View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Server für wissenschaftliche Schriften der Hochschule Hannover

Verzögerungen in paketbasierten Kommunikationsnetzen

von Rudolf Nocker

Inhalt
Bei der Übermittlung von Nachrichtensignalen in Kommunikationsnetzen treten neben der physikalischen Laufzeit weitere Verzögerungen auf, welche durch technische Bearbeitungsvorgänge (Quellencodierung, Verschlüsselung, Kanalcodierung, Leitungscodierung, Bedienungsvorgänge, Vermittlungsvorgänge) verursacht werden. In dieser Arbeit wird die mittlere Ende-zu-Ende-Durchlaufverzögerung zwischen zwei Endgeräten in Abhängigkeit von den Eigenschaften und Parametern der Endgeräte und des paketbasierten Kommunikationsnetzes allgemein berechnet.

Insbesondere wird der Mindestwert der mittleren Verzögerung in einem paketbasierten Kommunikationsnetz für Sprache (Voice over Internet Protocol, VoIP) und Daten analysiert. Die Wirkung einer Priorisierung der Sprachpakete wird untersucht. Zur Berechnung der Verzögerungen in einem Vermittlungsknoten wird das Bedienungsmodell M / G / 1 / ∞ mit Priorisierung verwendet. Die Anwendung der theoretischen Ergebnisse erfolgt auf ein stark vereinfachtes Modell-Netz mit typischen Parametern für die Netzkanten und Netzknoten. Dessen mittlere Ende-zu-Ende-Durchlaufverzögerung für Sprachsignale wird unter verschiedenen Randbedingungen (eine oder zwei Verkehrsklassen, mit oder ohne Priorisierung, unterschiedliche Parameter des Datenverkehrs) numerisch berechnet und graphisch veranschaulicht.

Schlagworte
Paketbasierte Sprachkommunikation, Next Generation Network (NGN), IP-Netz, Ende-zu-Ende-Durchlaufverzögerung, Voice over IP (VoIP), Dienstgüte, Quality of Service (QoS), Paketvermittlung, Bedienungstheorie, Wartesystem M / G / 1 / ∞ mit Priorisierung.

Autor
Prof. Dr.-Ing. Rudolf Nocker lehrt Kommunikationstechnik und Kommunikationssysteme an der Fakultät 1 (für Elektro- und Informationstechnik) der Fachhochschule Hannover.

Kontakt
Rudolf.Nocker@fh-hannover.de

Dieser Bericht wurde im April 2008 als Online-Veröffentlichung publiziert vom SHAKER-Verlag, DOI 10.2370 / 395_268.

Vorwort

In Kommunikationsnetzen treten zusätzlich zur entfernungsabhängigen, physikalischen Laufzeit der Nachrichtensignale weitere Verzögerungen auf, welche durch technische Bearbeitungsvorgänge (Quellencodierung, Verschlüsselung, Kanalcodierung, Leitungscodierung, Multiplex-Verfahren, Bedienungsvorgänge, Vermittlungsvorgänge) verursacht werden. Die Kommunikationsqualität (Dienstgüte, Quality of Service, QoS) für die Kommunikationsvorgänge in einem Kommunikationsnetz (beispielsweise für eine Fernsprech-Verbindung zwischen zwei Teilnehmern A und B) wird stark reduziert, falls die Gesamt-Verzögerungszeit von A nach B bestimmte Grenzwerte überschreitet.

In dieser Arbeit wird die Ende-zu-Ende-Durchlaufverzögerung (end to end delay, one way delay) vom Eingang des Endgeräts A zum Ausgang des Endgeräts B in Abhängigkeit von den wesentlichen Eigenschaften und Parametern des Kommunikationsnetzes allgemein berechnet. Dabei werden vorwiegend Kommunikationsnetze mit paketbasierter Übermittlung betrachtet, da bei diesen zwingend (durch die Paketisierung und die Art des Vermittlungsvorgangs) höhere Durchlaufverzögerungen auftreten als bei klassischen Kommunikationsnetzen mit Durchschalte-Vermittlungstechnik. Insbesondere wird die mindestens auftretende Ende-zu-Ende-Durchlaufverzögerung für ein paketisiertes Sprachsignal (Voice over Internet Protocol, VoIP) in einem paketbasierten Netz für Sprache und Daten betrachtet.

Der beschriebene Rechenweg zur Berechnung der Ende-zu-Ende-Durchlaufverzögerung setzt verbindungsorientierte Betriebsweise in der Netzwerkschicht (diese wird auch als Vermittlungsschicht oder OSI-Layer 3 bezeichnet) voraus. Die Ergebnisse gelten jedoch näherungsweise auch für verbindungslose Betriebsweise in der Netzwerkschicht (datagram-Betriebsweise, IP-Protokoll), wenn stabile und stationäre Netzbedingungen vorliegen. Als Subnetz-Technologie für die OSI-Schichten 1 und 2 wurde Switched-Ethernet-Technik unterstellt.

Für die Sprachsignal-Digitalisierung wurde Standard-Pulscodemodulation angenommen, da dann minimale Verzögerungen beim Umsetzvorgang auftreten. Zusätzliche Verzögerungen durch aufwendigere Quellencodierungs-Verfahren oder zusätzliche Signalverarbeitungs-Vorgänge in den Endgeräten (beispielsweise Verschlüsselung, Entschlüsselung) müssen ggf. zusätzlich berücksichtigt werden.

Eine sehr einfache Methode zur Verringerung der Durchlaufverzögerungen in einem paketbasierten Netz ist die Überdimensionierung aller Netzkanten und Netzknoten. Dadurch werden die mit ansteigender (Verkehrs-)Auslastung der Netzkanten ansteigenden Wartezeiten in den Netzknoten reduziert. Eine elegantere Methode ist die Priorisierung derjenigen Kommunikationsvorgänge, welche gegen Verzögerung besonders empfindlich sind. In einem paketbasierten Kommunikationsnetz für Sprache und Daten werden deshalb die Sprachpakete höher priorisiert als die Datenpakete.

Nachfolgend wird der Einfluss einer derartigen Priorisierung untersucht. Es wird allgemein und numerisch berechnet, welche Wartezeiten sich ohne bzw. mit Priorisierung in den Paket-Vermittlungsknoten bei typischen Verkehrs-Belastungen und Übertragungsgeschwindigkeiten ergeben. Dabei werden zwei Service-Klassen unterschieden, die Service-Klasse „Sprache“ mit hoher Priorität und die Service-Klasse „Daten“ (alle sonstigen Kommunikationsanwendungen in diesem Netz) mit niedriger Priorität.

In Kapitel 1 erfolgt zunächst eine Einführung in die Thematik, dann werden die prinzipiellen Möglichkeiten zur Verringerung der Ende-zu-Ende-Durchlaufverzögerung skizziert.

In Kapitel 2 wird auf die Dienstgüte für Sprach-Dialog-Anwendungen eingegangen. Die Sprach-Dialog-Dienstgüte wird definiert als Kombination von Sprach-Wiedergabe-Qualität (bewertet durch den MOS-Wert nach ITU-Empfehlung P.800) und maximal zulässiger Ende-zu-Ende-Verzögerungszeit. Die entsprechenden Zahlenwerte werden diskutiert.

In Kapitel 3 werden die Grundlagen zur Berechnung der Ende-zu-Ende-Durchlaufverzögerung in Kommunikationsnetzen zusammen gestellt. Die gesamte Ende-zu-Ende-Durchlaufverzögerung bei paketbasierter Kommunikation besteht aus der sendeseitigen Paketisierungszeit, der Gesamt-Serialisierungszeit, der Gesamt-Laufzeit, der Gesamt-Wartezeit, der Gesamt-Rechenzeit und zusätzlich (bei Sprachkommunikation) der empfangsseitigen Jitterbuffer-Verzögerungszeit (zum Ausgleich schwankender Durchlaufzeiten im Netz). Softwarebasierte Rechenzeiten in den Paket-Vermittlungsknoten werden nicht berücksichtigt.

In Kapitel 4 werden die Ergebnisse aus der Bedienungstheorie zur Berechnung der Wartezeiten in den Vermittlungsknoten aufbereitet. Zunächst werden die Ergebnisse zum Wartesystem $M / G / 1 / \infty$ ohne Priorisierung dargestellt. Damit können die Wartezeiten bei zwei Verkehrsklassen ohne Priorisierung berechnet werden. Danach werden die Ergebnisse zum Wartesystem $M / G / 1 / \infty$

mit Priorisierung diskutiert. Damit können die Wartezeiten für zwei Verkehrsklassen mit nicht unterbrechender Priorisierung berechnet werden. Die allgemeinen Ergebnisse werden durch allgemeine Diskussion und durch Zahlenbeispiele veranschaulicht.

Die Ergebnisse aus Kapitel 3 und Kapitel 4 ermöglichen bei gegebenen Eigenschaften der Netzknoten und Netzkanten eines Netzmodells sowie der Verkehrs-Auslastungen der Netzkanten durch einerseits Sprache und andererseits Daten die näherungsweise Berechnung der mittleren Ende-zu-Ende-Durchlaufverzögerung für den Sprachverkehr in einem paketbasierten Netz für Sprach- und Datenverkehr.

In Kapitel 5 werden die abgeleiteten Ergebnis-Formeln auf ein vereinfachtes Modell-Netz angewendet, um die numerische Berechnung der Ende-zu-Ende-Durchlaufverzögerung beispielhaft darzustellen. Das Modell-Netz hat 10 Vermittlungsknoten (Switches, Router) und alle Netzkanten weisen die Übertragungsgeschwindigkeit 100 Mbit/s auf. Die Entfernung AB im Netz zwischen den beiden Endgeräten sei 1000 km, die Paketisierung der Sprache wird mit typischen Parametern durchgeführt. Alle Parameter-Zahlenwerte des Netzmodells sind in Tabelle 5.2 zusammen gestellt. Das Modell-Netz entspricht „grob“ einer „nationalen Fernsprechverbindung“ über ein paketbasiertes Netz (VoIP-Fernsprechverbindung).

Die mittlere Ende-zu-Ende-Durchlaufverzögerung im Modell-Netz wird in Abhängigkeit von der Verkehrsauslastung der Netzkanten durch Sprachverkehr (Teil-Auslastung ρ_1) und Datenverkehr (Teil-Auslastung ρ_2) numerisch berechnet. Dabei werden mehrere Fälle (ohne oder mit Priorisierung des Sprachverkehrs, unterschiedliche Längen der Datenpakete) betrachtet. Die numerischen Ergebnisse für die Ende-zu-Ende-Durchlaufverzögerung in Abhängigkeit von den Teil-Auslastungen durch Sprache und Daten werden durch 3D-Diagramme graphisch veranschaulicht.

Hannover, im März 2008 Rudolf Nocker

Inhaltsverzeichnis

1 Einführung

Kommunikationsqualität in Kommunikationsnetzen

Die Dienstgüte eines Kommunikationsnetzes (Quality of Service, QoS) bewertet die Qualität einer definierten Kommunikations-Dienstleistung eines Kommunikationsnetzes (beispielsweise für Sprach-, Daten-, Bildkommunikation). Ein wesentliches Teil-Kriterium bei der Bewertung der Dienstgüte ist die Verzögerungszeit, welche ein Nachrichtensignal beim Netz-Durchgang erfährt.

In Kommunikationsnetzen treten ergänzend zur Laufzeit der elektromagnetischen Signale weitere Verzögerungen auf, welche durch technische Bearbeitungsvorgänge (Quellencodierung, Kanalcodierung, Leitungscodierung, Multiplex-Verfahren, Bedienungsvorgänge, Vermittlungsvorgänge) verursacht werden. Die Kommunikationsqualität (Dienstgüte, Quality of Service QoS) wird stark beeinträchtigt, falls die Gesamt-Verzögerungszeit bestimmte Grenzwerte überschreitet.

Beispielsweise ist für Sprach-Dialog-Kommunikation eine Ende-zu-Ende-Verzögerung „vom Mikrofon zum Hörer“ von maximal 25 ms erwünscht, maximal 150 ms gelten als noch akzeptabel, maximal 400 ms sind „bedingt akzeptabel“ (Bild 2.3). Für andere Internet-Anwendungen (beispielsweise Filetransfer, „Internet-Surfen“) sind dagegen Verzögerungszeiten im Sekunden-Bereich zulässig.

In dieser Arbeit wird untersucht, welche Mindestwerte der Ende-zu-Ende-Durchlaufverzögerung in Kommunikationsnetzen mit Durchschalte- oder Paket-Vermittlungstechnik bei vorgegebenen Verkehrsbelastungen der Netzkanten sowie vorgegebenen Eigenschaften der Endgeräte und Netzknoten auftreten.

Ende-zu-Ende-Verzögerungszeit

Eine vergleichende Abschätzung der Ende-zu-Ende-Verzögerung für das Durchschalte- und das Paket-Vermittlungsprinzip wurde in [NOCK05] durchgeführt, jedoch wurden dabei nur deterministische Anteile der Verzögerungszeit berücksichtigt. Die (vom verwendeten Vermittlungsprinzip unabhängige) physikalische Signal-Laufzeit ist bei terrestrischer Kommunikation sehr klein und wird nicht als störend empfunden. Beispielsweise ergibt sich pro 1000 km Kabelstrecke (elektrisch oder optisch) eine Laufzeit von rund 5 ms. Bei maximal 20 000 km terrestischer Entfernung (halber Erdumfang) ergeben sich bei kabelgebundener Übertragung maximal etwa 100 ms für die physikalische Signallaufzeit. Nur bei der Übertragung über geostationäre Satelliten ergeben sich höhere Signallaufzeiten.

Durchschalte-Vermittlungstechnik

Beim Durchschalte-Vermittlungsprinzip treten ergänzend zur physikalischen Signallaufzeit kleine konstante Durchschalte-Verzögerungszeiten in den Durchschalte-Vermittlungsknoten auf. Diese können mit maximal 1 ms pro Durchschalte-Vermittlung nach oben abgeschätzt werden. Bei einer interkontinentalen Durchschalte-Verbindung sind selten mehr als 10 Durchschalte-Vermittlungsknoten beteiligt, letztere erzeugen dann maximal 10 ms zusätzliche Durchschalte-Verzögerung.

Durchschalte-Vermittlungsnetze weisen resultierend konstante, kleine Ende-zu-Ende-Verzögerungszeiten auf und eignen sich deshalb hervorragend für Sprach-Dialog-Kommunikation. In [NOCK05] wurden die Ende-zu-Ende-Verzögerungszeiten in Durchschalte-Netzen für typische Beispiele berechnet.

Paket-Vermittlungstechnik

Beim Paket-Vermittlungsprinzip treten additiv zur physikalischen Signallaufzeit zusätzliche sendeseitige Paketisierungs-Verzögerungen in der Nachrichtenquelle (typische Größenordnung 20 ms), zusätzliche Verzögerungen durch Serialisierungsvorgänge in den Netzknoten, zusätzliche Wartezeiten in den Paket-Vermittlungsknoten und zusätzliche Verzögerungszeiten im empfangsseitigen Jitter-Buffer (typisch 5 ms bis 20 ms) auf.

Die Wartezeiten in den Paket-Vermittlungsknoten sind Zufallswerte, welche sowohl von der Übertragungsgeschwindigkeit der Ausgangsleitungen (dadurch wird die Serialisierungs-Dauer definiert) als auch von der Verkehrsbelastung der Ausgangsleitungen (dadurch wird die im Paket-Vermittlungsknoten auftretende Wartezeit bestimmt) abhängen.

Die derzeit verwendete Internet-Technik (IP Version 4) behandelt alle Daten-Pakete völlig gleichberechtigt. Es werden weder Reservierungs-Mechanismen noch Priorisierungs-Mechanismen eingesetzt. Es ergibt sich dann eine „bestmögliche Übermittlung“ (best effort service). Erst die nächste Internet-Generation (IP Version 6) ermöglicht eine Priorisierung ausgewählter Datenpakete.

Methoden zur Verringerung der Durchlaufverzögerung

Eine Verringerung der Durchlauf-Verzögerungen (im Vergleich zur existierenden Technik) ist prinzipiell möglich durch:

1) Überdimensionierung des Netzes (overprovisioning);
2) Einführung von Priorisierungs-Mechanismen (priorization);
3) Einführung von Reservierungs-Mechanismen (resource reservation);

Überdimensionierung (overprovisioning)
Eine Verringerung der Wartezeiten in den Paket-Vermittlungsknoten ist beispielsweise dadurch möglich, dass die Übertragungsgeschwindigkeiten aller Netzkanten (elektrische oder optische Leitungen, Funkstrecken) und die Verarbeitungsgeschwindigkeit der angrenzenden Netzknoten (switches, router) stark überdimensioniert werden (overprovisioning). Diese Methode ist immer anwendbar und erfordert keine Veränderung der bisher verwendeten Funktionsprinzipien (Übermittlungs-Protokolle, Algorithmen in den Vermittlungsknoten).

Bei Überdimensionierung (overprovisioning) sind im betrachteten (Teil-)Netz keine zusätzlichen administrativen Maßnahmen nötig. Dies gewährleistet eine unverändert einfache Wartung und Fehlersuche und vermeidet Abhängigkeiten von herstellerspezifischen Sonderlösungen [IMHO06].

Alle anderen Methoden zur Verringerung der Wartezeiten in Paket-Vermittlungsknoten erfordern technische Veränderungen in allen Netzknoten (Endknoten, Vermittlungsknoten) und erfordern erhöhten Aufwand für die Netz-Administration. Grundsätzliche Möglichkeiten sind die Einführung von Priorisierungs-Mechanismen oder die Einführung von Reservierungs-Mechanismen.

Priorisierungs-Mechanismen (priorization)
Eine Reduzierung der Wartezeiten für ausgewählte Kommunikations-Anwendungen in einem Paket-Vermittlungsknoten ist möglich durch eine Priorisierung der zugehörigen Datenpakete. Die bevorzugte Behandlung ausgewählter Kommunikations-Anwendungen verschlechtert die Dienstgüte für die restlichen (schlechter priorisierten) Kommunikations-Anwendungen. Im einfachsten Fall kann die Priorisierung durch das Endgerät festgelegt werden. Die erforderlichen technischen Veränderungen in den Paket-Vermittlungsknoten sind dann nur gering, in lokalen Netzen wird eine derartige Priorisierung bereits angewendet (IEEE-Standard 802.1p).

Reservierungs-Mechanismen (resource reservation)
Die Einführung von Reservierungs-Mechanismen in Paket-Vermittlungsnetzen ermöglicht die Bereitstellung von Kommunikationsdiensten mit Dienstgüte-Garantie. Dies erfordert jedoch hohen technischen Aufwand im Netz. Das Resource Reservation Protocol (RSVP) der Internet-Protokollfamilie verfolgt diesen Ansatz.

Abgrenzung

Nachfolgend wird der Einfluss einer Priorisierung bei paketbasierter Sprachsignal-Übermittlung in einem paketbasierten Netz für Sprache und Daten untersucht. Es wird berechnet, welche Wartezeiten sich ohne bzw. mit Priorisierung in den Paket-Vermittlungsknoten bei typischen Verkehrs-Belastungen und Übertragungsgeschwindigkeiten ergeben. Dabei werden nur zwei Service-Klassen unterschieden:

1)	Priorität 1 (hohe Priorität)	für Sprach-Dialog-Kommunikation;
2)	Priorität 2 (niedrige Priorität)	für alle sonstigen Kommunikations-Anwendungen;

2 Dienstgüte bei Sprach-Dialog-Anwendungen

Vorbemerkung
Die Dienstgüte Q (Quality of Service, QoS) bewertet die Qualität einer ausgewählten Kommunikations-Dienstleistung eines Kommunikationsnetzes (für beispielsweise für Sprach-, Daten-, Bildkommunikation). Die Dienstgüte Q ist mathematisch betrachtet eine mehrdimensionale Größe:

$[Q] = [q_1, q_2, \ldots q_n];$

Jede Komponente q_i (mit i = 1, 2, … n) beschreibt die normierte Güte einer vereinbarten, wichtigen Teil-Eigenschaft. Bei $q_i = 0$ sei die Teil-Eigenschaft i überhaupt nicht erfüllt, bei $q_i = 1$ sei die Teil-Eigenschaft i vollständig erfüllt. Die q_i-Werte können zusätzlich durch einen Gewichts-Faktor g_i gewichtet werden:

$[Q_g] = [g_1 \cdot q_1, g_2 \cdot q_2, \ldots g_n \cdot q_n];$

Für optimale (also maximale) Dienstgüte muss der Betrag des Vektors [Q] bzw. (bei Gewichtung der Teil-Eigenschaften) der Betrag des Vektors $[Q_g]$) maximiert werden.

Wichtige Teil-Eigenschaften der Dienstgüte sind beispielsweise die Kommunikations-Bandbreite (bzw. der Daten-Durchsatz), der Signal-Geräusch-Abstand bzw. die Bitfehlerrate, die Verbindungsaufbauzeit (bei verbindungsorientierter Kommunikation), die Wiedergabe-Güte für die Kommunikationsinhalte (beispielsweise die Sprachsignal-Wiedergabegüte bei Sprachkommunikation). Eine besonders wichtige Teil-Eigenschaft der Dienstgüte eines Kommunikationsnetzes ist die Ende-zu-Ende-Verzögerungszeit. Nachfolgend wird diese Verzögerungszeit analysiert.

Während beispielsweise beim „Internet-Surfen“ eine Zweiweg-Verzögerungszeit (zwischen Absendung der Anforderung in das Netz und Empfangen der Antwort aus dem Netz) im Sekunden-Bereich akzeptabel sein kann, ist für eine Sprach-Dialog-Anwendung (Fernsprechen, Sprach-Dialog mit einem Rechner) nur eine Einweg-Verzögerungszeit (Ende-zu-Ende-Verzögerungszeit) von maximal 150 ms zulässig, ansonsten wird diese Verzögerung als störend empfunden.

Nachfolgend werden die Mindest-Anforderungen an die Sprachsignal-Wiedergabe-Qualität sowie die zulässigen Grenzwerte für die Ende-zu-Ende-Verzögerungszeit für die Sprachkommunikation zusammengestellt. Hierbei werden In-

halte aus [NOCK05, Kap.3] verwendet, ergänzende Literaturhinweise sind dort angegeben.

