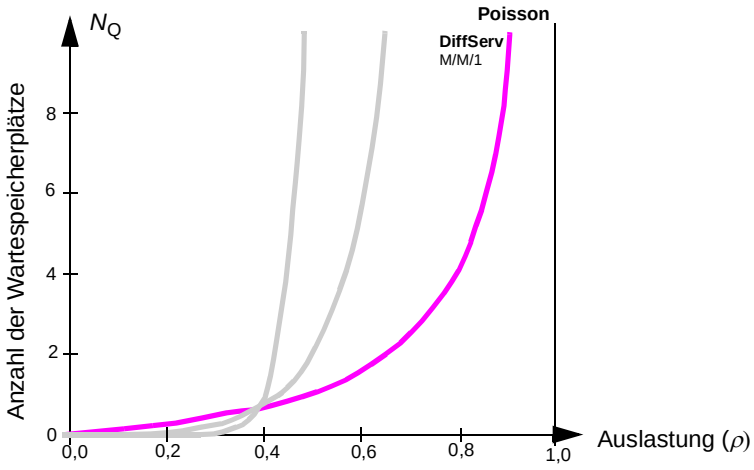
Konfiguration



Modell



Auslastung



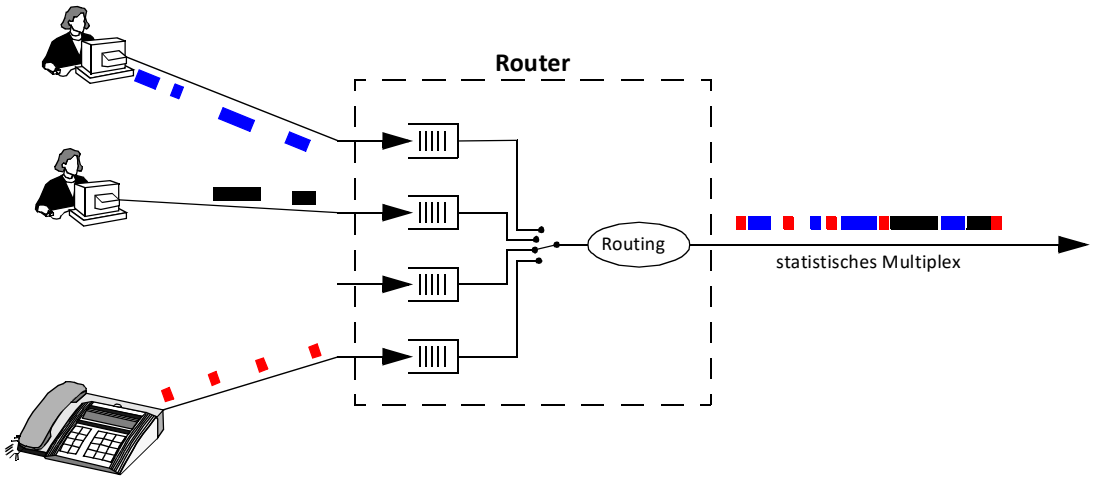
Formeln

$$\rho = \frac{\lambda}{\mu} \qquad N_Q = \frac{\rho^2}{1-\rho} \qquad \bar{t}_W = \frac{\rho}{\mu - \lambda}$$

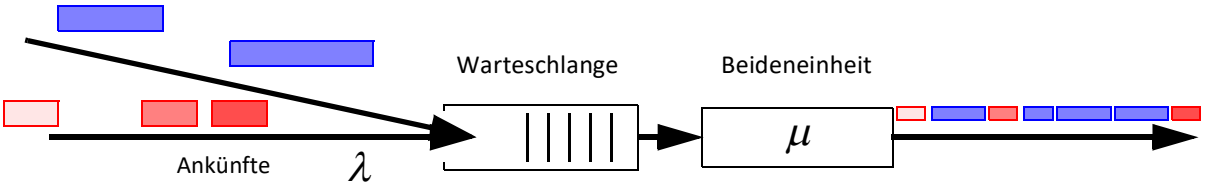
Die hier verwendeten Formeln vereinfachen das DiffServ-Prinzip auf eine einfache M/M/1-Warteschlange, für eine genauere Berechnung sind eigene Formeln für die bevorrechtigten Pakete und die benachteiligten Pakete zu verwenden [Tra05].

Überdimensionierung (kein QoS)

Konfiguration



Modell



λ

RTP- plus Daten-Pakete

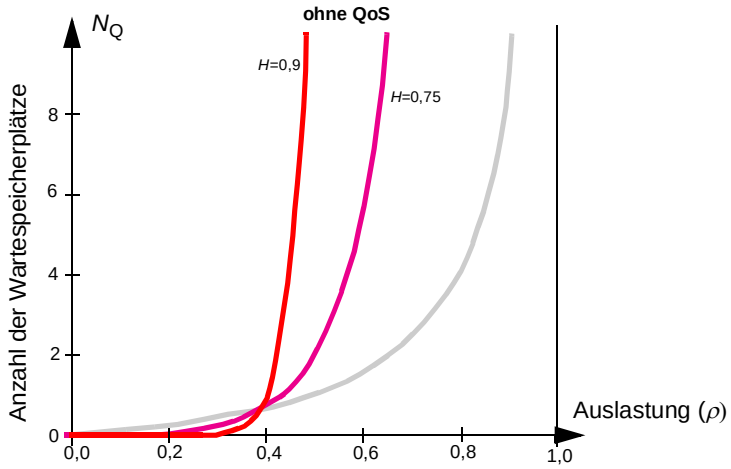
Zufällige Pakete-Rate
(selbstähnlicher Verkehr)

M/M/1
$H = 0,8$

μ

RTP-Pakete + Bearbeitung
der Datenpakete die mittlere Paketlänge
und die Bearbeitungszeit sind
unterschiedlich

Auslastung



Formeln

$$\rho = \frac{\lambda}{\mu} \qquad N_Q = \frac{\rho^{(1/(2*(1-H)))}}{(1-\rho)^{H/(1-H)}} \qquad \bar{t}_W = \frac{N_Q}{\lambda}$$

Die hier verwendeten Formeln vereinfachen die Berechnung, da mit zunehmenden VoIP-Verkehr der Hurst-Faktor kleiner wird.

A.4 Quellen, Literatur

[Gri04]: Grimm, C., Schlüchtermann, G.: Verkehrstheorie in IP-Netzen, Modelle, Berechnungen, statistische Methoden. Hüthig Verlag, Heidelberg, 2004. ISBN 3-8266-5047-6.

[Her08]: Herheuser, R.: Planung von Vermittlungsnetzen. Eine Einführung. vdf Hochschulverlag AG, Zürich, 2008. UTB-ISBN 978-3-8252-8394-0.

[Kle75]: Kleinrock, L.: Queueing Systems, Volume I: Theory. John Wiley-Interscience Publication, New York, Chichester, Brisbane, Toronto, 1975. ISBN 0-471-49110-1.

[Klot11]: Klotz, B.: Berechnung der mitteren Wartezeit, interne Aufzeichnungen. Stuttgart, 2011.

[McDysan, Paw 2002]: McDysan, D. E. und Paw, D.: ATM & MPLS Theory & Application. McGraw-Hill/ Osborne, Berkeley, USA, 2002.

[Sie09]: Siegmund, G.: Technik der Netze, Band 2: Neue Ansätze: SIP in IMS und NGN. 6. Auflage. Hüthig Verlag, Heidelberg, 2009. ISBN 978-3-7785-4063-3.

[Sie10]: Siegmund, G.: Technik der Netze, Band 1: Grundlagen, Verkehrstheorie, ISDN/GSM/IN. 6. Auflage. VDE Verlag, Berlin, Offenbach, 2010. ISBN 978-3-8007-3219-7.

[Tra05]: Tran-Gia, P.: Einführung in die Leistungsbewertung und Verkehrstheorie. 2. Auflage. Oldenbourg Wissenschaftsverlag, München, 2005. ISBN 3-486-57882-0.