Bewertung der Sprachsignal-(Wiedergabe-)Qualität

Die Sprachsignal-Wiedergabe-Qualität (beispielsweise einer Fernsprech-Verbindung) kann mit dem sog. MOS-Wert bewertet werden [ITUP]. Dabei werden vielen, repräsentativ ausgewählten Versuchspersonen Sprachproben zur Bewertung vorgespielt. Jede Versuchsperson beurteilt die Verständlichkeit einer Sprachprobe nach der in Bild 2.1 angegebenen Punkte-Skala. Die durchschnittliche Bewertung aller Versuchspersonen (Mean Opinion Score, MOS) ergibt den resultierenden MOS-Wert für die Sprachprobe. Der MOS-Wert beschreibt nur die Qualität einer Sprachsignal-Wiedergabe.

MOS-Wert	Bedeutung	Beschreibung der Störwirkung
5	excellent	no effort required Störeffekte nicht wahrnehmbar
4	good	no appreciable effort required Störeffekte kaum wahrnehmbar
3	fair	moderate effort required Störeffekte wahrnehmbar, leicht störend
2	poor	considerable effort required Störeffekte störend
1	bad	no meaning understood with reasonable effort Störeffekte sehr störend, unangenehm

Bild 2.1:
Bewertung der Sprachsignal-Qualität nach ITU-Empfehlung P.800.
MOS Mean Opinion Score.

Verzögerungen sind bei einer Sprach-Dialog-Kommunikation sehr störend, wenn bestimmte Grenzwerte überschritten werden. Aus dem Sprach-Dialog wird dann ein „Wechsel-Gespräch“ (Halbduplex-Modus), wie dies beispielsweise aus Raumfahrt-Anwendungen bekannt ist. Die Verzögerungszeiten eines Kommuni-

kationsnetzes werden durch die MOS-Bewertung nicht erfasst. Zulässige Verzögerungszeiten bei Vollduplex-Sprachkommunikation müssen deshalb zusätzlich definiert werden.

Sprachsignal-Wiedergabe-Qualität existierender Vermittlungsnetze
Für die Sprachsignal-Wiedergabe-Qualität in existierenden Vermittlungsnetzen ergeben sich die in Bild 2.2 angegebenen MOS-Werte.

MOS-Wert	Vermittlungsnetz
4.4 … 4.5	ISDN-Netz PCM-Sprache (8 kHz Abtastfrequenz, 8 Bit/Abtastwert)
3.5 ... 4.0	Analog-Fernsprechnetz (Analogkanal 0.3 bis 3.4 kHz)
3.6	Digitales Mobilfunknetz (GSM-Standard)

Bild 2.2:
Sprachsignal-Wiedergabe-Qualität (MOS-Wert) bei existierenden Vermittlungsnetzen.

Zulässige Verzögerungszeiten bei Sprach-Dialog-Kommunikation
Für eine Echtzeit-Sprachkommunikation (wie beim „telefonieren“) sollen im Idealfall keine „wahrnehmbaren“ Verzögerungszeiten auftreten. Falls dies aus technischen oder physikalischen Gründen nicht realisierbar ist, sollen bestimmte Maximalwerte nicht überschritten werden. Diese Grenzwerte werden in der ITU-Empfehlung G.114 [ITUG] beschrieben.

Bild 2.3 zeigt die zulässige (Einweg-) Ende-zu-Ende-Verzögerungszeit (end to end delay, one way delay, t_{ee}) vom Mikrofon des Fernsprech-Teilnehmers A bis zum Hörer des Fernsprech-Teilnehmers B. Die Mindest-Pausendauer bei Dialog-Kommunikation (round trip delay, t_{rtd}) ist um mindestens den Faktor 2 größer als die hier betrachtete (Einweg-) Ende-zu-Ende-Verzögerungszeit.

$$t_{rtd} \geq 2 \cdot t_{ee}$$

Ende-zu-Ende-Verzögerungszeiten unter 25 ms sind nicht wahrnehmbar, größere Verzögerungszeiten sind wahrnehmbar. Verzögerungszeiten zwischen 25 ms und 150 ms gelten als akzeptabel, da noch nicht störend. Verzögerungszeiten zwischen 150 ms und 400 ms sind bedingt akzeptabel, da für empfindliche Benutzer

bereits (mehr oder weniger stark) störend. Höhere Ende-zu-Ende-Verzögerungszeiten als 400 ms sind inakzeptabel.

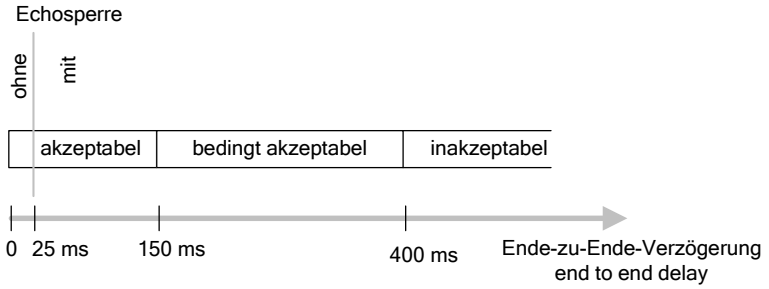


Bild 2.3:
Zulässige Ende-zu-Ende-Verzögerungszeiten für die Vollduplex-Sprachkommunikation.

Für $t_{ee} > 25$ ms wird der Einsatz einer Echosperre empfohlen, um Störungen des Gesprächsablaufs zu verhindern. Eine Echosperre erhöht die Dämpfung für die Übertragungsrichtung BA, wenn der A-Teilnehmer in der Übertragungsrichtung AB spricht (und umgekehrt). Bei kleinen Verzögerungszeiten kann anstatt einer Echosperre ein Echokompensations-Filter (echo cancellation) eingesetzt werden. Bei der Echokompensation wird ein Echo durch ein adaptives Digitalfilter (weitgehend) kompensiert .

Die Ende-zu-Ende-Verzögerungszeiten liegen für nationale Verbindungen im Digital-Fernsprechnetz bzw. ISDN-Netz unter 25 ms und sind somit vom Fernsprech-Teilnehmer nicht wahrnehmbar. Erst bei interkontinentalen Verbindungen über geostationäre Satelliten ergeben sich Ende-zu-Ende-Verzögerungszeiten bis maximal 400 ms. Damit dieser Grenzwert nicht überschritten wird, war in klassischen Fernsprechnetzen der „double hop“ über zwei Satelliten (beispielsweise Europa nach USA, USA nach Japan) nicht zulässig.

Erwünschte Eigenschaften zukünftiger Sprach-Vermittlungsnetze

Zukünftige Sprach-Vermittlungsnetze sollten bezüglich der Sprachsignal-Wiedergabe-Qualität im Idealfall ISDN-Qualität (MOS-Wert 4.4), mindestens aber die Qualität des früheren Analog-Fernsprechnetzes (MOS-Wert 3.5) aufweisen.

Als Ende-zu-Ende-Verzögerung (one way delay) für Sprach-Dialog-Kommunikation sind maximal 50 ms erwünscht, bis 150 ms gelten als „akzeptabel". Ende-zu-Ende-Verzögerungen größer als 150 ms sind nur „bedingt akzeptabel" (da für manche Teilnehmer bereits störend) und sollten möglichst vermieden werden, wenn eine hohe „Sprach-Dialog-Qualität" gewünscht wird.

Definition der Sprach-Dialog-Qualität

Nachfolgend wird eine „sehr gute" bzw. „ausreichende" Sprach-Dialog-Qualität eines Kommunikationsnetzes wie folgt durch einerseits Schranken für den MOS-Wert und andererseits Schranken für die Ende-zu-Ende-Verzögerungszeit (one way delay) definiert:

Sprach-Dialog-Qualität	MOS-Wert	Ende-zu-Ende-Verzögerung	Bemerkung
Sehr gut	MOS ≥ 4.4	$t_{ee} \leq 50$ ms;	Wie nationale ISDN-Verbindung!
Ausreichend	3.5≤ MOS <4.4	$t_{ee} \leq 150$ ms;	Wie internationale Analog-Verbindung!

Eine derart definierte Sprach-Dialog-Qualität berücksichtigt sowohl die Sprachsignal-Wiedergabe-Qualität als auch die (einen Sprach-Dialog) störende Durchlaufverzögerung eines Kommunikationsnetzes. Eine „sehr gute" Sprach-Dialog-Qualität gemäß obiger Definition entspricht etwa der Qualität einer nationalen ISDN-Fernsprech-Verbindung. Eine ausreichende Sprach-Dialog-Qualität gemäß obiger Definition entspricht etwa der Qualität einer internationalen Fernsprech-Verbindung im früheren Analog-Fernsprechnetz.

ETSI (European Telecommunications Standard Institute) definiert die Qualitätsklasse „best" für Sprachkommunikation durch eine Sprachsignal-Wiedergabe-Qualität wie bei Standard-Pulscodemodulation (nach ITU-Empfehlung G.711, entspricht ISDN-Sprachsignal-Qualität, also MOS ≥4.4) und eine Ende-zu-Ende-Verzögerungszeit von maximal 150 ms sowie eine Verbindungs-Aufbauzeit von maximal 4 s [SIEG02].

3 Ende-zu-Ende-Verzögerungen in Kommunikationsnetzen

Die wesentlichen Formeln zur Berechnung der Durchlaufverzögerung in Vermittlungsnetzen wurden in [NOCK05, Kap. 4.6] abgeleitet. Diese Grundlagen werden zunächst dargestellt, da sie zum Verständnis der nachfolgenden Ausführungen zwingend erforderlich sind. Anschließend werden die Minimalwerte, Mittelwerte und Maximalwerte der Ende-zu-Ende-Durchlaufverzögerung allgemein berechnet und Abschätzungen für verschiedene Sonderfälle vorgenommen.

3.1 Definition und Modellbildung

Die Ende-zu-Ende-Durchlaufverzögerung ist die Verzögerung eines Nachrichtensignals zwischen dem Ausgang der Nachrichtenquelle und dem Eingang der Nachrichtensenke unter Berücksichtigung aller Einflüsse.

Bei Sprachkommunikation wird der „elektrische“ Weg vom Mikrofon-Ausgang des Teilnehmers A (hier liegt erstmals ein elektrisches Signal vor) zum Hörer-Eingang des Teilnehmers B (hier liegt letztmalig ein elektrisches Signal vor) betrachtet. Bei Sprach-Kommunikation enthält die Ende-zu-Ende-Durchlaufverzögerung somit alle Teilverzögerungen zwischen Mikrofon-Ausgang des Endgeräts A und Hörer-Eingang des Endgeräts B.

Die Ende-zu-Ende-Durchlaufverzögerung (diese beinhaltet die Endgeräte-Verzögerungen) ist streng zu unterscheiden von der Durchlaufverzögerung des Vermittlungsnetzes (der Latenz des Vermittlungsnetzes, beinhaltet keine Endgeräte-Verzögerungen).

Bei jedem Übermittlungsvorgang ergibt sich eine unvermeidliche „physikalische“ Grund-Verzögerung. Diese ist durch die endliche Ausbreitungsgeschwindigkeit elektromagnetischer Signale bedingt und ist unabhängig vom verwendeten Vermittlungsprinzip.

Zu dieser Grund-Verzögerung addieren sich Verzögerungszeiten, welche abhängig vom verwendeten Vermittlungsprinzip sind. Weitere zusätzliche Verzögerungen können durch Signalverarbeitungs-Vorgänge (Quellencodierung, Filterung, Verschlüsselung) in den Endgeräten sowie softwarebasierte Verfahren in den Vermittlungsknoten auftreten, diese zusätzlichen Verzögerungen werden nachfolgend nicht berücksichtigt.

Physikalische Signal-Laufzeit

Die Ausbreitungsgeschwindigkeit elektromagnetischer Signale im Vakuum ist eine universelle Naturkonstante und beträgt rund $c_0 = 300\,000$ km/s. Für das Übertragungsmedium „Luft" kann mit dem gleichen Wert gerechnet werden. Dies ergibt eine Laufzeit pro Längeneinheit von 3.33 µs / km.

Die Ausbreitungsgeschwindigkeit hochfrequenter elektromagnetischer Signale auf elektrischen Kabelleitungen (mit Isoliermaterial der Dielektrizitätszahl ε_r und der Permeabilitätszahl μ_r) oder optischer Signale (dies sind elektromagnetische Signale mit extrem hoher Frequenz) in Lichtwellenleitern (mit dem Brechungsindex n des optischen Übertragungsmediums) ist:

$$v = \frac{c_0}{\sqrt{\varepsilon_r \cdot \mu_r}} = \frac{c_0}{n}; \qquad \text{mit} \qquad c_0 \approx 300\,000 \text{ km/s};$$

Für kunststoffisolierte elektrische Kabelleitungen (beispielsweise Polyethylen-Isolierung mit $\varepsilon_r = 2.3$ und $\mu_r = 1$) oder Lichtwellenleiter (beispielsweise Glasfasern mit $n \approx 1.5$) folgt eine Ausbreitungsgeschwindigkeit von rund $v = 200\,000$ km/s. Dies ergibt eine Laufzeit pro Längeneinheit von rund 5 µs/km.

Übertragungs-Overhead bei Datenpaket-Übermittlung

Bei der Bildung eines Datenpakets wird ein Nutzbit-Block mit n_n Bit durch n_z zusätzliche Bit (einen vorlaufenden „Header" und manchmal zusätzlich einen nachlaufenden „Trailer") zu einem Datenpaket mit $n_{ges} = n_n + n_z$ Bit ergänzt. Im Vermittlungsnetz entsteht dadurch ein Bitraten-Mehrbedarf, welcher als „Protokoll-Overhead" bezeichnet wird. Für eine Nachrichtenquelle mit der (Netto-) Übertragungsgeschwindigkeit EF_a folgt deshalb eine resultierende (Brutto-) Übertragungsgeschwindigkeit EF_{ges} im Vermittlungsnetz von:

$$EF_{ges} = EF_a \cdot \frac{n_{ges}}{n_n} = EF_a \cdot \left(1 + \frac{n_z}{n_n}\right);$$

Für die Sprachdaten-Übermittlung in IP-Netzen (VoIP-Technik) werden die Protokolle RTP (Realtime Transport Protocol), UDP (User Datagram Protocol) und IP (Internet Protocol) verwendet. Als Subnetz-Technik wird nachfolgend stets Switched-Ethernet-Technik vorausgesetzt.

Der resultierende Umfang eines Datenpakets beinhaltet dann 58 zusätzliche Bytes, bestehend aus 14 Byte Ethernet-Header und 4 Byte Ethernet-Trailer,

20 Byte IP-Header, 8 Byte UDP-Header und 12 Byte RTP-Header. Nachfolgend werden diese Zahlenwerte für die numerischen Berechnungen verwendet.

Die Technik von VoIP-Systemen wird beispielsweise in [BADA07, NOEL03] ausführlich dargestellt, zu Verzögerungen in paketbasierten Netzen wird ergänzend auf [CISCO] verwiesen. Die Grundlagen von Durchschalte- und Paketvermittlungsnetzen werden in [NOCK05, SIEG02, SIEG02b] dargestellt.

Modell zur Berechnung der Ende-zu-Ende-Durchlaufverzögerung

Nachfolgend wird der Weg einer Nachricht von einem Start-Knoten (dem paketbasierten Endgerät A) über einen oder mehrere Paket-Vermittlungsknoten zum Ziel-Knoten (dem paketbasierten Endgerät B) betrachtet.

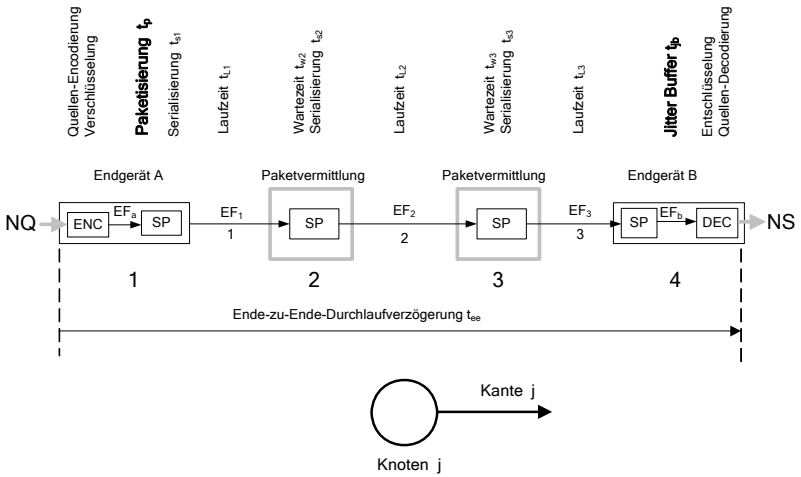


Bild 3.1:
Ende-zu-Ende-Durchlaufverzögerung in einem Paket-Vermittlungsnetz mit 4 Knoten.
EF Übertragungsgeschwindigkeit (Bitrate); NQ Nachrichtenquelle; NS Nachrichtensenke; ENC Encoder; DEC Decoder; SP Speicher; Rechenzeiten und „sonstige Verzögerungszeiten" sind nicht eingetragen.

Ein Weg durch ein Telekommunikationsnetz ist ein linienförmiger Graph mit mindestens 3 Knoten (Endgerät A, Vermittlungs-Knoten, Endgerät B) und mindestens 2 Kanten. Im allgemeinen Fall besteht der Wege-Graph aus $N \geq 3$ Kno-

ten (hiervon 2 Endgeräte, N-2 Vermittlungs-Knoten) und N-1 Kanten. Vom Start-Knoten (Endgerät A) bis zum Ziel-Knoten (Endgerät B) werden die Knoten mit 1, 2, ... N nummeriert, entsprechend werden die Kanten (Übertragungswege) mit 1, 2, ... N-1 nummeriert.

Bild 3.1 zeigt einen Weg mit N = 4 Knoten (2 Endknoten, 2 Vermittlungsknoten) und N-1 = 3 Kanten, welche entsprechend obiger Beschreibung nummeriert sind. Die Ende-zu-Ende-Durchlaufverzögerung t_{ee} (end to end) vom Eingang des paketbasierten Endgeräts A zum Ausgang des paketbasierten Endgeräts B ist eingetragen.

Da verbindungsorientierter Netzbetrieb vorausgesetzt wird, durchläuft jedes Datenpaket einer Kommunikationsbeziehung dieselben Knoten. Empfangsseitig liegt somit immer auch die korrekte Datenpaket-Reihenfolge vor. Zunächst wird vorausgesetzt, dass alle Datenpakete gleiche Priorität aufweisen.

Hinweis:
Im „Internet“ wird die datagram-Betriebsweise (das Internet-Protokoll IP ist ein verbindungsloses Protokoll für die Netzwerkschicht, auch als OSI-Layer 3 oder als Vermittlungsschicht bezeichnet) verwendet. Dabei kann prinzipiell jedes Datenpaket einer Kommunikationsbeziehung einen anderen Weg durch das Netz benutzen. Im Normalfall (stabile Verhältnisse beim Netzbetrieb vorausgesetzt) wird jedoch auch bei datagram-Betriebsweise fast immer derselbe Weg durch das Kommunikationsnetz benutzt.

Die nachfolgenden Betrachtungen gelten zwar exakt nur für verbindungsorientierte Netzwerk-Protokolle, können jedoch (stabile und stationäre Verhältnisse beim Netzbetrieb vorausgesetzt) mit guter Näherung auch auf Kommunikationsnetze mit verbindungslosen Netzwerk-Protokollen (beispielsweise IP-Protokoll) angewendet werden.

3.2 Berechnung der Ende-zu-Ende-Durchlaufverzögerung ohne Jitter-Buffer

Allgemeine Formulierung

Die Ende-zu-Ende-Durchlaufverzögerung t_{ee} eines Datenpakets durch ein Paket-Vermittlungsnetz mit verbindungsorientierter Betriebsweise setzt sich additiv aus den Teil-Verzögerungszeiten aller Knoten und Kanten des Weges von A nach B zusammen (siehe Bild 4.17) und ergibt sich zu:

$$t_{ee} = \sum_{(j)} t_{v,j} = \sum_{(j)} \left(t_{v,node,j} + t_{v,edge,j} \right);$$

Dabei ist $t_{v,node,j}$ bzw. $t_{v,edge,j}$ die Teilverzögerung durch Knoten j bzw. Kante j. Die Summation ist über alle relevanten Werte von j durchzuführen. Schlüsselt man die Teilwerte entsprechend ihrer technischen Ursache auf, erhält man:

$$t_{ee} = t_{ums,ab} + t_p + t_{jb} + \sum_{j=1}^{N-1} (t_{sj} + t_{wj} + t_{rj} + t_{Lj}) \; ;$$

Bedeutung der Formelzeichen

$t_{ums,ab}$ **Umsetzverzögerung** (Quellencodierung, Coding Delay) im Knoten 1 und N (Endgerät A und Endgerät B).

t_p Paketisierungszeit im Knoten 1, dem Endgerät A.

t_{jb} **Jitterbuffer-Verzögerungszeit** im Knoten N, dem Endgerät B (ist nur für isochrone Signale nötig, beispielsweise PCM-Sprache).

t_{sj} Serialisierungszeit im Knoten j;

t_{wj} **Wartezeit** im Knoten j;

t_{rj} Rechenzeit oder sonstige Verzögerungszeit im Knoten j;

l_j Länge der Kante j;

v_j Ausbreitungsgeschwindigkeit für die Kante j;

t_{Lj} **Laufzeit** für die Kante j: $t_{Lj} = l_j / v_j$;

EF_j Übertragungsgeschwindigkeit (Bitrate) für die Kante j;

EF_a Übertragungsgeschwindigkeit der Nachrichtenquelle im Endgerät A;

EF_b $= EF_a$;
Übertragungsgeschwindigkeit der Nachrichtensenke im Endgerät B;

Hinweis zur Jitterbuffer-Verzögerung

Die Jitterbuffer-Verzögerungszeit t_{jb} im Endgerät B (bei VoIP-Sprachkommunikation erforderlich, nicht bei der Datenübertragung) wird in Kapitel 3.3 diskutiert und zunächst nicht weiter berücksichtigt.

Hinweis zur Umsetz-Verzögerung

Die Umsetz-Verzögerung $t_{ums,ab}$ ist die Summe der Zeitdauern für die Quellen-Encodierung im Endgerät A und die zugehörige Quellen-Decodierung im Endgerät B. Nachfolgend wird hierfür Standard-Pulscodemodulation vorausgesetzt, dann ist die Umsetzverzögerung (mit Abtastperiodendauer t_a = 125 µs) $t_{ums,ab} = 2 \cdot t_a$ = 0.25 ms. Dieser kleine Teilwert wird nachfolgend gegenüber der rund 100-fach größeren Paketisierungszeit t_p vernachlässigt. Für andere Quellencodierungsverfahren wird in Kapitel 5 eine Tabelle der zugehörigen Umsetz-Verzögerungszeiten (Coding Delay durch Quellencodierung) angegeben.

Berechnung der Teil-Verzögerungszeiten

Die Anzahl der Nutzbit pro Datenpaket und die Gesamtanzahl der Bit pro Datenpaket sei:

n_n Anzahl der Nutzbit pro Datenpaket;

n_{ges} (Gesamt-)Anzahl der Bit pro Datenpaket (Header plus Nutzbit plus ggf. Trailer);

Damit folgt:

$t_p = n_n / EF_a$;

$t_{sj} = n_{ges} / EF_j$;

t_{wj} = **Zufallswert ≥ 0**; abhängig von Verkehrs-Belastung der Kante j und der Abfertigungs-Strategie im Knoten j, wird später berechnet;

t_{rj} = deterministischer Wert, abhängig vom Rechen-Aufwand im Knoten j für das verwendete Übermittlungs-Protokoll;

$t_{Lj} = l_j / v_j = l_j / (200\,000$ km/s$) = l_j \cdot (5$ µs/km); (Leitungs-Übertragung)

$= l_j / (300\,000$ km/s$) = l_j \cdot (3.33$ µs/km); (Funk-Übertragung)

Physikalische Grund-Laufzeit

Wenn teilweise leitungsgebundene Übertragung und teilweise Funk-Übertragung vorliegt, müssen die unterschiedlichen Ausbreitungsgeschwindigkeiten für die jeweiligen Kanten berücksichtigt werden. Wenn einheitlich leitungsgebundene (elektrische oder optische) Übertragung oder einheitlich Funk-Übertragung vorliegt, ist die Ausbreitungsgeschwindigkeit für alle Kanten gleich groß und wird nachfolgend mit v (ohne Indizierung) bezeichnet.

Die Gesamt-Laufzeit t_{Lges} ist dann:

$$t_{Lges} = \sum_{j=1}^{N-1} t_{Lj} = \sum_{j=1}^{N-1} \frac{l_j}{v_j} = \frac{l_{ges}}{v} = \begin{cases} l_{ges} \cdot [5\,\mu s/km] & \text{(Leitung)}; \\ l_{ges} \cdot [3.33\,\mu s/km] & \text{(Funk)}; \end{cases}$$

Ende-zu-Ende-Durchlaufverzögerung ohne Jitter Buffer

Die Ende-zu-Ende-Durchlaufverzögerung eines Datenpakets ergibt sich ohne empfangsseitigen Jitter Buffer zu:

$$t_{ee,PV} = \frac{l_{ges}}{v} + t_p + \sum_{j=1}^{N-1}(t_{sj} + t_{wj} + t_{rj}) \; ;$$

Nach Berechnung der Summenwerte folgt:

$$t_{ee,PV} = \left(\frac{l_{ges}}{v} + t_p + t_{s,ges} + t_{r,ges} \right) + t_{w,ges} \; ;$$

Die Ende-zu-Ende-Durchlaufverzögerung eines Datenpakets durch ein Paket-Vermittlungsnetz besteht somit aus zwei Anteilen, einem deterministischen Anteil (runde Klammer) und einem stochastischen Anteil (die Gesamt-Wartezeit $t_{w,ges}$ ist eine Zufallsgröße ≥ 0).

Der deterministische Anteil besteht aus der physikalischen Gesamt-Laufzeit l_{ges}/v, der sendeseitigen Paketisierungszeit t_p, der Gesamt-Serialisierungszeit $t_{s,ges}$ und der Gesamt-Rechenzeit $t_{r,ges}$.

Nachfolgend werden bei numerischen Berechnungen die Rechenzeiten $t_{r,ges}$ nicht weiter berücksichtigt. Einerseits sind sie produktabhängig (firmenspezifisch), andererseits können sie zukünftig durch eine weiter steigende Rechenleistung in den Netzknoten „beliebig“ reduziert werden.

Die Gesamt-Serialisierungsdauer $t_{s,ges}$ ist die Summe der individuellen Serialisierungsdauern in den Netzknoten 1 bis N-1.

$$t_{s,ges} = \sum_{j=1}^{N-1} t_{sj} \; ;$$

Die oben angegebene Ende-zu-Ende-Durchlaufverzögerung enthält noch nicht die empfangsseitige, zusätzliche Verzögerung durch den Jitter-Buffer im Endge-

rät B, welcher zum Ausgleich der von Sprachpaket zu Sprachpaket unterschiedlichen Verzögerungswerte eingesetzt werden muss. Dies wird im nachfolgenden Teilabschnitt genauer betrachtet.

3.3 Jitter-Buffer-Dimensionierung für paketbasierte Vollduplex-Sprachkommunikation

Funktionsweise des „Jitter Buffer“

Bei Vollduplex-Sprachkommunikation ist die quellenseitige Übertragungsgeschwindigkeit EF_a und die senkenseitige Übertragungsgeschwindigkeit EF_b identisch. Für alle nachfolgenden Überlegungen zur Sprachkommunikation gilt deshalb immer $EF_b = EF_a$.

Ein „Jitter Buffer“ ist eine Empfangsspeicher-Erweiterung zum Ausgleich schwankender Datenpaket-Durchlaufzeiten durch das Netz. Dieser Ausgleich ist erforderlich, wenn **isochrone Digitalsignale** (beispielsweise PCM-codierte Sprachsignale) über ein paketbasiertes Vermittlungsnetz übertragen werden. Vom Netz werden die Datenpakete in den Empfangsspeicher mit hoher Übertragungsgeschwindigkeit **burstartig eingelesen**, zur Nachrichtensenke muss der Inhalt des Empfangsspeichers mit niedriger Übertragungsgeschwindigkeit **kontinuierlich und somit unterbrechungsfrei ausgelesen** werden.

Um trotz schwankender Datenpaket-Durchlaufzeiten einen kontinuierlichen, unterbrechungsfreien Auslesevorgang sicherstellen zu können, muss der Start des Auslesevorgangs verzögert werden. Die (auf die Paketisierungsdauer t_p normierte) Verzögerung t_{jb} des Auslesebeginns gegenüber dem Empfangszeitpunkt des ersten Datenpakets wird nachfolgend als „Vorlauf β“ bezeichnet. Mit steigendem Vorlauf verbessert sich der kontinuierliche Datenfluss zur Nachrichtensenke, die unerwünschte Verzögerungszeit steigt aber an.

Bei optimaler Dimensionierung kann ein vorzeitiger Speicher-Leerlauf (Speicher-Unterlauf, buffer underrun) oder ein Speicher-Überlauf gerade noch nicht auftreten. Die Verzögerung des Sprachsignals und die erforderliche Empfangsspeicher-Kapazität sind dann minimal, gleichzeitig kann der Auslesevorgang soeben noch kontinuierlich und unterbrechungsfrei erfolgen.

Dimensionierung des Jitter-Buffers

In [NOCK05] wurde die Auslegung eines Jitter-Buffers allgemein behandelt. Für eine genaue Ableitung hierzu wird darauf verwiesen. Die dabei auftretenden Größen sind:

$BUF = n / n_n =$	normierter Füllungsstand des Jitter-Buffer-Speichers;
$\beta = t_{jb} / t_p =$	normierte Vorlaufzeit;
$\varepsilon(k) = t_{w,ges}(k) / t_p =$	normierte Gesamt-Wartezeit für das k-te Paket;
$\varepsilon_{max} = [\ \varepsilon(k)\ ]_{max} =$	Maximalwert aller $\varepsilon(k)$-Werte;

Wegen $\Delta\varepsilon := \varepsilon_{max} - \varepsilon_{min} = \varepsilon_{max} - 0 = \varepsilon_{max}$ ist der Maximalwert ε_{max} identisch mit der normierten Schwankungsbreite $\Delta\varepsilon$ der möglichen Wartezeiten. Um einen Leerlauf des Jitter-Buffer-Speichers zu vermeiden, muss ein Mindest-Vorlauf $\beta_{min} = \varepsilon_{max}$ verwendet werden:

$$\beta \geq \beta_{min} = \varepsilon_{max}$$

Wird ein Vorlauf $\beta > \beta_{min}$ verwendet, verbleibt der Empfangsspeicher-Füllungsstand (wie erwünscht) immer im positiven Bereich. Als maximaler Empfangsspeicher-Füllungsstand ergibt sich dann $(1+\varepsilon_{max}+\beta)$. Für den Fall $\beta > \beta_{min} = \varepsilon_{max} \leq 1$ kann dann weder ein Speicher-Leerlauf noch ein Speicher-Überlauf eintreten. Als allgemeine Vorschrift zur Auslegung des Jitter-Buffers ergibt sich (ausführliche Ableitung siehe [NOCK05]):

$$\beta > \beta_{min} = \varepsilon_{max} \leq 1; \qquad \Rightarrow \qquad 0 < BUF \leq 1 + \varepsilon_{max} + \beta;$$

Für den (extremen) Fall $\varepsilon_{max} = 1$ ist eine normierte Empfangsspeicher-Kapazität $(2 + \beta) \geq (2+1) = 3$ stets ausreichend.

3.4 Berechnung der Ende-zu-Ende-Durchlaufverzögerung mit Jitter Buffer

Das Datenpaket k (mit $k = 1, 2, \ldots$) wird vom Startknoten (Knoten 1, Endgerät A) zum Zeitpunkt $k \cdot t_p$ abgesendet. Beim Durchlauf durch das Vermittlungsnetz erfährt das Datenpaket k die konstante Verzögerung $const = (l_{ges}/v + t_{s,ges} + t_{r,ges})$ und zusätzlich die zufällige Gesamt-Wartezeit $t_{w,ges}(k)$. Somit wird das Datenpaket k zum Zeitpunkt $const + k \cdot t_p + t_{w,ges}(k)$ in den Empfangsspeicher des Zielknotens (Knoten N, Endgerät B) eingelesen.

Bei einem Vorlauf β startet der empfangsseitige Auslesevorgang um $\beta \cdot t_p$ verzögert nach dem tatsächlichen Eintreff-Zeitpunkt des ersten Datenpakets, also zum Zeitpunkt $[const + 1 \cdot t_p + t_{w,ges}(1)] + \beta \cdot t_p$. Diese Verzögerung bestimmt die Ende-zu-Ende-Durchlaufverzögerung des gesamten Sprachsignals. Somit

folgt (nachfolgend ist noch einmal die Umsetz-Verzögerung mit aufgeführt) als Ergebnis:

$$t_{ee,PV} = t_{ums,ab} + \left(\frac{l_{ges}}{v} + t_{s,ges} + t_{r,ges} \right) + t_p \cdot (1+\beta) + t_{w,ges}(1);$$

Die Größe $t_{w,ges}(1)$ ist die Gesamt-Wartezeit für das erste Nutzdatenpaket beim Durchlauf durch das Netz. Der Term $t_p \cdot (1 + \beta)$ in obiger Gleichung beinhaltet die sendeseitige Paketisierung $(1 \cdot t_p)$ und die empfangsseitige Verzögerung im Jitter Buffer $(t_{jb} := \beta \cdot t_p)$.

Die Umsetzverzögerung wird nachfolgend nicht weiter berücksichtigt, da bei der nachfolgend unterstellten Standard-PCM der zugehörige Teilwert sehr klein ist, siehe hierzu den Hinweis in Kapitel 3.2.

3.5 Minimalwert der Ende-zu-Ende-Durchlaufverzögerung (best case)

Bei einem leerlaufenden Netz (mit der Verkehrs-Auslastung 0) treten keine Wartezeiten auf. Es kann dann $t_{w,ges}(1) = 0$ gesetzt werden. Damit ergibt sich als Mindest-Durchlaufverzögerung:

$$t_{ee,PV,min} = \left(\frac{l_{ges}}{v} + t_{s,ges} + t_{r,ges} \right) + \left[t_p \cdot (1+\beta) \right];$$

Der erste Anteil (in runden Klammern) ist die unvermeidliche physikalische Laufzeit, diese kann durch technische Maßnahmen nicht reduziert werden. Der zweite Anteil (in eckigen Klammern) ist durch das Paket-Vermittlungsprinzip bedingt. Der Summenwert ist eine untere Schranke für die Ende-zu-Ende-Durchlaufverzögerung, wird hier aber vereinfacht als Minimalwert bezeichnet. Real ist die Ende-zu-Ende-Durchlaufverzögerung immer höher als der durch diese Formel berechenbare Wert.

Für Datenübertragung entfällt der Jitter Buffer, dann ist $\beta = 0$ zu setzen. Für Vollduplex-Sprachübertragung ist ein Jitter Buffer und somit $\beta > 0$ zwingend erforderlich. Der theoretisch zulässige Mindestwert für den Vorlauf β ist $\beta_{min} = \varepsilon_{max} = \max(t_{w,ges} / t_p)$. Für Abschätzungen wird nachfolgend mit $\beta = 0.5$ oder $\beta = 1$ gerechnet.

3.6 Mittelwert der Ende-zu-Ende-Durchlaufverzögerung (average case)

Als Erwartungswert (für viele unterschiedliche Sprachverbindungen!) der Ende-zu-Ende-Durchlaufverzögerung bei paketbasierter Übermittlung mit empfangsseitigem Jitter Buffer und einem Vorlauf von β Paketisierungsdauern ergibt sich mit den Abkürzungen

$EW[t_{ee,PV}] := t_{ee,PV,avg}$; $EW[t_{w,ges}(1)] := t_{w,ges,avg}$;

nachfolgendes Ergebnis für die Ende-zu-Ende-Durchlaufverzögerung:

$$t_{ee,PV,avg} = \left(\frac{l_{ges}}{v} + t_{s,ges} + t_{r,ges} \right) + t_p \cdot (1+\beta) + t_{w,ges,avg} ;$$

Die Größe $t_{w,ges,avg}$ ist die **mittlere** Gesamt-Wartezeit des ersten Nutzdatenpakets, wobei die Mittelung über viele unterschiedliche Sprachverbindungen (mit gleichem Weg durch das Netz) durchgeführt wird. Bei stationären Bedingungen im Netz entspricht dies der mittleren Gesamt-Wartezeit für beliebige Nutzdatenpakete (mit gleichem Weg durch das Netz).

Die mittlere Gesamt-Wartezeit $t_{w,ges,avg}$ ergibt sich unter der Annahme (und nur dann), dass die Wartezeiten in den Vermittlungsknoten voneinander statistisch unabhängig sind, durch Addition der Teil-Wartezeiten in jedem Knoten.

$$t_{w,ges,avg} = \sum_{j=1}^{N-1} t_{wj,avg} ;$$

Hinweis:
In obiger Formel startet der Summations-Index bei 1, weil beispielsweise bei Multimedia-Kommunikation auch im Multimedia-Endgerät A eine Wartesituation vorliegen kann, ebenso wie im letzten Vermittlungsknoten (Knoten N-1) vor der „Anschlussleitung“ zum Multimedia-Endgerät B. Bei reiner Sprachkommunikation entfallen diese Wartesituationen in den Knoten 1 und N-1 und die entsprechenden Teil-Wartezeiten sind dann mit 0 anzusetzen.

Im allgemeinen wird die mittlere Wartedauer $t_{wj,avg}$ in jedem Vermittlungsknoten unterschiedlich sein, weil die Randbedingungen für jeden Vermittlungsknoten unterschiedlich sein werden. Für vereinfachte Planungs-Berechnungen kann mit gleichen Randbedingungen für jeden Vermittlungsknoten gerechnet werden. Auf die dann möglichen Vereinfachungen wird nachfolgend noch eingegangen.

3.7 Maximalwert der Ende-zu-Ende-Durchlaufverzögerung (worst case)

Im ungünstigsten Fall muss ein Datenpaket maximal lange warten:

$$t_{ee,PV,max} = \left(\frac{l_{ges}}{v} + t_{s,ges} + t_{r,ges} \right) + t_p \cdot (1+\beta) + t_{w,ges,max} ;$$

Die maximale Gesamt-Wartezeit ergibt sich durch Addition der Teil-Maximalwerte für jeden Knoten:

$$t_{w,ges,max} = \sum_{j=1}^{N-1} t_{wj,max} ;$$

Der Hinweis aus Kapitel 3.6 gilt für diesen Fall entsprechend.

3.8 Vereinfachungen für Sonderfälle

Bei gleicher Übertragungsgeschwindigkeit auf allen Netzkanten ist (mit t_s als Serialisierungsdauer pro Serialiserungsvorgang):

$t_{s,ges} = (N\text{-}1) \cdot t_s$;

Bei gleichen Randbedingungen in allen Vermittlungsknoten (gleiche Übertragungsgeschwindigkeit der Netzkanten, gleiche Belastungswerte) folgt mit $t_{w,avg}$ als mittlere Wartezeit pro Vermittlungsknoten und N_W als Anzahl der „Netzknoten mit Wartesituation“ folgende vereinfachte Beziehung:

$t_{w,ges,avg} = N_W \cdot t_{w,avg}$;

Hinweise zu obigen Vereinfachungen:

- N ist die Anzahl aller Knoten, also der Endgeräte (Endknoten) und der Vermittlungsknoten (Transitknoten, Vermittlungsknoten), siehe Bild 3.1. Eine Serialisierung der Daten erfolgt im Endgerät A (Knoten 1) und in allen $N_{PVK} = (N\text{-}2)$ Vermittlungsknoten. Resultierend ergeben sich somit immer $(N\text{-}1) = (N_{PVK}+1)$ Serialisierungs-Vorgänge, wie oben angegeben.
- Die Anzahl der „Netzknoten mit Wartesituation“ N_W ist abhängig von den Randbedingungen. Bei Multimedia-Kommunikation liegt auch im Multimedia-Endgerät A und im letzten Vermittlungsknoten eine Wartesituation vor, weil dann mehrere „Medienströme“ konkurrierend (quasi parallel) übertragen werden. Unter diesen Randbedingungen ergibt sich $N_W = (N\text{-}1)$.

- Nach Bild 3.1 sind die Endgeräte A und B über individuell genutzte „Anschlussleitungen“ an die zugehörigen Transitknoten angeschaltet. Wird reine Sprachkommunikation zwischen dem Sprach-Endgerät A und dem Sprach-Endgerät B betrachtet, dann liegt weder im Endgerät A noch im letzten Vermittlungsknoten eine Wartesituation vor. In diesem Fall ist die Anzahl der Netzknoten mit Wartesituation $N_W = (N-3)$. Die Durchlaufverzögerung wird dann etwas kleiner als bei Multimedia-Kommunikation.
- Im vereinfachten Netz-Modell in Kapitel 5 (siehe Bild 5.1) wird reine Sprachkommunikation zwischen zwei Sprach-Endgeräten A und B betrachtet, welche über individuell genutzte „Anschlussleitungen“ an die Paketvermittlungsknoten angeschaltet sind. Es ist dann wie oben beschrieben: $N_W = (N-3) = (N_{PVK}-1)$;

3.9 Maximal-Durchlaufverzögerung beim Durchschalte-Vermittlungsprinzip

Die abgeleitete Formel für die Ende-zu-Ende-Durchlaufverzögerung beim Paket-Vermittlungsprinzip kann verwendet werden, um die Ende-zu-Ende-Durchlaufverzögerung beim Durchschalte-Vermittlungsprinzip nach oben abzuschätzen.

Bei einem Durchschalte-Vermittlungsnetz entfallen die Paketisierungs-Zeit im Startknoten, alle Serialisierungs-Zeiten, alle Wartezeiten und die empfangsseitige Verzögerungszeit durch den Jitter Buffer.

Bei Zeitmultiplex-Durchschaltevermittlung beträgt die Durchschalte-Verzögerungzeit t_d (vorwiegend verursacht durch die Zeitstufen des Zeitmultiplex-Koppelnetzes) in jeder Durchschaltevermittlung nur wenige Pulsrahmen-Dauern T_r mit $T_r = T_a$ (Pulsrahmendauer T_r gleich Abtast-Periodendauer T_a). Bei Standard-Pulscodemodulation ist $f_a = 8$ kHz und somit $T_a = 125$ µs. Die Durchschalte-Verzögerung im Vermittlungsknoten j (mit j = 2, 3, ... N-1) ist damit $t_{dj} = n_j \cdot T_a$ mit $T_a = 125$ µs.

Wird für eine **grobe Abschätzung nach oben** mit maximal **$n_{max} = 8$ Rahmendauern** für jeden Vermittlungsknoten gerechnet, ergibt sich die maximale Durchschalte-Verzögerungszeit in jedem Durchschalte-Vermittlungsknoten zu

$$t_{d,max} = n_{max} \cdot T_a = 8 \cdot 125\ \mu s = 1\ ms;$$

Somit folgt mit $t_{d,max} = 1$ ms (die Durchschalte-Verzögerung kann als „sonstige Verzögerungszeit“ $t_{rj} = t_{d,max} = 1$ ms eines Durchschalte-Vermittlungsknotens betrachtet werden) folgende Abschätzung nach oben (nachfolgend vereinfacht

als Maximalwert bezeichnet) für die Ende-zu-Ende-Durchlaufverzögerung beim Durchschalte-Vermittlungsprinzip (mit $N \geq 3$):

$$t_{ee,DV,max} = \left(\frac{l_{ges}}{v} \right) + \left[(N-2) \cdot t_{d,max} \right];$$

mit $t_{d,max} \leq 1$ ms; $N \geq 3$; $(N\text{-}2) = N_{DVK} \geq 1$;

$(N\text{-}2) = N_{DVK} \geq 1$ ist die Anzahl der durchlaufenen Durchschalte-Vermittlungsknoten. Der erste Anteil in runden Klammern ist die unvermeidliche physikalische Grund-Laufzeit, diese kann durch technische Maßnahmen nicht verringert werden kann. Der zweite Anteil in eckigen Klammern ist durch das Durchschalte-Vermittlungsprinzip bedingt.

Der zugehörige Zahlenwert ist selbst für $N = 12$ (also 10 Vermittlungsknoten, eine ungünstige Annahme) mit maximal 10 ms sehr klein. Bei $l_{ges} = 1000$ km (siehe Eingangs-Parameter für die in Kapitel 5 berechneten Modell-Netze) folgt eine Gesamt-Laufzeit von 5 ms, die Ende-zu-Ende-Durchlaufverzögerung bei 10 Durchschalte-Vermittlungsknoten wird dann maximal $(5+10 \cdot 1)$ ms = 15 ms. Diese sehr geringe konstante Verzögerung kann von den Fernsprech-Teilnehmern nicht wahrgenommen werden.

3.10 Kettenschaltung unterschiedlicher Teilnetze

Besonders ungünstig ist eine Kettenschaltung mehrerer Teilnetze mit unterschiedlicher Technik und großer Gesamt-Entfernung AB. Zu den großen Laufzeiten (wegen großer Entfernung) addieren sich dann die Verzögerungen durch vielfache Vermittlungsvorgänge und mehrfache Nutzdaten-Umsetzvorgänge an den Schnittstellen zwischen Teilnetzen mit unterschiedlicher Technik.

Beispiel:
Es wird eine interkontinentale Fernsprech-Verbindung betrachtet, welche folgende Teilnetze durchläuft:

A) Teilnetz A:
Paketbasierte Verbindung in einer VoIP-Nebenstellenanlage
(typische Verzögerung ≥ 30 ms);

B) Teilnetz B:
Durchschalte-Verbindung im nationalen ISDN-Netz
(typische Verzögerung ≥ 10 ms);

C) Teilnetz C:
Übertragung über einen geostationären Satelliten
(typische Verzögerung ≥ 330 ms bei rund 100 000 km Funkstrecke insgesamt);

D) Teilnetz D:
Paketbasierte Fernsprechverbindung im nationalen Netz
(typische Verzögerung ≥ 50 ms);

E) Teilnetz E:
Durchschalte-Verbindung in einem Mobilfunknetz nach GSM-Standard
(typische Verzögerung ≥ 20 ms wegen GSM-Rahmendauer 20 ms);

Die resultierende Ende-zu-Ende-Durchlaufverzögerung beträgt dann mindestens 440 ms. Werden zusätzlich in der VoIP-Nebenstellenanlage (Teilnetz A) schnurlose Endgeräte mit DECT-Technik verwendet, addieren sich zusätzliche Verzögerungen durch die DECT-Funktechnik (typische Verzögerung ≥ 10 ms wegen DECT-Rahmendauer 10 ms). Die resultierende Ende-zu-Ende-Durchlaufverzögerung steigt dann auf mindestens 450 ms. Ende-zu-Ende-Durchlaufverzögerungen > 400 ms sind nach ITU-Empfehlung G.114 absolut unzulässig.

3.11 Zusammenfassung

Unabhängig vom verwendeten Vermittlungsprinzip ergibt sich abhängig von der Entfernung im Netz eine Grundverzögerung durch die endliche Ausbreitungsgeschwindigkeit elektromagnetischer Signale. Diese physikalische Laufzeit ist unvermeidbar und unabhängig von der im Netz verwendeten Vermittlungstechnik. Hinzu kommen zusätzliche Verzögerungszeiten, welche abhängig vom verwendeten Vermittlungsprinzip sind:

In einem Durchschalte-Vermittlungsnetz treten keine nennenswerten zusätzlichen Verzögerungszeiten auf. Die bei Zeitmultiplex-Koppelnetzen in Zeitstufen auftretende Durchschalte-Verzögerung liegt in der Größenordnung von wenigen Abtastperiodendauern und ist bei Sprachkommunikation nicht wahrnehmbar.

In einem Paket-Vermittlungsnetz treten immer zusätzliche Verzögerungszeiten auf. Diese sind durch die Paketisierung, die Serialisierung, die Wartezeiten beim Vermittlungsvorgang und den empfangsseitigen Jitter-Buffer bedingt. Der Jitter-Buffer kann bei Datenübertragung entfallen. Die Ende-zu-Ende-Durchlaufverzögerung für Sprachsignale wird weitgehend durch die quellenseitige Paketisierungs-Verzögerung und den senkenseitigen Jitter Buffer bestimmt. Hinzu kom-

men zufällige Wartezeiten in den Paket-Vermittlungsknoten, welche in Kap. 4 berechnet werden.

4 Wartezeit-Verzögerungen in Paket-Vermittlungsknoten

4.1 Vorbemerkungen

Als „Paket-Vermittlungsknoten“ wird nachfolgend jeder „Netzknoten“ bezeichnet, welcher nach dem „store and forward“-Vermittlungsprinzip arbeitet. Dies können „Switches“ (solche arbeiten mit Schicht-2-Adressen, also MAC-Adressen) oder „Router“ (solche arbeiten mit Schicht-3-Adressen, im Internet sind dies die IP-Adressen) sein.

Das Bedienungssystem „Paket-Vermittlungsknoten“ beinhaltet mehrere Eingangs-Forderungsströme, deren exakte statistische Eigenschaften nicht bekannt sind. Die eintreffenden Forderungen werden (abhängig von den Adressen-Informationen und den Routing-Vorgaben) auf verschiedene Ausgangsrichtungen (Abnehmer) zugeordnet. Die Warteplatz-Anzahl jeder Ausgangs-Warteschlange ist begrenzt. Bei der Abfertigung sind eventuell mehrere Prioritäten zu berücksichtigen. Die Bediendauer-Verteilungen der einzelnen Forderungs-Klassen sind meist nicht genau bekannt. Der Gesamt-Bedienvorgang kann deshalb nicht exakt modelliert werden, sondern ist (mit entsprechendem Aufwand) nur messtechnisch exakt erfassbar. Rechnerisch sind immer nur (mehr oder weniger gute) Annäherungen an die Realität möglich.

Nachfolgend wird vorausgesetzt, dass jede Ausgangsrichtung eines Paket-Vermittlungsknotens einzeln betrachtet werden kann. Für jede Ausgangsrichtung soll als Eingangs-Prozess näherungsweise ein poissonscher Forderungsstrom (Zufallsverkehr 1.Art, Pure Chance Traffic) vorliegen. Die Warteplatz-Anzahl der Ausgangs-Warteschlange sei so groß, dass Forderungen niemals zu Verlust gehen können. Die Bediendauern seien entweder konstant oder exponentiell verteilt. Bei anderer Bediendauer-Verteilung sei Mittelwert und Variationskoeffizient der Bediendauer bekannt. Falls mehr als eine Verkehrsklasse vorliegt, wird statistische Unabhängigkeit der Forderungsströme vorausgesetzt. Als Abfertigungsstrategie innerhalb einer Verkehrsklasse wird immer FCFS (first come first served) vorausgesetzt.

Bei exponentiell verteilter Bediendauer ergeben obige Voraussetzungen das Erlang-Wartesystem $M / M / 1 / \infty$ mit einem Abnehmer. Bei anderen Bediendauer-Verteilungen kann die mittlere Wartedauer mit der sogenannten Khintchine-Pollazcek-Formel für das Bedienungssystem $M / G / 1 / \infty$ (ohne Priorisierung) berechnet werden.

Nachfolgend werden zunächst die Grundlagen zu Bedienungssystemen beschrieben, welche die (zumindest näherungsweise) Berechnung von Wartezeit-Problemen ermöglichen. Bei verbleibenden Fragen zu dieser Thematik wird auf [ROBE94, HAZO95, TRAN96] verwiesen, weitere Literaturhinweise sind im Text eingefügt. Zunächst wird das Bedienungssystem $M / G / 1 / \infty$ ohne Priorisierung, anschließend das Bedienungssystem $M / G / 1 / \infty$ mit nicht unterbrechender Priorisierung betrachtet.

Hinweis zu den verwendeten Formelzeichen:
In diesem Kapitel werden die mittleren Bediendauern mit t_m bezeichnet. Die mittlere Bediendauer t_m eines Datenpakets in einem Paketvermittlungs-Knoten ist identisch mit der Serialisierungsdauer t_s dieses Datenpakets.

Das Formelzeichen μ_n wird nachfolgend für das n-te Moment einer Zufallsvariablen verwendet und nicht (wie in der Bedienungstheorie-Literatur auch oft üblich) für die Belegungsende-Einfallrate der n-ten Service-Klasse.

4.2 Wartesystem (M / G / 1 / ∞) ohne Priorisierung

4.2.1 Beschreibung

Das Wartesystem $M / G / 1 / \infty$ weist folgende Eigenschaften auf:

- Poissonscher Forderungsstrom mit der Einfallrate λ,
- beliebig verteilte Bediendauer τ_b mit Mittelwert t_m und Streuung σ_b;
- genau ein Abnehmer ($N = 1$),
- unendlich viele Warteplätze ($S = \infty$).

Als Bedien-Strategie wird nachfolgend immer FCFS (first come first served) unterstellt. Die Ergebnisse für die mittlere Wartedauer und alle daraus abgeleiteten Größen gelten jedoch unverändert für alle Abfertigungsstrategien, soweit diese unabhängig von der Bediendauer der einfallenden Forderungen sind. Die hier nicht betrachtete Streuung der zufälligen Wartedauer ist dagegen stark abhängig von der verwendeten Bedienstrategie.

Bild 4.1 zeigt das Blockschaltbild des Wartesystems $M / G / 1 / \infty$. Eine einfallende Forderung wird sofort bedient, wenn keine Forderung im System vorhanden ist. Andernfalls wird eine einfallende Forderung am Ende der Warteschlange eingeordnet und (bei der hier voraus gesetzten Bedienstrategie FCFS) erst dann bedient, wenn alle früher im Wartesystem eingetroffenen Forderungen das Wartesystem verlassen haben.

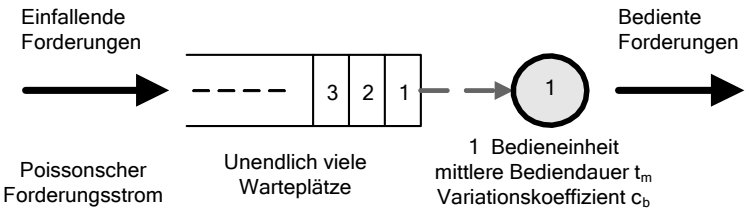


Bild 4.1:
Blockschaltbild des Wartesystems M / G / 1 / ∞.

4.2.2 Wartewahrscheinlichkeit p_w

Bei einem Wartesystem mit einem Abnehmer (N = 1) muss das Angebot $A = \lambda \cdot t_m < 1$ sein (mit λ = Forderungs-Einfallrate, t_m = mittlere Bediendauer), damit sich ein stationärer Zustand einstellen kann. Da keine Verluste auftreten, muss die Abnehmer-Belastung y = A sein. Die Auslastung (utilization) des Abnehmers ist definiert zu $\rho = y / N = y / 1 = y$. Somit gilt für ein Wartesystem mit genau einem Abnehmer immer:

$$A = y = \rho = \lambda \cdot t_m < 1.$$

Die Wartewahrscheinlichkeit ist die Wahrscheinlichkeit, dass eine einfallende Forderung einen belegten Abnehmer sieht. Wenn der einzige vorhandene Abnehmer die Belastung ρ (mit $\rho < 1$) aufweist, sieht eine einfallende Forderung diesen Abnehmer mit der Wahrscheinlichkeit ρ belegt. Damit ist die Wartewahrscheinlichkeit:

$$p_w = \rho = \lambda \cdot t_m ;$$

4.2.3 Mittlere Wartedauer t_w (KP-Formel, Schreibweise A)

Die nachfolgende Formel für die mittlere Wartedauer wird als Khintchine-Pollazcek-Formel bezeichnet, nachfolgend wird die Kurzbezeichnung „KP-Formel“ verwendet. Zur Ableitung der KP-Formel wird auf [ROBE94, HAZO95, KADE95, TRAN96] verwiesen.

Für die mittlere Wartedauer t_w einer Forderung unter Berücksichtigung aller Forderungen ergibt sich beim Wartesystem $M / G / 1 / \infty$ das nachfolgende Ergebnis:

(Formel 4.1)

$$\frac{t_w}{t_m} = \frac{\rho}{1-\rho} \cdot \left(\frac{1+c_b^2}{2} \right);$$

Dabei ist c_b der Variationskoeffizient der Bediendauer:

$$c_b := \frac{\text{Streuung der Bediendauer}}{\text{Mittelwert der Bediendauer}} = \text{Variationskoeffizient der Bediendauer;}$$

Die KP-Formel gilt für **beliebige Bediendauer-Verteilungen**! Neben der Belastung ρ (auch als Auslastung oder utilization bezeichnet) des Abnehmers geht nur der Variationskoeffizient der Bediendauer (c_b = Streuung / Mittelwert) in die Berechnung der mittleren Wartedauer ein.

Für die KP-Formel werden verschiedene Schreibweisen verwendet. Um dem Leser den Vergleich mit anderen Veröffentlichungen zu diesem Thema zu erleichtern, werden nachfolgend zwei häufig verwendete Schreibweisen abgeleitet.

4.2.4 Mittlere Wartedauer t_w (KP-Formel, Schreibweise B)

Unter Verwendung von $\rho = \lambda \cdot t_m$ und den nachfolgenden Umformungen

$$t_w = \frac{\rho \cdot t_m \cdot (1+c_b^2)}{2 \cdot (1-\rho)} ;$$

$\rho \cdot t_m \cdot (1+c_b^2) = \rho \cdot t_m + \rho \cdot t_m \cdot c_b^2 = \rho \cdot t_m + \lambda \cdot t_m^2 \cdot (\sigma_b^2 / t_m^2) = \rho \cdot t_m + \lambda \cdot \sigma_b^2;$

kann die mittlere Wartedauer beim Wartesystem $M / G / 1 / \infty$ wie folgt formuliert werden:

(Formel 4.2)

$$t_w = \frac{\rho \cdot t_m + \lambda \cdot \sigma_b^{\,2}}{2 \cdot (1-\rho)} ;$$

Bedeutung der Formelzeichen:

t_w mittlere Wartedauer einer Forderung;

ρ Belastung der Bedieneinheit (utilization, Auslastung);

t_m mittlere Bediendauer für eine Forderung;
λ Rufeinfallrate, mittlere Anzahl der Forderungen je Zeiteinheit;
σ_b Streuung der Bediendauer;

4.2.5 Mittlere Wartedauer t_w (KP-Formel, Schreibweise C)

Der Steinersche Satz lautet in üblicher mathematischer Notation $\sigma^2 = \mu_2 - \mu_1^2$. Dabei ist μ_1 das erste Moment, μ_2 das zweite Moment und σ^2 die Varianz einer Zufallsgröße. Anwendung des Steinerschen Satzes auf die Zufallsgröße „Bedienzeit" τ_b und Umstellen auf die hier verwendeten Bezeichnungen ergibt:

$\sigma_b^2 = EW[\tau_b^2] - (EW[\tau_b])^2 = EW[\tau_b^2] - t_m^2$;

Somit kann der Zähler der KP-Formel (Schreibweise B) wie folgt geschrieben werden:

$[\ \rho \cdot t_m + \lambda \cdot \sigma_b^2\ ] = \lambda \cdot t_m^2 + \lambda \cdot EW[\tau_b^2] - \lambda \cdot t_m^2 = \lambda \cdot EW[\tau_b^2]$

Damit ergibt sich für die mittlere Wartedauer beim Wartesystem M / G / 1 / ∞ folgender Ausdruck:

$$t_w = \frac{\lambda \cdot EW[\tau_b^{\,2}]}{2 \cdot (1-\rho)} = \frac{\rho}{2 \cdot (1-\rho)} \cdot \frac{EW[\tau_b^{\,2}]}{t_m} \ ;$$

(Formel 4.3)

$$\frac{t_w}{t_m} = \frac{\rho}{2 \cdot (1-\rho)} \cdot \frac{EW[\tau_b^{\,2}]}{t_m^{\,2}} \ ;$$

Bedeutung der Formelzeichen:

t_w mittlere Wartedauer einer Forderung;
ρ Belastung der Bedieneinheit (utilization, Auslastung);
t_m mittlere Bediendauer für eine Forderung;
τ_b Zufallsgröße Bediendauer;
$EW[\tau_b^2]$ Mittelwert „μ_2" der quadrierten Bediendauer (quadratischer Mittelwert von τ_b);

Hinweis:
In Formel 4.3 stellt $\rho / [2 \cdot (1-\rho)]$ die normierte mittlere Wartedauer t_w / t_m bei

konstanter Bediendauer t_m dar. Dies wird nachfolgend bei der Diskussion zu M / D / 1 / ∞ noch einmal erläutert.

Der zweite Term ist der Quotient aus zweitem Moment und quadriertem erstem Moment der Bediendauer:

$\mu_2 / \mu_1^2 = EW[\tau_b^2] / \{EW[\tau_b]\}^2 \geq 1.$

Wegen $\sigma_b^2 \geq 0$ ist immer $\mu_2 \geq \mu_1^2$, somit ist der Quotient $\mu_2 / \{\mu_1\}^2 \geq 1$. Der Quotient $\mu_2 / \{\mu_1\}^2$ gibt demnach an, um welchen Faktor die mittlere Wartedauer t_w (bei gleich bleibendem Mittelwert t_m der Bediendauer) bei beliebig verteilter Bediendauer größer ist als bei konstanter Bediendauer.

Bei konstanter Bediendauer ergibt sich eine minimale mittlere Wartedauer, bei jeder anderen Verteilung der Bediendauer (bei gleich bleibendem Mittelwert der Bediendauer) vergrößert sich die mittlere Wartedauer.

4.2.6 Warteschlangen-Länge y_w und mittlere Wartedauer t_{ww}

Allgemein gültig für jedes Wartesystem ist das Little Theorem:

$$y_w = \lambda \cdot t_w ;$$

Die Wartebelastung y_w (= mittlere Warteschlangen-Länge) ist die mittlere Anzahl der (wartenden) Forderungen in der Warteschlange. Aus dem Little-Theorem ableitbar gilt für jedes Wartesystem folgende Beziehung zwischen mittlerer Wartedauer t_w (Mittelwert über alle Forderungen) und mittlerer Wartedauer t_{ww} der wartenden Forderungen (Mittelwert nur über wartende Forderungen):

$$t_{ww} = \frac{t_w}{p_w} ;$$

Damit können aus der Wartewahrscheinlichkeit $p_w = \rho$ und der mittleren Wartedauer t_w die mittlere Warteschlangen-Länge (auch als Wartebelastung bezeichnet) y_w sowie die mittlere Wartedauer der wartenden Forderungen t_{ww} berechnet werden.

4.2.7 Mittlere Durchlaufdauer (Systemzeit)

Die Durchlaufdauer gibt an, wie lange eine Forderung im Wartesystem verbleibt:

Durchlaufdauer = Wartedauer + Bediendauer;

Aus den mittleren Wartedauern t_w, t_{ww} folgen durch Addition der mittleren Bediendauer t_m die mittleren Durchlaufdauern t_{BS}, t_{BSw}.

$t_{BS} = t_w + t_m$; Mittlere Durchlaufdauer bei Berücksichtigung aller Forderungen;

$t_{BSw} = t_{ww} + t_m$; Mittlere Durchlaufdauer der wartenden Forderungen;

4.2.8 Ergebnis-Übersicht zu M / G / 1 / ∞

$\rho = \lambda \cdot t_m < 1;$ Belastung des Abnehmers (utilization); ($A = y = \rho$ bei Wartesystem mit $N = 1$)

$p_w = \rho;$ Wartewahrscheinlichkeit;

$$\frac{t_w}{t_m} = \frac{\rho}{1-\rho} \cdot \left(\frac{1 + c_b^2}{2} \right);$$ Normierte mittlere Wartedauer für alle Forderungen;

$$\frac{t_{ww}}{t_m} = \frac{1}{\rho} \cdot \frac{t_w}{t_m};$$ Normierte mittlere Wartedauer der wartenden Forderungen;

$$\frac{t_{BS}}{t_m} = 1 + \frac{t_w}{t_m};$$ Normierte mittlere Durchlaufzeit für alle Forderungen;

$$\frac{t_{BSw}}{t_m} = 1 + \frac{t_{ww}}{t_m};$$ Normierte mittlere Durchlaufzeit der wartenden Forderungen;

$$y_w = \lambda \cdot t_w = \frac{\rho^2}{1-\rho} \cdot \left(\frac{1 + c_b{}^2}{2} \right);$$ Mittlere Warteschlangen-Länge (Wartebelastung);

4.2.9 Sonderfälle

Wartesystem M / D / 1/ ∞

Beim Wartesystem M / D / 1 / ∞ ist die Bediendauer **konstant** mit Mittelwert t_m, ansonsten gelten vollständig die Aussagen zu M / G / 1 / ∞. Als Bedien-Strategie wird FCFS (first come first served) unterstellt, die Bemerkung zu anderen Bedienstrategien bei M / G / 1 / ∞ gelten auch hier. Bei konstanter

Bediendauer τ_b folgt $\sigma_b = 0$ und somit $c_b = 0$. Aus den Formeln für M / G / 1 / ∞ folgt dann:

(Formel 4.4)

$$\frac{t_w}{t_m} = \frac{\rho}{2 \cdot (1 - \rho)} \; ;$$

Wartesystem M / M / 1 / ∞

Beim Wartesystem M / M / 1 / ∞ ist die Bediendauer **exponentiell verteilt** mit Mittelwert t_m, ansonsten gelten vollständig die Aussagen zu M / G / 1 / ∞. Als Bedien-Strategie wird FCFS (first come first served) unterstellt, die Bemerkung bei M / G / 1 / ∞ zu anderen Bedienstrategien gelten auch hier. Bei exponentiell verteilter Bediendauer τ_b folgt $\sigma_b = EW[\tau_b] = t_m$ und somit folgt für den Variationskoeffizient der Bediendauer c_b = Streuung / Mittelwert = t_m / t_m = 1. Der Term $(1+c_b^2) / 2$ aus der allgemeinen Lösung wird dann 2, somit folgt für die normierte mittlere Wartedauer:

(Formel 4.5)

$$\frac{t_w}{t_m} = \frac{\rho}{1 - \rho} \; ;$$

Einfluss der Bediendauer-Verteilung auf die Wartedauer

Division der beiden letzten Ergebnis-Formeln ergibt:
(t_w bei exponentiell verteilter Bediendauer) / (t_w bei konstanter Bediendauer) = 2;

Bei gleichem Mittelwert t_m der Bediendauer ist bei exponentiell verteilter Bediendauer die mittlere Wartedauer t_w doppelt so groß wie bei konstanter Bediendauer. Wenn die Verteilung der Bediendauer nicht bekannt ist, rechnet man oft mit exponentieller Verteilung, weil sich dann für die berechnete Wartedauer eine Abschätzung nach oben ergibt.

4.3 Wartesystem (M / G / 1 / ∞) mit zwei Verkehrsklassen ohne Priorisierung

4.3.1 Ableitung zu zwei Verkehrsklassen ohne Priorisierung

Nachfolgend wird Mischverkehr bestehend aus zwei Verkehrs-Klassen ohne Priorisierung vorausgesetzt. Jede Forderungs-Klasse soll **konstante** Bediendauer aufweisen.

Die Verkehrs-Klasse 1 mit der konstanten Bediendauer t_{m1} erzeugt die Auslastung ρ_1, die Verkehrsklasse 2 mit der konstanten Bediendauer t_{m2} erzeugt die Auslastung ρ_2, die Gesamt-Auslastung durch beide Verkehrsklassen ist dann $\rho = \rho_1 + \rho_2$. Natürlich muss die Gesamt-Auslastung $\rho < 1$ sein.

Es liegt dann eine Zweipunkt-Verteilung der Bediendauer vor. Da die KP-Formel für das Wartesystem M / G / 1 / ∞ für beliebige Bediendauer-Verteilungen gültig ist, kann sie auch auf diesen Fall angewendet werden. Mit der nachfolgend verwendeten Hilfsgröße

$Q = t_{m2} / t_{m1}$

(als Quotient der mittleren Bediendauern der beiden Verkehrs-Klassen) ergeben sich die Wahrscheinlichkeiten für eine Forderung der Klasse 1 bzw. 2:

$$p_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{\lambda_1 \cdot t_{m1}}{\lambda_1 \cdot t_{m1} + \lambda_2 \cdot t_{m1}} = \frac{\rho_1}{\rho_1 + \rho_2 / Q} = \frac{Q \cdot \rho_1}{Q \cdot \rho_1 + \rho_2} \quad ;$$

$$p_2 = \frac{\lambda_2}{\lambda_1 + \lambda_2} = \frac{\lambda_2 \cdot t_{m2}}{\lambda_1 \cdot t_{m2} + \lambda_2 \cdot t_{m2}} = \frac{\rho_2}{Q \cdot \rho_1 + \rho_2} = \frac{\rho_2}{Q \cdot \rho_1 + \rho_2} = 1 - p_1 \; ;$$

(mit $Q = t_{m2} / t_{m1}$).

Damit können gewöhnlicher Mittelwert und quadratischer Mittelwert der Bediendauer berechnet werden:

$$EW[\tau_b] = t_m = p_1 \cdot t_{m1} + p_2 \cdot t_{m2} \; ;$$
$$EW[\tau_b^2] = p_1 \cdot t_{m1}^2 + p_2 \cdot t_{m2}^2 \; ;$$

Einsetzen der obigen Wahrscheinlichkeiten ergibt nach kurzer Zwischenrechnung:

$$EW[\tau_b] = t_{m1} \cdot \frac{Q \cdot (\rho_1 + \rho_2)}{Q \cdot \rho_1 + \rho_2} \; ; \qquad EW[\tau_b{}^2] = t_{m1}{}^2 \cdot \frac{Q \cdot \rho_1 + Q^2 \cdot \rho_2}{Q \cdot \rho_1 + \rho_2} \; ;$$

Damit folgt unter Anwendung der Schreibweise C für die Kintchine-Pollaczek-Formel mit $\rho = \rho_1 + \rho_2$ folgendes Ergebnis für die mittlere Wartedauer:

$$t_w = \frac{\rho}{2 \cdot (1-\rho)} \cdot \frac{EW[\tau_b{}^2]}{t_m} = \frac{\rho}{2 \cdot (1-\rho)} \cdot \frac{\rho_1 + Q \cdot \rho_2}{(\rho_1 + \rho_2)} \cdot t_{m1} = \frac{\rho_1 + Q \cdot \rho_2}{2 \cdot (1-\rho)} \cdot t_{m1} \; ;$$

Normierung der mittleren Wartedauer t_w auf die konstante Bediendauer t_{m1} der Verkehrsklasse 1 ergibt folgendes Ergebnis für die mittlere Wartedauer $t_{w1} = t_{w2} = t_w$ (Mittelwert über alle Forderungen):

(Formel 4.6)

$$\frac{t_w}{t_{m1}} = \frac{\rho_1 + Q \cdot \rho_2}{2 \cdot [1 - (\rho_1 + \rho_2)]} \qquad \text{mit} \qquad \rho = \rho_1 + \rho_2 \; ; \qquad Q = \frac{t_{m2}}{t_{m1}} \; ;$$

Die mittlere Wartedauer ist für beide Verkehrsklassen gleich (also $t_{w1} = t_{w2} = t_w$), da nach Voraussetzung beide Verkehrsklassen gleichberechtigt behandelt (also nicht unterschieden) werden.

4.3.2 Beispiel zu zwei Verkehrsklassen ohne Priorisierung

Gegeben:
Für die Beispiel-Berechnung werden folgende Verkehrs-Parameter verwendet:

Verkehr der Klasse 1: $\rho_1 = 0.1$, $t_{m1} = 1$;
Verkehr der Klasse 2: $\rho_2 = 0.4$; $t_{m2} = 6$;
Damit ergibt sich: $\rho = 0.5$; $t_{m2} / t_{m1} = Q = 6$;

Gesucht:
Welche mittlere Wartedauer ergibt sich für folgende Fälle:

a) Nur Verkehr der Klasse 1 vorhanden;
b) Nur Verkehr der Klasse 2 vorhanden;
c) Verkehr der Klassen 1 und 2 gleichzeitig vorhanden;

Berechnung:
a) Nur Verkehr der Klasse 1 vorhanden:

Mit $t_{m1} = 1$ und $\rho_1 = 0.1$ folgt: $t_{w1}/t_{m1} = 0.1 / (2 \cdot 0.9) = 0.056$;

b) Nur Verkehr der Klasse 2 vorhanden:
Mit $t_{m2} = 6$ und $\rho_2 = 0.4$ folgt:
$t_{w2}/t_{m2} = 0.4/(2 \cdot 0.6) = 0.333$; $t_{w2}/t_{m1} = [t_{w2}/t_{m2}] \cdot (t_{m2}/t_{m1}) = 0.333 \cdot 6 = 2.0$;

c) Verkehr der Klassen 1 und 2 gleichzeitig vorhanden
Mit $t_{m1} = 1$, $\rho_1 = 0.1$ sowie $t_{m2} = 6$, $\rho_2 = 0.4$ und somit $Q = 6 / 1 = 6$ ergibt sich durch Einsetzen in die oben abgeleitete Formel:

$t_w/t_{m1} = (0.1+6 \cdot 0.4) / (2 \cdot 0.5) = 2.50$;

Diskussion der Ergebnisse
Es war zu erwarten, dass im Fall c wegen Anwesenheit beider Verkehrsklassen die mittlere Wartedauer größer sein muss als in den Fällen a bzw. b. Für die Verkehrsklasse 1 hat sich die mittlere Wartedauer von 0.056 auf 2.5 erhöht, für die Verkehrsklasse 2 hat sich die mittlere Wartedauer von 2.0 auf 2.5 erhöht. Alle Zahlenwerte gelten nur für die angegeben Verkehrs-Parameter.

Hinweis
Im Fall c ist nach den obigen Ableitungen

$$EW[\tau_b] = t_{m1} \cdot \frac{Q \cdot (\rho_1 + \rho_2)}{Q \cdot \rho_1 + \rho_2} = 1 \cdot \frac{6 \cdot (0.1 + 0.4)}{6 \cdot 0.1 + 0.4} = 3.0;$$

$$EW[\tau_b^2] = t_{m1}^2 \cdot \frac{Q \cdot \rho_1 + Q^2 \cdot \rho_2}{Q \cdot \rho_1 + \rho_2} = 1 \cdot \frac{6 \cdot 0.1 + 36 \cdot 0.4}{6 \cdot 0.1 + 0.4} = 15.0;$$

Mit dem Steinerschen Satz folgt die Streuung der Bediendauer zu:

$\sigma_b = \mathrm{sqrt}(15 - 3 \cdot 3) = \mathrm{sqrt}(6) = 2.45$;

Somit ist der Variationskoeffizient der Bediendauer:

$c_b = \sigma_b / t_m = 2.45 / 3 = 0.82$;

Der hier vorliegende Variationskoeffizient der Bediendauer liegt damit nahe am Variationskoeffizient von 1 für exponentiell verteilte Bediendauer. Hätte man die Ergebnis-Formeln für M / M / 1 / ∞ mit $t_m = 3.0$ und $\rho = 0.5$ verwendet, würde sich für die mittlere Wartedauer näherungsweise ergeben:

$$t_w = t_m \cdot \frac{\rho}{1 \cdot (1 - \rho)} = 3 \cdot \frac{0.5}{0.5} = 3.0; \qquad t_w/t_{m1} = 3/1 = 3.0;$$

Die vereinfachte Berechnung mit exponentiell verteilter Bediendauer ergibt eine etwas zu große mittlere Wartedauer (wegen $c_b = 0.82 < 1$). Für eine Abschätzung der Wartedauer nach oben wird deshalb häufig bei $c_b \approx 1$ näherungsweise mit den einfacheren Formeln für exponentiell verteilte Bediendauer gerechnet.

4.3.3 Numerische Auswertung zu zwei Verkehrsklassen ohne Priorisierung

Bild 4.2 zeigt die normierte mittlere Wartedauer t_w/t_{m1} bei Mischverkehr mit zwei Verkehrsklassen. Auf der Abszisse ist die Auslastung ρ_1 durch die Verkehrs-Klasse 1 (hier die Verkehrsklasse mit kurzer Bediendauer) aufgetragen.

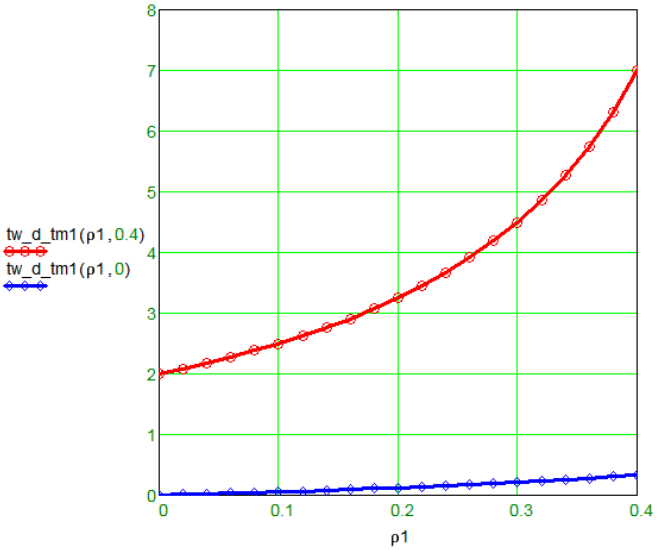
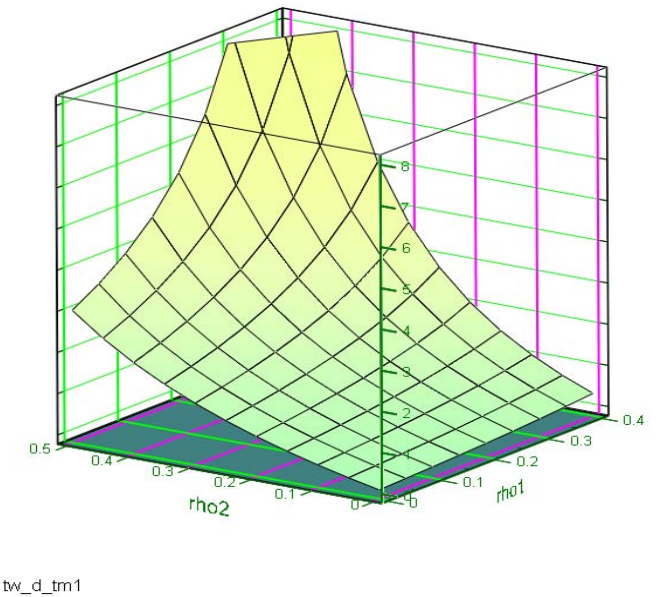


Bild 4.2:
Normierte mittlere Wartedauer bei Mischverkehr **ohne** Priorisierung.
Graph unten: $\rho 1 \in [0, 0.4]$ und $\rho 2 = 0$;
Graph oben: $\rho 1 \in [0, 0.4]$ und $\rho 2 = 0.4$.

Der untere Graph gilt für den Fall $\rho_2 = 0$, dann liegt nur die Verkehrsklasse 1 vor. Dies entspricht dem Fall a aus dem Zahlenbeispiel. Der obere Graph gilt für den Fall $\rho_2 = 0.4$. Für $\rho_1 = 0$ hat der obere Graph den Funktionswert 2.0, dies entspricht der Situation des Zahlenbeispiels Fall b. Der obere Graph zeigt bei-

spielsweise, dass die die normierte Wartedauer für $\rho_1 = 0.4$ und $\rho_2 = 0.4$ genau 7.0 ist.

Bild 4.3 zeigt die normierte mittlere Wartedauer t_w/t_{m1} in Abhängigkeit von den Belastungswerten ρ_1 und ρ_2 als 3D-Diagramm. Nach rechts ist die ρ_1-Achse angetragen, nach links die ρ_2-Achse. Die normierte mittlere Wartedauer t_w/t_{m1} wird durch die 3D-Fläche dargestellt. Ein Schnitt bei $\rho_2 = 0$ durch dieses Diagramm ergibt den unteren Graph aus Bild 4.2, ein Schnitt bei $\rho_2 = 0.4$ durch dieses 3D-Diagramm ergibt den oberen Graph aus Bild 4.2. Als normierte Wartedauer ergibt sich für $\rho_1 = 0.4$ und $\rho_2 = 0.5$ der Wert 17 (im Bild nicht mehr sichtbar).



tw_d_tm1

Bild 4.3:
Normierte mittlere Wartedauer tw / tm1 bei Mischverkehr ohne Priorisierung.
Verkehrs-Klasse 1: Konstante Bediendauer tm1 = 1, Auslastung $\rho 1$;
Verkehrs-Klasse 2: Konstante Bediendauer tm2 = 6, Auslastung $\rho 2$;.
$\rho 1 \in [0, 0.4]$ nach rechts, $\rho 2 \in [0, 0.5]$ nach links aufgetragen.

4.4 Wartesystem (M / G / 1 / ∞) mit zwei Verkehrsklassen mit Priorisierung

Vorbemerkung

Bisher wurde unterstellt, dass alle Forderungen auf Bedienung gleichberechtigt behandelt werden. Bei der Bedienstrategie FCFS erfolgt dann die Bearbeitung in der Reihenfolge des Eintreffens unabhängig von der Art der Forderung.

Üblich ist es, zur Verbesserung der Service-Güte für solche Dienste, welche gegen Verzögerungen empfindlich sind (beispielsweise Sprach-Dialog-Anwendungen, Voice over IP), Prioritäten einzuführen. Höher priorisierte Forderungen werden dann bevorzugt behandelt gegenüber niedriger priorisierten Forderungen.

4.4.1 Wartesystem mit Priorisierung

Bei Wartesystemen mit Priorisierung unterscheidet man unterbrechende (preemptive) und nichtunterbrechende (nonpreemptive) Priorisierung. Bei Internet-Anwendungen kommt eine nichtunterbrechende Priorisierung zum Einsatz. Eine eintreffende Forderung höchster Priorität muss dabei so lange warten, bis ein soeben noch laufender Bedienvorgang für eine niedriger priorisierte Forderung (und die Bedienung aller früher eingetroffenen Forderungen höchster Priorität) abgeschlossen ist. Wenn ein soeben laufender Bedienvorgang für eine niedrig priorisierte Forderung lange andauert, kann eine Forderung höchster Priorität erheblich verzögert werden.

Die **mittlere** Wartedauer beim Wartesystem M / G / 1 / ∞ mit Priorisierung in P Service-Klassen kann exakt berechnet werden. In [ROBE94] wird eine kommentierte Übersicht über die Original-Veröffentlichungen zu dieser Thematik gegeben.

Leider wird in der Bedienungstheorie keine einheitliche Bezeichnungsmethodik verwendet. Im angegebenen Schrifttum werden unterschiedliche Bezeichner verwendet, außerdem werden die Prioritäten unterschiedlich nummeriert (manchmal aufsteigend, manchmal absteigend).

Hier wird folgende Nummerierung der Prioritäten verwendet:

1 = höchste Priorität, 2 = niedrigere Priorität, … P = niedrigste Priorität.

Beschreibung des Bedien-Vorgangs bei zwei Service-Klassen und nicht unterbrechender Priorisierung

Bild 4.4 zeigt das Wartesystem M / G / 1 / ∞ mit Priorisierung und zwei Service-Klassen. Beispielsweise werden (von einem Klassifizierer am Eingang des Bedienungssystems) die hoch priorisierten Forderungen (der Service-Klasse 1) in die obere Warteschlange eingereiht, die niedrig priorisierten Forderungen (der Service-Klasse 2) in die untere Warteschlange.

Wird die Bedieneinheit frei, werden (vom Scheduler am Ausgang des Bedienungssystems) zuerst **alle** Forderungen in der Warteschlange 1 bedient. Eine wartende Forderung in der Warteschlange 2 wird nur dann bedient, falls bei Freiwerden der Bedieneinheit die Warteschlange 1 leer ist. Diese Abfertigungsstragie wird als „priority queuing" bezeichnet.

Wenn eine Forderung der Klasse 2 soeben bedient wird, muss eine eintreffende Forderung der Klasse 1 den soeben laufenden Bedienvorgang abwarten (da nicht unterbrechende Priorität vorliegt), bevor sie bedient werden kann. Dieser Fall ist kritisch, falls die mittlere Bediendauer der niedrig priorisierten Forderungen hoch ist gegenüber der mittleren Bediendauer der hoch priorisierten Forderungen.

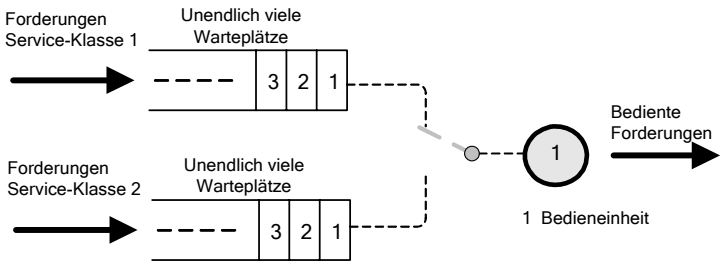


Bild 4.4:
Blockschaltbild des Wartesystems M / G / 1 / ∞ mit Prioritäten.

Hinweis:
Die oben beschriebene Abfertigungsstrategie wird als „nonpreemptive priority queuing" bezeichnet und ergibt für die bevorzugten Forderungen kleinere mitt-

lere Wartedauern als die bei technischen Realisierungen häufig verwendete Abfertigungsstrategie „weighted round robin“. Die nachfolgenden Berechnungen liefern somit immer Mindestwerte (für die mittlere Wartedauer), welche durch technisch realisierte Systeme nicht unterschritten werden können.

4.4.2 Ergebnis-Formeln für nicht unterbrechende Priorisierung

Die Analyse des Wartesystems M / G / 1 / ∞ mit Priorisierung ist mathematisch aufwendig. Beispielsweise werden in [HAZO95, KADE95, LANG92] die mittleren Wartezeiten für nicht unterbrechende Priorisierung und für unterbrechende Priorisierung abgeleitet. In [ROBE94] wird eine kommentierte Übersicht zu den Original-Veröffentlichungen zu diesem Themenbereich gegeben. Nachfolgend werden diese theoretischen Ergebnisse zu Wartesystemen mit Prioritäten-Abfertigung so aufbereitet, dass sie auf den hier interessierenden Fall (zwei Verkehrsklassen, konstante Bediendauer in jeder Verkehrsklasse, festes Verhältnis der Bediendauern, nicht unterbrechende Priorität) anwendbar sind.

Zunächst werden P Prioritätsklassen (mit nicht unterbrechender Priorität, mit 1 = höchste Priorität, mit P = niedrigste Priorität) unterstellt. Nach [LANG92] ergibt sich (nach Anpassung auf die in dieser Arbeit verwendete Bezeichnungsmethodik) für die mittlere Wartedauer der Forderungen mit Priorität k (mit k = 1, 2, … P):

(Formel 4.7)

$$t_{wk} = \frac{\sum_{i=1}^{P} \lambda_i \cdot EW[\tau_{bi}^{\,2}]}{2 \cdot \left(1 - \sum_{i=1}^{k-1} \lambda_i \cdot t_{mi}\right) \cdot \left(1 - \sum_{i=1}^{k} \lambda_i \cdot t_{mi}\right)}; \qquad k = 1, 2, \ldots P; \qquad 1 = \text{höchste Priorität};$$

In obiger Formel ist λ_i die Ankunftsrate der Forderungen mit Priorität i und τ_{bi} die zufällige Bediendauer für diese Forderungen. Die mittlere Bediendauer $EW[\tau_{bi}]$ für die Service-Klasse i wird nachfolgend mit t_{mi} bezeichnet, also $t_{mi} = EW[\tau_{bi}]$. Durch Anwendung des Steinerschen Satzes und Verwendung der Belastungen (utilization) $\rho_i = \lambda_i \cdot t_{mi}$ kann Formel 4.7 umgeformt werden zu:

(Formel 4.8)

$$t_{wk} = \frac{\sum_{i=1}^{P}(\rho_i \cdot t_{mi} + \lambda_i \cdot \sigma_{bi}{}^2)}{2 \cdot \left(1 - \sum_{i=1}^{k-1} \rho_i\right) \cdot \left(1 - \sum_{i=1}^{k} \rho_i\right)}; \qquad k = 1, 2, ... P; \qquad 1 = \text{höchste Priorität};$$

4.4.3 Nicht unterbrechende Priorisierung mit zwei Service-Klassen

Nachfolgend erfolgt eine Beschränkung auf zwei Service-Klassen P = 2. Hohe Priorität 1 erhalten die kurzen (gegen Verzögerungen sehr empfindlichen) Sprach-Pakete. Niedrige Priorität 2 erhalten alle übrigen Kommunikations-Anwendungen, beispielsweise die Daten-Anwendungen. Für den Fall N = 2 ergeben sich mit der Abkürzung $\rho = \rho_1 + \rho_2$ (= Gesamt-Belastung des Wartesystems) aus (Formel 4.8) folgende Ergebnisse für die mittleren Wartedauern:

(Formel 4.9)

$$t_{w1} = \frac{(\rho_1 \cdot t_{m1} + \lambda_1 \cdot \sigma_{b1}{}^2) + (\rho_2 \cdot t_{m2} + \lambda_2 \cdot \sigma_{b2}{}^2)}{2 \cdot (1 - \rho_1)};$$

(Formel 4.10)

$$t_{w2} = \frac{(\rho_1 \cdot t_{m1} + \lambda_1 \cdot \sigma_{b1}{}^2) + (\rho_2 \cdot t_{m2} + \lambda_2 \cdot \sigma_{b2}{}^2)}{2 \cdot (1 - \rho_1) \cdot (1 - \rho)} = \frac{t_{w1}}{(1 - \rho)};$$

Bedeutung der Formelzeichen (mit i = 1, 2):

Index 1 Service-Klasse 1 mit hoher Priorität;

Index 2 Service-Klasse 2 mit niedriger Priorität;

t_{wi} mittlere Wartedauer für Forderungen der Service-Klasse i;

ρ_i Belastung (utilization) durch Service-Klasse i;

ρ $= \rho_1 + \rho_2$; ρ ohne Index ist die Summen-Belastung;

t_{mi} mittlere Bediendauer für Forderungen der Service-Klasse i;

λ_i Einfallrate für Forderungen der Service-Klasse i;

σ_{bi} Streuung der Bediendauer bei Service-Klasse i;

Division der beiden Ergebnis-Formeln ergibt:

$$\frac{t_{w1}}{t_{w2}} = (1-\rho) \; ; \qquad \text{mit} \qquad \rho = \rho_1 + \rho_2 \; ;$$

Die mittlere Wartedauer für die hoch priorisierten Forderungen ist um den Faktor $(1-\rho) < 1$ kleiner als die mittlere Wartedauer für die niedrig priorisierten Forderungen.

Kurze Ergebnis-Diskussion

Bei $\lambda_2 = 0$ wird $\rho_2 = 0$, es liegt dann nur noch die Service-Klasse 1 vor. Für die mittlere Wartedauer muss sich dann die Wartezeit-Formel für M / G / 1 / ∞ ohne Priorisierung ergeben:

$$t_{w1} = \frac{\rho_1 \cdot t_{m1} + \lambda_1 \cdot \sigma_{b1}^{\,2}}{2 \cdot (1-\rho_1)} \; ;$$

Bei $\lambda_1 = 0$ wird $\rho_1 = 0$ und $\rho = \rho_1 + \rho_2 = \rho_2$, es liegt dann nur noch die Service-Klasse 2 vor. Für die mittlere Wartedauer muss sich dann die Wartezeit-Formel für M / G / 1 / ∞ ohne Priorisierung ergeben:

$$t_{w2} = \frac{\rho_2 \cdot t_{m2} + \lambda_2 \cdot \sigma_{b2}^{\,2}}{2 \cdot (1-\rho_2)} \; ;$$

Ein Vergleich mit den Ergebnis-Formeln für M / G / 1 / ∞ ohne Priorisierung (in Schreibweise B) bestätigt die Übereinstimmung.

Sonderfall: Konstante Bediendauer für beide Service-Klassen

Nachfolgend wird unterstellt, dass die Bediendauer für jede Service-Klasse konstant ist, aber unterschiedliche Werte aufweist. Die Ergebnisse für die mittlere Wartedauer vereinfachen sich dann wegen $\sigma_{b1} = \sigma_{b2} = 0$ zu:

$$t_{w1} = \frac{(\rho_1 \cdot t_{m1} + \rho_2 \cdot t_{m2})}{2 \cdot (1-\rho_1)} \; ;$$

$$t_{w2} = \frac{(\rho_1 \cdot t_{m1} + \rho_2 \cdot t_{m2})}{2 \cdot (1-\rho_1) \cdot (1-\rho)} = \frac{t_{w1}}{(1-\rho)} \; ; \qquad \frac{t_{w2}}{t_{w1}} = \frac{1}{(1-\rho)} \; ;$$

mit $\rho = \rho_1 + \rho_2 \; ;$

Nach Normierung der mittleren Wartedauern auf die mittlere Bediendauer t_{m1} der Service-Klasse 1 ergibt sich unter Verwendung des Quotienten $Q := t_{m2}/t_{m1}$ folgendes Ergebnis:

$$\frac{t_{w1}}{t_{m1}} = \frac{\rho_1}{2\cdot(1-\rho_1)} + \frac{t_{m2}}{t_{m1}}\cdot\frac{\rho_2}{2\cdot(1-\rho_1)} = \frac{\rho_1 + Q\cdot\rho_2}{2\cdot(1-\rho_1)};$$

(Formel 4.11)

$$\frac{t_{w1}}{t_{m1}} = \frac{\rho_1 + Q\cdot\rho_2}{2\cdot(1-\rho_1)} = \left[\frac{\rho_1}{2\cdot(1-\rho_1)}\right]\cdot\left(1 + Q\cdot\frac{\rho_2}{\rho_1}\right);$$

(Formel 4.12):

$$\frac{t_{w2}}{t_{m1}} = \frac{t_{w1}}{t_{m1}}\cdot\frac{1}{(1-\rho)} = \frac{t_{w1}}{t_{m1}}\cdot\frac{1}{(1-\rho_1-\rho_2)};$$

Wie bereits allgemein abgeleitet gilt der Zusammenhang:

$$\frac{t_{w2}}{t_{w1}} = \frac{1}{1-\rho} = \frac{1}{1-\rho_1-\rho_2} > 1;$$

Hinweis zu Formel 4.11:
In Formel 4.11 gibt der zweite Term (in der runden Klammer) an, um welchen Faktor sich die mittlere Wartedauer der hoch priorisierten Forderungen bei Mischverkehr mit nicht unterbrechender Priorisierung gegenüber einer isolierten Verkehrsabwicklung der Verkehrsklasse 1 erhöht.

Haben die niedrig priorisierten Forderungen eine hohe normierte Bediendauer $Q = t_{m2} / t_{m1}$, dann führt bereits eine geringe Zusatzbelastung $\rho_2 << 1$ durch die niedrig priorisierten Forderungen zu einer starken Erhöhung der mittleren Wartdauer für die hoch priorisierten Forderungen (verglichen mit den Ergebnissen für eine isolierte Verkehrsabwicklung). Beispielsweise folgt für $\rho_1 = 0.1$, $\rho_2 = 0.4$, $Q = 6$ ein Faktor von

$1 + Q\cdot\rho_2 / \rho_1 = 1 + (6\cdot 0.4) / 0.1 = 25;$

Hinweis zu Formel 4.12:
Formel 4.12 zeigt, dass die mittlere Wartedauer t_{w2} der „benachteiligten“ Forde-

rungen um den Faktor $[1 / (1-\rho)] > 1$ größer ist als die mittlere Wartedauer t_{w1} der „bevorzugten“ Forderungen. Beispielsweise folgt für eine Gesamt-Auslastung $\rho = \rho_1+\rho_2 = 0.5$ der Zusammenhang $t_{w2} = 2 \cdot t_{w1}$.

Vergleich mit den mittleren Wartedauern ohne Priorisierung

Ein Vergleich mit Formel 4.6 (Mischverkehr ohne Priorisierung) ergibt:

$$\frac{t_{w1}(\text{mit Prio.})}{t_{w1}(\text{ohne Prio.})} = \frac{1-\rho}{1-\rho_1} = \frac{1-\rho_1-\rho_2}{1-\rho_1} = 1-\frac{\rho_2}{1-\rho_1} < 1;$$

$$\frac{t_{w2}(\text{mit Prio.})}{t_{w2}(\text{ohne Prio.})} = \frac{1}{1-\rho_1} > 1;$$

Ein Vergleich der mittleren Wartedauern ohne bzw. mit Priorisierung ergibt:

Für die hoch priorisierte Verkehrsklasse 1 (bevorzugte Forderungen) ist die mittlere Wartedauer um den Faktor $(1-\rho) / (1-\rho_1) < 1$ kleiner geworden.

Für die niedrig priorisierte Verkehrsklasse 2 (benachteiligte Forderungen) ist die mittlere Wartedauer um den Faktor $1 / (1-\rho_1) > 1$ größer geworden.

Die Vorteile für die hoch priorisierte (bevorzugte) Verkehrsklasse 1 werden durch Nachteile der niedrig priorisierten (benachteiligten) Verkehrsklasse 2 erkauft. Der exakte Zusammenhang kann mathematisch als „M / G / 1 – Erhaltungssatz“ formuliert werden, hierzu wird auf die angegebene Literatur verwiesen.

Beispiel:

Bei beispielsweise $\rho_1 = 0.1$ und $\rho_2 = 0.4$ wird durch eine Priorisierung der Verkehrsklasse 1 die mittlere Wartedauer der hoch priorisierten Forderungen (Verkehrsklasse 1) um den Faktor $0.5 / 0.9 = 0.555 \approx 0.56$ kleiner (um rund 44% reduziert) als bei „Gleichbehandlung“ der beiden Verkehrsklassen.

Die mittlere Wartedauer der niedrig priorisierten Forderungen (Verkehrsklasse 2) wird bei diesen Auslastungen um den Faktor $1 / 0.9 = 1.11$ höher (um rund 11% erhöht) als bei Gleichbehandlung der beiden Verkehrsklassen.

Als Quotient der mittleren Wartedauern t_{w2} und t_{w1} der beiden Verkehrsklassen ergibt sich (wie oben allgemein abgeleitet):

$t_{w2} / t_{w1} = 1 / (1-\rho) = 1 / 0.5 = 2 = (1.11 / 0.555)$.

4.4.4 Beispiel zu zwei Verkehrsklassen mit Priorisierung

Gegeben:

Die Bediendauern seien in jeder Verkehrsklasse konstant.

Verkehr der Klasse 1: $\rho_1 = 0.1$, $t_{m1} = 1$;

Verkehr der Klasse 2: $\rho_2 = 0.4$; $t_{m2} = 6$;

Damit ergibt sich: $\rho = 0.5$; $t_{m2} / t_{m1} = Q = 6$;

Gesucht:

1) Mittlere Wartedauer für Forderungen der Klasse 1 bei isolierter Verkehrsabwicklung;
2) Mittlere Wartedauer für Forderungen der Klasse 2 bei isolierter Verkehrsabwicklung;
3) Mittlere Wartedauer für Forderungen bei Mischverkehr ohne Priorisierung;
4) Mittlere Wartedauer für Forderungen der Klasse 1 und 2 bei Mischverkehr mit Priorisierung sowie Gesamt-Mittelwert der Wartedauer für alle Forderungen.

Berechnung:

1) $t_{w1} / t_{m1} = \rho_1 / [2\cdot(1-\rho_1)] = 0.056$;
2) $t_{w2} / t_{m1} = (t_{w2} / t_{m2})\cdot(t_{m2}/t_{m1}) = Q\cdot \rho_2 / [2\cdot(1-\rho_2)] = 6\cdot 0.4 / 1.2 = 2.0$;
3) $t_w / t_{m1} = (\rho_1+Q\cdot\rho_2) / ([2\cdot(1-\rho_1-\rho_2)] = (0.1+2.4) / 1 = 2.50$;
4) $t_{w1} / t_{m1} = (\rho_1\cdot Q\cdot\rho_2) / [2\cdot(1-\rho_1)] = (0.1+2.4) / 1.8 = 1.389$;
   $t_{w2} / t_{m1} = [t_{w1}/t_{m1}] / (1-\rho) = 1.389 / (1-0.5) = 2.778$;
   $t_{w2} / t_{w1} = 2.778 / 1.389 = 2$;

Bei den vorgegebenen Belastungen durch Verkehrs-Klasse 1 und Verkehrs-Klasse 2 treten Forderungen der Klasse 1 mit Wahrscheinlichkeit 0.6 auf, solche der Klasse 2 mit Wahrscheinlichkeit 0.4 (siehe früheres Beispiel). Damit ergibt sich bei den vorgegebenen Belastungen der Gesamt-Mittelwert zu:
$t_w / t_{m1} = 0.6\cdot 1.389 + 0.4\cdot 2.778 = 1.95$;

Diskussion:

Bild 4.5 veranschaulicht die „Entwicklung“ der mittleren Wartedauer für das berechnete Beispiel mit $\rho_1 = 0.1$, $\rho_2 = 0.4$, $t_{m1} = 1$, $t_{m2} = 6$ (und somit Q =

$t_{m2} / t_{m1} = 6$). **Bild 4.5 gilt nur für diese Zahlenwerte der Belastungen ρ_i** (mit i = 1, 2) durch die beiden Verkehrs-Klassen.

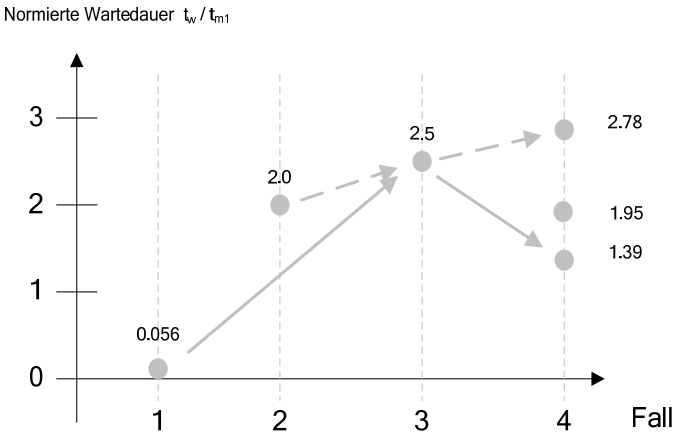


Bild 4.5:
Entwicklung der mittleren Wartedauer für verschiedene Abfertigungs-Szenarien.
Erläuterung der Voraussetzungen und der Randbedingungen für Fall 1 bis 4
siehe Text:
Fall 1: Nur Klasse-1- Verkehr mit 0.1 Erl;
Fall 2: Nur Klasse-2-Verkehr mit 0.4 Erl;
Fall 3: Klasse-1- und Klasse-2-Verkehr gleichzeitig, ohne Priorisierung;
Fall 4: Klasse-1- und Klasse-2-Verkehr gleichzeitig, mit Priorisierung;

Im Fall 1 ist nur die Verkehrsklasse 1 vorhanden (also Gesamt-Belastung 0.1 Erl), die normierte mittlere Wartedauer ist dann 0.056. Im Fall 2 ist nur die Verkehrsklasse 2 vorhanden (also Gesamt-Belastung 0.4 Erl), die normierte mittlere Wartedauer ist dann 2.0. Im Fall 3 sind beide Verkehrsklassen gleichzeitig vorhanden (also Gesamt-Belastung 0.5 Erl), die Abwicklung erfolgt jedoch **ohne Priorisierung**. Die normierte mittlere Wartedauer ist dann 2.50. Im Fall 4 sind beide Verkehrsklassen gleichzeitig vorhanden (also Gesamt-Belastung 0.5 Erl), jedoch ist die **Verkehrsklasse 1 nicht unterbrechend priorisiert** gegenüber Verkehrsklasse 2.

Für die „benachteiligten“ Pakete der Verkehrsklasse 2 erhöht sich die mittlere Wartedauer von 2.50 auf 2.78, wird also um rund 11% vergrößert. Für die hoch priorisierten Pakete der Verkehrsklasse 1 reduziert sich die mittlere Wartedauer von 2.50 auf 1.39, wird also stark reduziert.

Der Gesamt-Mittelwert über alle Pakete sinkt von 2.50 auf 1.95 und wird somit um rund 22% verringert. Diese Verringerung des Gesamt-Mittelwerts ergibt sich, weil die kurzen Pakete (hier 60% Anteil) gegenüber den langen Paketen (hier 40% Anteil) bevorzugt werden.

4.4.5 Numerische Auswertung zu zwei Verkehrsklassen mit Priorisierung

Bild 4.6 zeigt die normierte mittlere Wartedauer t_{w1}/t_{m1} der hoch priorisierten Forderungen bei Mischverkehr mit zwei Verkehrsklassen. Auf der Abszisse ist die Auslastung ρ_1 durch die Verkehrs-Klasse 1 (die hoch priorisierte Verkehrsklasse mit kurzer Bediendauer) aufgetragen, auf der Ordinate die normierte Wartedauer.

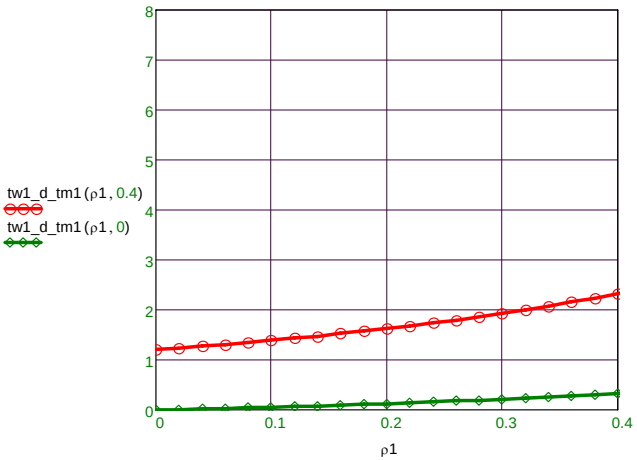


Bild 4.6:
Normierte mittlere Wartedauer der hoch priorisierten Forderungen bei Mischverkehr **mit** Priorisierung.
Graph oben: $\rho 1 \in [0, 0.4]$ und $\rho 2 = 0.4$.
Graph unten: $\rho 1 \in [0, 0.4]$ und $\rho 2 = 0$;

Der untere Graph gilt für den Fall $\rho_2 = 0$, es liegt dann nur die Verkehrsklasse 1 mit der Auslastung ρ_1 vor. Der obere Graph gilt für den Fall $\rho_2 = 0.4$, es liegt dann Verkehrsklasse 1 mit Auslastung ρ_1 und Verkehrsklasse 2 mit Auslastung 0.4 vor. Beispielsweise zeigt der obere Graph, dass die die normierte Wartedauer bei $\rho_1 = 0.4$ und $\rho_2 = 0.4$ genau 2.33 ist.

Die Ergebnisse in Bild 4.6 können mit den Ergebnissen in Bild 4.2 verglichen werden. Dort ist dieselbe Situation ohne Priorisierung dargestellt. Durch die Priorisierung hat sich beispielsweise die normierte Wartedauer für die hoch priorisierten Forderungen bei $\rho_1 = 0.4$ und $\rho_2 = 0.4$ von 7.0 (siehe Bild 4.2) auf 2.33 (siehe Bild 4.6) verringert. Zur besseren Vergleichbarkeit wurde in Bild 4.6 dieselbe Skalierung verwendet wie in Bild 4.2.

Bild 4.7 zeigt die normierte mittlere Wartedauer t_{w1}/t_{m1} der hoch priorisierten Forderungen bei Mischverkehr mit zwei Verkehrsklassen in Abhängigkeit von den Belastungswerten ρ_1 und ρ_2 als 3D-Diagramm. Nach rechts ist die ρ_1-Achse angetragen, nach links die ρ_2-Achse. Die normierte mittlere Wartedauer t_{w1}/t_{m1} der hoch priorisierten Verkehrsklasse wird durch die 3D-Fläche dargestellt.

Ein Schnitt bei $\rho_2 = 0$ durch dieses Diagramm ergibt den unteren Graph aus Bild 4.6, ein Schnitt bei $\rho_2 = 0.4$ durch dieses 3D-Diagramm ergibt den oberen Graph aus Bild 4.6.

Die Ergebnisse in Bild 4.7 können mit den Ergebnissen aus Bild 4.3 verglichen werden. In Bild 4.3 ist dieselbe Situation ohne Priorisierung dargestellt. Als normierte Wartedauer ergibt sich für $\rho_1 = 0.4$ und $\rho_2 = 0.5$ der Wert 2.83 (das Maximum der 3D-Fläche). Durch die Priorisierung hat sich beispielsweise die normierte Wartedauer für die hoch priorisierten Forderungen bei $\rho_1 = 0.4$ und $\rho_2 = 0.5$ von 17 (siehe Diskussion zu Bild 4.3) auf 2.83 (das Maximum der 3D-Fläche in Bild 2.7) verringert. Zur besseren Vergleichbarkeit der Ergebnisse wurde in Bild 4.7 dieselbe Skalierung verwendet wie in Bild 4.3

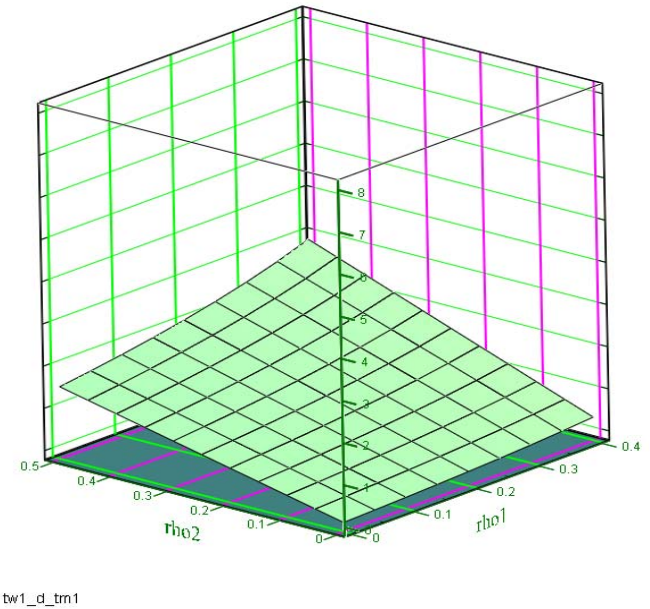


Bild 4.7:
Normierte mittlere Wartedauer t_{w1} / t_{m1} der hoch priorisierten Forderungen bei Mischverkehr **mit** Priorisierung.
Verkehrs-Klasse 1: Konstante Bediendauer $t_{m1} = 1$, Auslastung ρ_1;
Verkehrs-Klasse 2: Konstante Bediendauer $t_{m2} = 6$, Auslastung ρ_2;.
$\rho_1 \in [0, 0.4]$ nach rechts, $\rho_2 \in [0, 0.5]$ nach links aufgetragen.
Verkehrs-Klasse 1 hat höchste Priorität.

4.5 Abschätzung der maximalen Wartedauer beim Wartesystem mit Priorisierung

Worst Case Situation

Im ungünstigsten Fall beginnt soeben vor dem Eintreffen einer hoch priorisierten Forderung (mit kurzer Bediendauer) ein Bedienvorgang für eine niedrig priorisierte Forderung (mit langer Bediendauer). Bei nicht unterbrechender Priorität muss dann bis zum Ende dieses (lang andauernden) Bedienvorgangs gewartet werden, bevor die hoch priorisierte Forderung bedient werden kann. Dies ergibt eine maximale Wartedauer von $t_{w1,max} = t_{m2}$ und somit:

$$\frac{t_{w1,max}}{t_{m1}} = \frac{t_{m2}}{t_{m1}} = Q; \qquad \text{(Näherung für } \rho_1 << 1);$$

Dieses Ergebnis ist nur **näherungsweise** bei kleinem ρ_1 gültig ($\rho_1 << 1$, z. B. $\rho_1 \leq 0.2$). Bei großem ρ_1 muss zusätzlich die Wartesituation innerhalb der Verkehrsklasse 1 berücksichtigt werden. Eine exakte Worst-Case-Analyse wird in [SCHM03] durchgeführt.

Beispiel
Mit den Zahlenwerten des letzten Beispiel ($\rho_1 = 0.1$, $\rho_2 = 0.4$, $t_{m2}/t_{m1} = Q = 6$) lag die normierte mittlere Wartedauer bei „isoliertem" Klasse-1-Verkehr bei 0.056. Im obigen Worst-Case-Fall folgt die normierte **maximale** Wartedauer:

$t_{w,max} / t_{m1} = t_{m2} / t_{m1} = Q = 6;$

Verglichen mit der **mittleren** Wartedauer bei isolierter Abwicklung der hoch priorisierten Forderungen (dies ergäbe für die hoch priorisierten Forderungen identische Wartezeiten wie bei unterbrechender Priorisierung) resultiert (bei den vorliegenden Paketlängen und den vorliegenden Auslastungen durch Sprache bzw. Daten) eine Erhöhung der Wartedauer um den Faktor $6 / 0.056 \approx 100$.

Verglichen mit der **mittleren** Wartedauer der hoch priorisierten Forderungen beim Wartesystem mit nicht unterbrechender Priorisierung ergibt sich (bei den vorliegenden Randbedingungen) eine Erhöhung um den Faktor $6 / 1.39 = 4.3$.

5 Netzberechnungen

In diesem Kapitel werden die bisher abgeleiteten allgemeinen Teil-Ergebnisse auf ein sehr einfaches Modell-Netz angewendet. Dies ermöglicht die numerische Berechnung der Ende-zu-Ende-Durchlaufverzögerung in einem Paketvermittlungsnetz in Abhängigkeit von dessen Parametern. Vorwiegend wurden Mittelwerte (Mittelwert-Analyse) berechnet, für einen Fall wurden näherungsweise die Maximalwerte (Maximum-Analyse, worst case analysis) berechnet.

Als Nachrichtenverkehr wird Mischverkehr aus zwei Verkehrsklassen vorausgesetzt. Die Verkehrsklasse 1 sei Echtzeitverkehr (z. B. Sprachverkehr) mit konstanter kurzer Paketlänge, die Verkehrsklasse 2 sei Datenverkehr mit konstanter langer Paketlänge. Die beiden Verkehrsklassen seien entweder gleichberechtigt (keine Priorisierung), alternativ sei die Verkehrsklasse 1 (Echtzeitverkehr, Sprachverkehr) nicht unterbrechend priorisiert gegenüber der Verkehrsklasse 2 (Datenverkehr). Für die maximalen Datenpaket-Längen werden die Zahlenwerte für die Ethernet-Technik verwendet.

Die Ende-zu-Ende-Durchlaufverzögerung in einem paketbasierten Telekommunikationsnetz unter Berücksichtigung aller wesentlichen Teilverzögerungen (wegen Paketisierung im Endgerät A, Jitter-Buffer-Verzögerung im Endgerät B, Serialisierung im Endgerät A und allen Vermittlungsknoten, physikalischer Signallaufzeit auf allen Netzkanten, Wartezeiten in allen Paketvermittlungsknoten) wird nachfolgend für die gegen Verzögerungen sehr empfindliche Verkehrsklasse 1 (Sprache) numerisch berechnet. Diese wesentlichen Teilverzögerungen sind allein durch das Bearbeitungs-Prinzip bedingt und können bei gegebenen technischen Daten für die Netzknoten und Netzkanten nicht vermieden oder verringert werden.

Mögliche zusätzliche Verzögerungen durch zusätzliche Signalverarbeitungs-Vorgänge (beispielsweise für Quellencodierung, für softwarebasierte Verschlüsselung) sowie durch Rechenzeiten in den Vermittlungsknoten (beispielsweise für die Kanalcodierung, für softwarebasiertes Routing usw.) sind von der Realisierung der eingesetzten Produkte abhängig und werden nicht berücksichtigt. Für eine Berücksichtigung müssten Hersteller-Produkt-Daten vorhanden sein, außerdem würden die Ergebnisse dann nur für diese speziellen Produkte gelten. Solche Software-Rechenzeiten können außerdem zukünftig durch weiter steigende Rechenleistungen „beliebig“ verringert werden.

Für die Sprachsignal-Digitalisierung wurde in dieser Untersuchung immer Standard-Pulscodemodulation (PCM-Sprache nach ITU-Empfehlung G.711 wie im

ISDN-Durchschaltenetz) unterstellt. Als Verzögerung durch den Analog-Digital-Umsetzvorgang sowie den Digital-Analog-Umsetzvorgang ergibt sich dann jeweils eine Abtastperiodendauer und somit resultierend $t_{ums,ab} = 2 \cdot T_a = 2 \cdot 0.125$ ms = 0.25 ms. Dieser kleine Wert wurde nicht weiter berücksichtigt, weil er gegenüber der „großen“ Paketisierungszeit im Endgerät praktisch vernachlässigbar ist.

Werden aufwendigere Quellencodierungs-Verfahren als Standard-PCM nach ITU-Empfehlung G.711 eingesetzt, müssen zusätzliche Verzögerungszeiten im Endgerät berücksichtigt werden. Eine Übersicht zu den Verzögerungszeiten und MOS-Werten einiger Quellencodierungsverfahren gibt Tabelle 5.1 [NOEL03, CISCO].

Standard	Bezeichnung	Bitrate [kbit / s]	Auflösung [bit]	NF-Bandbreite [kHz]	MOS-Wert	Rahmengröße [ms]	Typische Verzögerung [ms]
G.711	PCM	64	8	3.4	4.4	sample	0.125
G.721	ADPCM	32	8	3,4	4.2	sample	0.125
G.722	SB-ADPCM	64	16	7	4.5	sample	1.625
G.722.1	Transform-Codierung	24/32	16	7	%	20	40
G.729	CS-ACELP	8	8	3.4	3.9	10	15

Tabelle 5.1:
Typische Verzögerungszeiten für die Sprachsignal-Digitalisierung (Quellencodierung).
PCM Pulscodemodulation;
ADPCM Adaptive Differenz-Pulscodemodulation;
SB-ADPCM Subband-ADPCM;
CS-ACELP Conjugate Structure Algebaric Code Excited Linear Prediction;

Die in Tabelle 5.1 angegebenen Quellencodierungs-Verzögerungszeiten sind bei Anwendung der entsprechenden Quellencodierungs-Verfahren zusätzlich zu den nachfolgend berechneten Verzögerungswerten zu berücksichtigen. Ändert sich zusätzlich die Sprachpaket-Größe (gegenüber dem für die Modell-Rechnungen verwendeten Zahlenwert), sind auch die Verzögerungs-Berechnungen anzupassen.

5.1 Parameter für die Netzberechnungen

Die in früheren Kapiteln abgeleiteten Formeln sind (unter Beachtung der jeweils genannten Voraussetzungen) allgemein gültig. Für jeden Knoten bzw. jede Kante des Netzes sind real unterschiedliche Parameter möglich. Damit die Anzahl der Eingangs-Parameter überschaubar bleibt, wurden für die nachfolgende numerische Auswertung dieselben Parameter für jede Netzkante und für jeden Paketvermittlungsknoten gewählt. Es können dann die in Kapitel 3.9 beschreibenen Vereinfachungen bei der Berechnung der Gesamt-Wartezeit benutzt werden. Für die Parameter der Nachrichtenquelle und Nachrichtensenke wurden typische Zahlenwerte gewählt.

Tabelle 5.2 zeigt die für alle nachfolgenden Berechnungen verwendeten Parameter. Mit den in Tabelle 5.2 angegebenen Zahlenwerten wurden verschiedene Netz-Modelle durchgerechnet. Die Netzmodelle unterscheiden sich im Wesentlichen dadurch, dass nur eine Verkehrsklasse, zwei Verkehrsklassen, keine Priorisierung oder (nicht unterbrechende) Priorisierung in den Paketvermittlungsknoten zu berücksichtigen ist.

Bild 5.1 zeigt den betrachteten Weg durch das Paketvermittlungsnetz mit den wichtigsten Parametern. Die für die Berechnungen gewählte Entfernung von 1000 km entspricht etwa der maximalen Entfernung zweier (Fernsprech-)Teilnehmer innerhalb Deutschlands.

5.2 Berechnung wichtiger Zahlenwerte

Deterministische Teilverzögerungen

Mit den in Tabelle 5.2 angegebenen Eingangs-Parametern ergeben sich folgende deterministische Teilverzögerungen:

Paketisierung:	(200·8 bit) / (64 kbit/s) = 25 ms ;
Jitter-Buffer mit $\beta = 0.5$:	0.5·25 ms = 12.5 ms;
Physikalische Signallaufzeit: (bei Kabel-Übertragung)	(1000 km) / (200 000 km/s) = 5 ms;
Gesamt-Serialisierungsdauer:	(10+1)·(258·8 bit) / 100 Mbit/s = 0.227 ms;
Summe:	42.73 ms;

Die oben berechneten Teilverzögerungen sind konstant und unabhängig von der (Verkehrs-) Auslastung der Netzkanten. Der konstante, deterministische Anteil an der Ende-zu-Ende-Durchlaufverzögerung beträgt bei den verwendeten Eingangs-Parametern aus Tabelle 5.2 somit 42.73 ms.

Zu diesem deterministischen Mindestwert addieren sich die zufälligen Wartezeiten in den Paketvermittlungsknoten, welche von den Belastungen der Netzkanten und der Verkehrsaufteilung auf die beiden Verkehrsklassen sowie von der Bedienstrategie (ohne / mit Priorisierung, Art der Priorisierung) abhängig sind.

Die nachfolgend berechneten Ende-zu-Ende-Durchlaufverzögerungen weisen deshalb stets einen (durch die Zahlenwerte aus Tabelle 5.2 bedingten) Mindestwert von 42.73 ms bei unbelastetem Netz auf. Bei ansteigender Verkehrsbelastung ergeben sich die in den nachfolgenden Diagrammen dargestellten höheren Gesamtwerte.

Hinweis:
Beim Durchschalte-Vermittlungsprinzip ergibt sich bei 1000 km Entfernung im Netz bei insgesamt 10 Durchschalte-Vermittlungsknoten (mit jeweils einer Durchschalte-Verzögerung von < 1 ms) eine resultierende, konstante Ende-zu-Ende-Durchlaufverzögerung von (5+10·1) ms = 15 ms. Beim Vergleich verschiedener Vermittlungsverfahren sind diese 15 ms den nachfolgend berechneten Zahlenwerte für paketbasierte Netze gegenüber zu stellen.

Bezeichner	**Wert**	**Bedeutung**
N	12	Anzahl der Knoten des Netz-Graphs (Endgeräte plus Vermittlungsknoten)
N_{PVK}	10	= N-2 = Anzahl der Paketvermittlungs-Knoten
EF_a	64 kbit/s	Übertragungsgeschwindigkeit der digitalisierten Sprachquelle
EF	100 Mbit/s	Übertragungsgeschwindigkeit der Netzkanten
l_{ges}	1000 km	Gesamt-Entfernung AB im Netz
v	200 000 km/s	Ausbreitungsgeschwindigkeit auf elektrischen Kabeln oder optischen Lichtwellenleitern
n_n	200·8 bit	Anzahl der Nutzbit je Sprachpaket (Verkehrsklasse 1)
n_h	58·8 bit	Anzahl der Headerbit je Sprachpaket
β	0.50	„Vorlauf“ des Jitter-Buffers
$\rho = \rho_1+\rho_2$	< 1.0	Gesamt-Auslastung, Utilization , Belastung einer Netzkante in Erlang
ρ_1	[0, 0.3]	Auslastung durch die hoch priorisierte Verkehrsklasse 1 mit kleiner Bediendauer t_{m1}
ρ_2	[0, 0.6]	Auslastung durch die niedrig priorisierte Verkehrsklasse 2 mit großer Bediendauer t_{m2}
t_{m1}	wird berechnet	Konstante Bediendauer für Verkehrsklasse 1 (hoch priorisiert)
t_{m2}	wird berechnet	Konstante Bediendauer für Verkehrsklasse 2 (niedrig priorisiert)
t_m	wird berechnet	Mittelwert der Bediendauer über alle Forderungen aus Verkehrsklasse 1 und 2
Q	1 bis 40	Quotient $Q = t_{m2} / t_{m1}$ der Bediendauern von Verkehrsklasse 2 und Verkehrsklasse 1

Tabelle 5.2:
Parameter für die Berechnung der Netzmodelle

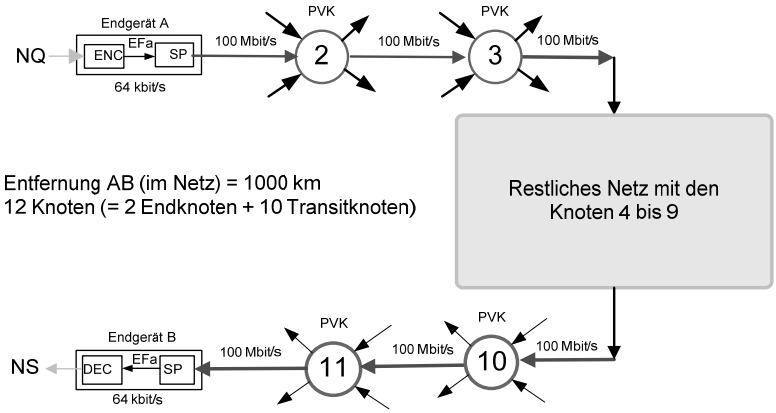


Bild 5.1:
Weg durch ein Modell-Paketvermittlungsnetz mit Parametern aus Tabelle 5.2. NQ Nachrichtenquelle; NS Nachrichtensenke; PVK Paketvermittlungsknoten; SP Speicher; Restliche Parameter siehe Tabelle 5.2.

Einfluss der Entfernung im Netz

Würde man beispielsweise mit einer Entfernung im Netz von 10 000 km rechnen (typischer Wert für eine Transatlantik-Verbindung), würde bei kabelgebundener (elektrischer oder optischer) Signalübertragung allein die zugehörige Laufzeit 50 ms (statt 5 ms bei 1000 km Entfernung) betragen. Die Ende-zu-Ende-Durchlaufverzögerung im Netz würde sich dann gegenüber den in den nachfolgenden Modellen berechneten Zahlenwerten um 45 ms erhöhen. Der deterministische, konstante Anteil der Ende-zu-Ende-Durchlaufverzögerung würde somit 87.73 ms betragen. Die bei ansteigender Auslastung der Netzkanten additiven Wartezeiten erhöhen den Gesamtwert in gleicher Weise wie bei den nachfolgend durchgerechneten Beispielen.

Zusammenhang Netzkanten-Auslastung und Sprachsignal-Transportkapazität

Ein Simplex-Sprachsignal erfordert im Netz (bei einer Sprachcodierung mit einer Netto-Übertragungsgeschwindigkeit EF_a) wegen des Protokoll-Overheads folgende Brutto-Übertragungsgeschwindigkeit:

$EF_{sp,sx} = EF_a \cdot (1+n_h / n_n)$;

Bei einer Netzkanten-Auslastung ρ und einer Übertragungsgeschwindigkeit EF pro Netzkante ist die Anzahl $N_{sp,sx}$ zulässiger Simplex-Sprachsignale:

$N_{sp,sx} = \text{floor}\ [\ (\rho \cdot EF) / EF_{sp,sx}\ ]$;

Beispiel:
Mit den nachfolgenden Parametern aus Tabelle 5.2

$EF_a = 64$ kbit / s, $n_n = 200$ byte, $n_h = 58$ byte, $EF = 100$ Mbit/s

folgt für eine Auslastung der Netzkanten von $\rho = 0.1$ (durch Sprachsignale):

$EF_{sp,sx} = 64 \text{ kbit/s} \cdot (1+58 / 200) = 82.56$ kbit / s;

$N_{sp,sx} = \text{floor}(10 \cdot 10^6 / 82.56 \cdot 10^3) = 121$;

Bei den gegebenen Zahlenwerten (100 Mbit/s pro Netzkante, PCM-Sprachcodierung ohne Sprach-Detektor) können pro 0.1 Erlang Auslastung einer Netzkante (durch Sprachverkehr) rund 120 Simplex-Sprachsignale transportiert werden.

Entsprechend obiger Ableitung ergeben sich pro 0.10 Erlang Auslastung einer Netzkante (bei ansonsten gleichen Voraussetzungen bezüglich der Sprachsignal-Digitalisierung) rund 12 Simplex-Sprachsignale bei 10 Mbit/s je Netzkante bzw. rund 1200 Simplex-Sprachsignale bei 1 Gbit/s.

Netzkanten-Übertragungsgeschwindigkeit	10 Mbit/s	100 Mbit/s	1 Gbit/s
Anzahl Simplex-Sprachsignale je 0.10 Erlang Auslastung (!) einer Netzkante	12	ca. 120	ca. 1200

Beispiel:
Bei 0.30 Erlang Auslastung einer 100 Mbit/s-Netzkante (durch Sprachverkehr) können ca. 360 Simplex-Sprachsignale pro Übertragungsrichtung transportiert werden.

5.3 Übersicht zu den Netz-Modellen

Modell A:
Mittelwert-Analyse, eine Verkehrsklasse
Nur die Verkehrsklasse 1 (Sprache) liegt vor. Als Wartedauer in den Paketvermittlungsknoten wird die **mittlere** Wartezeit berücksichtigt. Die **mittlere** Ende-zu-Ende-Durchlaufverzögerung für die Sprache wird berechnet.

Hinweis:
Bei Mischverkehr mit unterbrechender Priorität für die Sprachpakete würden (für die priorisierten Sprachsignale) ebenfalls die Ergebnisse aus Modell A gelten. Kleinere Verzögerungswerte als nach Modell A sind (bei gleicher Belastung durch die Verkehrsklasse 1) nicht möglich.

Modell B:
Mittelwert-Analyse, zwei Verkehrsklassen, ohne Priorisierung, Q = 6
Verkehrsklasse 1 (Sprache) und Verkehrsklasse 2 (Daten) liegen gleichzeitig vor. Ein Quotient $Q = t_{m2} / t_{m1} = 6$ entspricht näherungsweise der Situation, dass die Verkehrsklasse 1 durch kurze Sprachpakete (258 byte), die Verkehrsklasse 2 durch Ethernet-Pakete mit der Standard-Maximallänge 1512 byte realisiert wird.

In den Paketvermittlungsknoten erfolgt keine Unterscheidung der Verkehrsklassen (keine Priorisierung). Die **mittlere** Ende-zu-Ende-Durchlaufverzögerung für die Sprache wird berechnet.

Modell C:
Mittelwert-Analyse, zwei Verkehrsklassen, ohne Priorisierung, Q = 40
Ähnlich wie Modell B, jedoch wird ein Quotient $Q = t_{m2} / t_{m1} = 40$ verwendet. Dies entspricht der Situation, dass die Verkehrsklasse 1 durch kurze Sprachpakete (258 byte), die Verkehrsklasse 2 ausschließlich durch sehr lange Datenpakete (etwa 10000 byte, sog. Jumbo-Ethernetpaket) realisiert wird. Die **mittlere** Ende-zu-Ende-Durchlaufverzögerung für die Sprache wird berechnet.

Modell D:
Mittelwert-Analyse, zwei Verkehrsklassen, mit Priorisierung, Q = 6
Verkehrsklasse 1 (Sprache) und Verkehrsklasse 2 (Daten) liegen gleichzeitig vor. In den Paketvermittlungsknoten werden die Sprachpakete mit höchster Priorität (nicht unterbrechende Priorität) abgefertigt. Die **mittlere** Ende-zu-Ende-Durchlaufverzögerung für die priorisierte Sprache wird berechnet.

Modell E:
Mittelwert-Analyse, zwei Verkehrsklassen, mit Priorisierung, Q = 40
Ähnlich wie Modell D, jedoch mit Q = 40. Erklärung zum Q-Wert siehe Modell C.

Modell F:
Worst Case-Analyse, zwei Verkehrsklassen, mit Priorisierung, $0 \leq Q \leq 38.8$
Verkehrsklasse 1 (Sprache) und Verkehrsklasse 2 (Daten) liegen vor. In den Paketvermittlungsknoten werden die Sprachpakete mit höchster Priorität (nicht unterbrechende Priorität) abgefertigt. Im ungünstigsten Fall muss in jedem Paketvermittlungsknoten die vollständige Übertragung eines „maximal langen" Datenpakets abgewartet werden, erst dann kann das Sprachpaket übertragen werden.

Als „maximal langes Datenpaket" werden Datenpakete mit einer konstanten Länge bis 10 000 byte angenommen. Bei einer Sprachpaketlänge von 258 byte (siehe Hinweise zu den Eingangsparametern) entspricht dies einem Q-Wert von $0 \leq Q \leq 38.8$.

Die unter diesen Annahmen resultierende **maximale** Ende-zu-Ende-Durchlaufverzögerung (worst case Analyse) für die priorisierte Sprache wird näherungsweise (nur gültig bei geringer Auslastung ρ_1 durch den hoch priorisierten Sprachverkehr) berechnet.

5.4 Übersicht zu den Rechen-Ergebnissen

In Tabelle 5.3 sind die Ergebnisse zu den Modellen A bis F zusammen gestellt. Die Berechnung erfolgte unter Verwendung der Eingangs-Parameter aus Tabelle 5.2. Für die Wartezeiten wurde in Tabelle 5.3 der Maximalwert bei maximaler Auslastung der Netzkanten ($\rho_1 = 0.3$ durch Sprachsignale, $\rho_2 = 0.6$ durch Datensignale) angegeben. Die mittlere Wartezeit liegt zwischen 0 ms (bei Auslastung 0) und dem angegebenen Maximalwert (bei maximaler Auslastung aller Netzkanten).

Im Anhang (A bis F) sind die Modelle A bis F genau beschrieben, die Abhängigkeit der Ende-zu-Ende-Durchlaufverzögerung von den tatsächlichen Belastungswerten $[\rho_1, \rho_2]$ mit $0 \leq \rho_1 \leq 0.3$ und $0 \leq \rho_2 \leq 0.6$ wird durch 3D-Diagramme dargestellt.

Die Software-Rechenzeiten sind nicht berücksichtigt, da diese nicht allgemein berechenbar sind. Hierfür müssten die Daten der eingesetzten Switches und Rou-

ter bekannt sein, die Ergebnisse würden dann nur für das ausgewählte Produkt-Portfolio stimmen und wären nicht mehr allgemein gültig.

Die in Tabelle 5.3 angegebenen Zahlenwerte sind Mindestwerte, welche durch die verwendeten Verfahren (und die vorgegebenen Parameter für die Endgeräte, Netzkanten und Vermittlungsknoten) bedingt sind. Die wesentlichen Einschränkungen und Hinweise sind in den Bemerkungen zur Tabelle 5.3 zusammen gefasst.

Bemerkungen zu Tabelle 5.3:

1) Als Kurzbeschreibung für die Rechenmodelle ist angegeben:
   Anzahl der Service-Klassen
   Mit oder ohne Priorisierung
   Q = (Bediendauer Service-Klasse 2) / (Bediendauer Service-Klasse 1)
2) Die mittlere Gesamt-Wartezeit liegt in Abhängigkeit von der Belastung der Netzkanten durch Sprache und Daten zwischen 0 (bei Belastung 0) und dem angegebenen Zahlenwert (bei maximaler Belastung aller Netzkanten). Für Modell F ist die maximale Gesamt-Wartezeit (Näherungswert bei kleiner Auslastung durch Sprache) angegeben.
3) Für den „Vorlauf" des empfangsseitigen Jitterbuffers wurde mit $\beta = 0.5$ gerechnet. Die zugehörige Verzögerungszeit des Jitterbuffers ist dann $\beta \cdot t_p = 12.5$ ms
4) Für die Sprachsignal-Digitalisierung wurde Standard-Pulscodemodulation nach ITU-Empfehlung G.711 unterstellt. Für den Umsetzvorgang sind dann 0.125 ms pro Endgerät zu berücksichtigen, insgesamt somit 0.25 ms. Dieser kleine Wert wurde bei den Zahlenangaben in Tabelle 5.3 nicht berücksichtigt. Bei anderen Digitalisierungsverfahren sind zusätzliche Umsetz-Verzögerungen zu berücksichtigen, siehe Tabelle 5.1
5) Software-Rechenzeiten sind nicht berücksichtigt.

Modell (siehe Bem. 1)	**Paketisierungsdauer**	**Gesamt-Laufzeit**	**Gesamt-Serial.-Dauer**	**Jitterbuffer-Verzög. (siehe Bem. 3, 4)**	**SW-Gesamt-Rechen-zeit (siehe Bem. 5)**	**Determinist. Anteil**	**Mittlere Gesamt-Wartezeit (bei max. Auslastung, siehe Bem. 2)**	**Ende-zu-Ende-Durchlaufverzögerung (für Sprachsignale)**
Modell A 1 Service-Klasse	25 ms	5 ms	0.2 ms	12.5 ms	0 ms	42.7 ms	**0.9 ms**	43.6 ms
Modell B 2 Service-Klassen Ohne Priorisierung Q = 6	25 ms	5 ms	0.2 ms	12.5 ms	0 ms	42.7 ms	**3.6 ms**	46.3 ms
Modell C 2 Service-Klassen Ohne Priorisierung Q = 40	25 ms	5 ms	0.2 ms	12.5 ms	0 ms	42.7 ms	**22.6 ms**	65.3 ms
Modell D 2 Service-Klassen Mit Priorisierung Q = 6	25 ms	5 ms	0.2 ms	12.5 ms	0 ms	42.7 ms	**0.6 ms**	43.3 ms
Modell E 2 Service-Klassen Mit Priorisierung Q = 40	25 ms	5 ms	0.2 ms	12.5 ms	0 ms	42.7 ms	**3.6 ms**	46.3 ms
Modell F 2 Service-Klassen Mit Priorisierung Q = 40 Worst Case	25 ms	5 ms	0.2 ms	12.5 ms	0 ms	42.7 ms	**7.2 ms**	49.9 ms

Tabelle 5.3:
Ende-zu-Ende-Durchlaufverzögerungen für Sprachsignale im Modell-Netz bei verschiedenen Abfertigungs-Strategien in den Paket-Vermittlungsknoten. Genauere Angaben siehe Anhang A bis F und Text.

6 Zusammenfassung

Die abgeleiteten Ergebnisse ermöglichen die allgemeine und numerische Berechnung der Ende-zu-Ende-Durchlaufverzögerung für hoch priorisierte Sprachsignale in einem paketbasierten Kommunikationsnetz für Sprache und Daten. Die Berechnung setzt verbindungsorientierte Betriebsweise in der Netzwerkschicht (Vermittlungsschicht, OSI-Layer 3) voraus. Die Ergebnisse gelten jedoch näherungsweise auch für verbindungslose Betriebsweise in der Netzwerkschicht (datagram-Betriebsweise, IP-Protokoll), wenn stabile und stationäre Netzbedingungen vorliegen. Als Subnetz-Technologie für die OSI-Schichten 1 und 2 wurde Switched-Ethernet-Technik unterstellt.

Für die Sprachsignal-Digitalisierung wurde bei allen Berechnungen Standard-Pulscodemodulation nach ITU-Empfehlung G.711 unterstellt, dies ergibt minimale Verzögerungen in den Endgeräten. Werden andere Quellencodierungs-Verfahren eingesetzt, sind zusätzliche Verzögerungen (siehe Tabelle 5.1) für die Quellencodierung zu berücksichtigen.

Für die Berechnung müssen die Anzahl der durchlaufenen Vermittlungsknoten, die Parameter der Endgeräte (Sprachpaket-Parameter, Übertragungsgeschwindigkeit für ein Simplex-Sprachsignal) und der Netzkanten (Übertragungsgeschwindigkeit pro Übertragungsrichtung jeder Netzkante, Auslastung in Erlang durch Sprache, Auslastung in Erlang durch Daten) bekannt sein.

Für die Berechnung der Wartezeiten in den Vermittlungsknoten wurde nicht unterbrechende Priorisierung unterstellt. Software-Rechenzeiten wurden bei der Analyse nicht berücksichtigt. Insofern liegt eine Abschätzung der realen Verzögerungswerte nach unten vor. Real werden bei gleichen Parametern stets größere Verzögerungen vorliegen. Es sei außerdem noch einmal deutlich darauf hingewiesen, dass alle in Kapitel 5 berechneten Zahlenwerte nur für das in Kapitel 5 beschriebene Modell-Netz mit den Eigenschaften und Parametern nach Tabelle 5.2 gelten.

Der weit überwiegende Anteil der Ende-zu-Ende-Durchlaufverzögerung bei paketbasierter Sprachkommunikation ist durch die sendeseitige Paketisierung, die Signal-Laufzeiten im Netz und den empfangsseitigen Jitter-Buffer bedingt. Für das verwendete Modell-Netz folgt beispielsweise für die Ende-zu-Ende-Durchlaufverzögerung (one way delay) ein deterministischer Mindestwert (durch Paketisierung, Jitter-Buffer, Serialisierung, Laufzeiten, jedoch ohne zufällige Wartezeiten) bei 1000 km Entfernung im Netz von 42.3 ms. Dieser Mindestwert ist durch die verwendeten Bearbeitungs-Verfahren und durch physikalische

Grenzwerte bedingt und kann durch technische Maßnahmen nicht verringert werden, wenn die in Tabelle 5.2 definierten Parameter (für Paket-Größen, Netz-Eigenschaften) vorliegen. Rechenzeiten durch Software-Verarbeitungsvorgänge sind dabei nicht berücksichtigt.

Wird der Sprachverkehr gegenüber dem Datenverkehr nicht priorisiert, ergeben sich bei den gewählten Netz-Parametern (siehe Tabelle 5.2) mittlere Gesamt-Wartezeiten für den priorisierten Sprachverkehr bei maximaler Auslastung (Sprachverkehr 0.3 Erlang, Datenverkehr 0.6 Erlang) von maximal ca. 23 ms, siehe Modell C. Die resultierende mittlere Ende-zu-Ende-Durchlaufverzögerung würde unter diesen Randbedingungen auf insgesamt ca. 65 ms im Modell-Netz ansteigen.

Bei den vorliegenden Netz-Parametern (siehe Tabelle 5.2) für die Netzmodelle können die in den Paket-Vermittlungsknoten auftretenden Wartezeit-Verzögerungen immer dann vernachlässigt werden, wenn der Sprachverkehr gegenüber dem Datenverkehr (nicht unterbrechend) priorisiert wird. Bei priorisiertem Sprachverkehr ergeben sich (selbst dann, wenn als Datenverkehr im Extremfall nur Jumbo-Ethernet-Datenpakete mit 10 000 byte übertragen werden) mittlere Wartezeit-Verzögerungen (bei maximaler Auslastung) von kleiner 4 ms für den (hoch priorisierten) Sprachverkehr, siehe Modell E. Die resultierende mittlere Ende-zu-Ende-Durchlaufverzögerung beträgt für diesen Fall ca. 46 ms.

Bei einer vereinfachten Worst-Case-Betrachtung ist die maximale Wartezeit für den hoch priorisierten Sprachverkehr ca. 7 ms (näherungsweise, bei geringer Auslastung durch Sprachverkehr, nur Jumbo-Datenpakete), die resultierende maximale Ende-zu-Ende-Durchlaufverzögerung wird dann maximal ca. 50 ms, siehe Modell F.

Bei Priorisierung des Sprachverkehrs liegen im Modell-Netz bei einer Netz-Entfernung AB von 1000 km die Ende-zu-Ende-Durchlaufverzögerungen bei maximaler Auslastung im Wertebereich bis ca. 50 ms. Bei gleichen Randbedingungen ergibt sich für ein Durchschalte-Vermittlungsnetz eine konstante Ende-zu-Ende-Durchlaufverzögerung (als Abschätzung nach oben) von rund 15 ms.

Die berechneten Verzögerungswerte sind theoretische Mindestwerte, welche durch das Vermittlungsprinzip und die physikalischen Randbedingungen verursacht sind. Diese Mindestwerte können bei gegebenen Eigenschaften der Netzelemente (siehe Tabelle 5.2) durch technische Maßnahmen nicht verringert werden. Zusätzliche Verzögerungen durch beispielsweise Software-Bearbeitungsvorgänge in den Netzknoten sind dabei nicht berücksichtigt. Diese Verzögerun-

gen durch Software-Bearbeitungsvorgänge können durch eine (zukünftig weiter) steigende Rechenleistung „beliebig“ reduziert werden. Bei Ersatz von Software-Lösungen durch Signalverarbeitungs-Hardware können diese zusätzlichen Verzögerungen weitgehend vermieden werden.

Die abgeleiteten allgemeinen Formeln ermöglichen bei vorgegebenen Eigenschaften der Netzelemente (Endgeräte, Vermittlungsknoten, Übertragungswege) und vorgegebenen Verkehrs-Auslastungen der Übertragungswege durch Sprache und Daten die näherungsweise Berechnung der mittleren oder maximalen Ende-zu-Ende-Durchlaufverzögerung für den hoch priorisierten Sprachverkehr (bei nicht unterbrechender Priorisierung) in einem paketbasierten Kommunikationsnetz für Sprache und Daten.

Anhang

A Modell A: Mittelwert-Analyse, eine Verkehrsklasse

Für Modell A wird angenommen, dass nur eine Verkehrsklasse (Sprache) vorliegt. Als Wartedauer in den Paketvermittlungsknoten wird die mittlere Wartezeit berücksichtigt. Bild A.1 oben zeigt die mittlere Ende-zu-Ende-Durchlaufverzögerung in ms sowie deren Bestandteile in Abhängigkeit von der Verkehrsbelastung der Netzkanten. Die mittlere Ende-zu-Ende-Durchlaufverzögerung liegt im Wertebereich [42.7 ms, 43.6 ms].

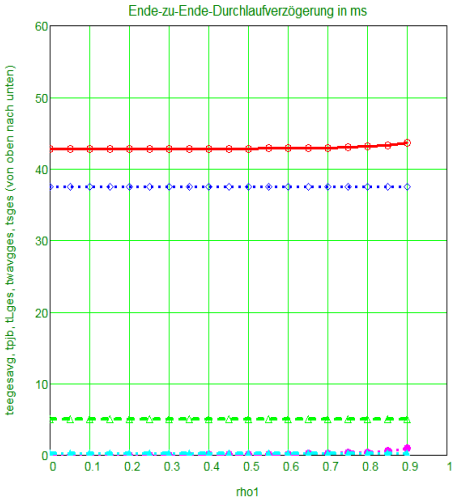


Bild A.1:
Modell A (Mittelwert-Analyse, eine Verkehrsklasse)
Mittlere Ende-zu-Ende-Durchlaufverzögerung und deren Bestandteile in Abhängigkeit von der Verkehrsbelastung ρ (Auslastung, utilization) der Netzkanten.
Eingangs-Parameter siehe Tabelle 5.2.

Bild A.1 zeigt von oben nach unten folgende Verzögerungen (in ms):

- Gesamtwert der Ende-zu-Ende-Durchlaufverzögerung $t_{ee,ges,avg}$;
- Verzögerung durch Paketisierung plus Jitterbuffer t_{pjb};
- Verzögerung durch physikalische Laufzeit $t_{L,ges}$;

- Verzögerung durch Serialisierung in den Netzknoten (unabhängig von ρ_1);
- Mittlere Gesamt-Wartedauer in allen Netzknoten (ansteigend mit ρ_1);

Die deterministischen Anteile ergeben den bereits berechneten Summenwert von 42.7 ms. Für Paketisierung und Jitterbuffer-Verzögerung fallen 37.5 ms an, die physikalische Signallaufzeit (bei der vorliegenden Entfernung von 1000 km) verursacht zusätzliche 5 ms, rund 0.23 ms entstehen durch die hier insgesamt 11 Serialisierungs-Vorgänge beim Auslesen der Paketspeicher auf die Netzkanten. Diese Zahlenwerte gelten ebenso für alle anderen Netzmodelle, da stets von denselben Eingangsdaten (Tabelle 5.2) ausgegangen wird.

Die Serialisierungsdauer für ein Datenpaket der Verkehrsklasse 1 beträgt bei den hier verwendeten Eingangs-Parametern (258·8 bit) / (100 Mbit/s) = 20.64 µs pro Serialisierungsvorgang. Bei einer (extremen) Verkehrsbelastung von 0.9 Erlang je Netzkante folgt damit pro Paketvermittlungsknoten eine mittlere Wartedauer von 20.64 µs·[0.9 / (2·0.1)] = 92.88 µs. Bei insgesamt 10 Paketvermittlungsknoten (und den getroffenen Annahmen über die Anschaltung der Endgeräte) resultiert eine mittlere Gesamt-Wartedauer von (10-1)·93 µs ≈ 0.84 ms. Damit ergibt sich der oben angegebene Maximalwert für die mittlere Ende-zu-Ende-Durchlaufverzögerung von (42.73+0.84) ms ≈ 43.6 ms in Bild 5.2.

Bei diesem Modell wurde (wie oben beschrieben) angenommen, dass kein zusätzlicher Datenverkehr vorliegt.

Günstigere Ergebnisse als bei Modell A sind (bei gleicher Belastung durch den hoch priorisierten Sprachverkehr) niemals möglich.

B Modell B: Mittelwert-Analyse, zwei Verkehrsklassen, ohne Priorisierung, Q = 6

Verkehrsklasse 1 (Sprache) und Verkehrsklasse 2 (Daten) liegen vor. Die Datenpakete sind um den Faktor Q = 6 länger als die Sprachpakete. In den Paketvermittlungsknoten erfolgt keine Unterscheidung der Verkehrsklassen (keine Priorisierung). Die mittlere Ende-zu-Ende-Durchlaufverzögerung für die Sprache wird berechnet.

Bild B.1 zeigt die mittlere Ende-zu-Ende-Durchlaufverzögerung in ms in Abhängigkeit von den Verkehrsbelastungen ρ_1 (durch Verkehrsklasse 1) und ρ_2 (durch Verkehrsklasse 2) jeder Netzkante. Die mittlere Ende-zu-Ende-Durchlaufverzögerung liegt im Wertebereich [42.7 ms, 46.3 ms]. Zur Berechnung der Wartezeiten wird auf Kapitel 4 verwiesen.

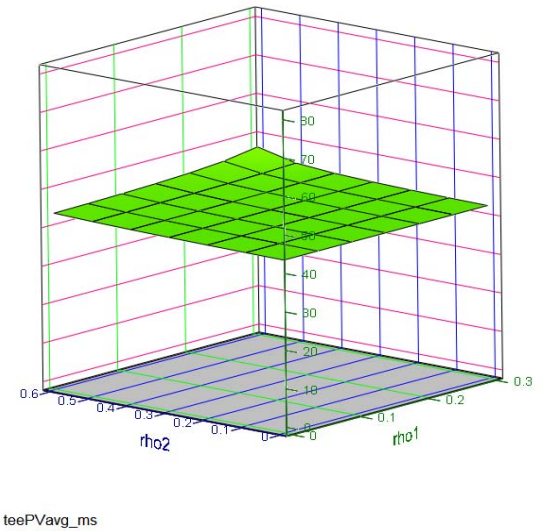


Bild B.1:
Modell B (Mittelwert-Analyse, zwei Verkehrsklassen ohne Priorisierung, Q = 6).
Mittlere Ende-zu-Ende-Durchlaufverzögerung in ms für Verkehrsklasse 1.
Eingangs-Parameter siehe Tabelle 5.2.

C Modell C:
Mittelwert-Analyse, zwei Verkehrsklassen, ohne Priorisierung, Q = 40

Modell C ist ähnlich wie Modell B, jedoch wird der Parameter Q = 40 verwendet. Bild 5.4 zeigt die mittlere Ende-zu-Ende-Durchlaufverzögerung in ms in Abhängigkeit von den Verkehrsbelastungen ρ_1 (durch Verkehrsklasse 1) und ρ_2 (durch Verkehrsklasse 2) jeder Netzkante. Die mittlere Ende-zu-Ende-Durchlaufverzögerung liegt im Wertebereich [42.7 ms, 65.3 ms]. Zur Berechnung der Wartezeiten wird auf Kapitel 4 verwiesen.

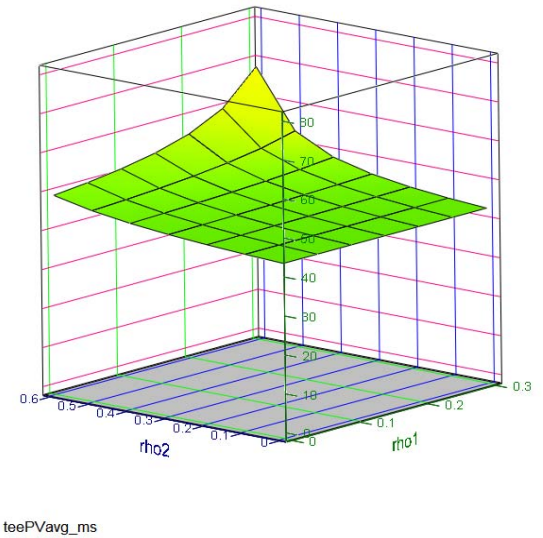


Bild C.1:
Modell C (Mittelwert-Analyse, zwei Verkehrsklassen ohne Priorisierung, Q = 40).
Mittlere Ende-zu-Ende-Durchlaufverzögerung in ms für Verkehrsklasse 1.
Eingangs-Parameter siehe Tabelle 5.2.

Hinweis: Hier ist die Schwankungsbreite der Verzögerungen so groß, dass der Vorlauf β des Jitter-Buffers von 0.5 auf 1 erhöht werden müsste, was zusätzliche 12.5 ms Verzögerung ergeben würde, siehe Hinweise zum Jitter-Buffer.

D Modell D:
Mittelwert-Analyse, zwei Verkehrsklassen, mit Priorisierung, Q = 6

Verkehrsklasse 1 (Sprache) und Verkehrsklasse 2 (Daten) liegen vor. Die Datenpakete sind um den Faktor Q = 6 länger als die Sprachpakete. In den Paketvermittlungsknoten werden die Sprachpakete mit höchster Priorität (nicht unterbrechende Priorität) abgefertigt. Die mittlere Ende-zu-Ende-Durchlaufverzögerung für die Sprachpakete wird berechnet. Bei unterbrechender Priorität für die Verkehrsklasse Sprache würden sich dieselben Werte wie bei Modell A ergeben.

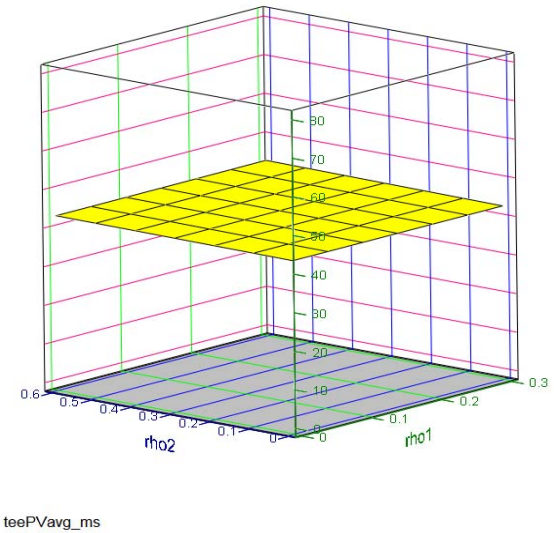


Bild D.1:
Modell D (Mittelwert-Analyse, zwei Verkehrsklassen mit Priorisierung, Q = 6).
Mittlere Ende-zu-Ende-Durchlaufverzögerung in ms für Verkehrsklasse 1.
Eingangs-Parameter siehe Tabelle 5.2.

Bild D.1 zeigt die mittlere Ende-zu-Ende-Durchlaufverzögerung in ms für die priorisierte Verkehrsklasse 1 in Abhängigkeit von den Verkehrsbelastungen ρ_1 (durch Verkehrsklasse 1) und ρ_2 (durch Verkehrsklasse 2) jeder Netzkante. Die mittlere Ende-zu-Ende-Durchlaufverzögerung liegt im Wertebereich [42.7 ms, 43.3 ms]. Zur Berechnung der Wartezeiten wird auf Kapitel 4 verwiesen.

E Modell E:
Mittelwert-Analyse, zwei Verkehrsklassen, mit Priorisierung, Q = 40

Verkehrsklasse 1 (Sprache) und Verkehrsklasse 2 (Daten) liegen vor. Die Datenpakete sind um den Faktor Q = 40 länger als die Sprachpakete. In den Paketvermittlungsknoten werden die Sprachpakete mit höchster Priorität (nicht unterbre-

chende Priorität) abgefertigt. Die mittlere Ende-zu-Ende-Durchlaufverzögerung für die Sprachpakete wird berechnet.

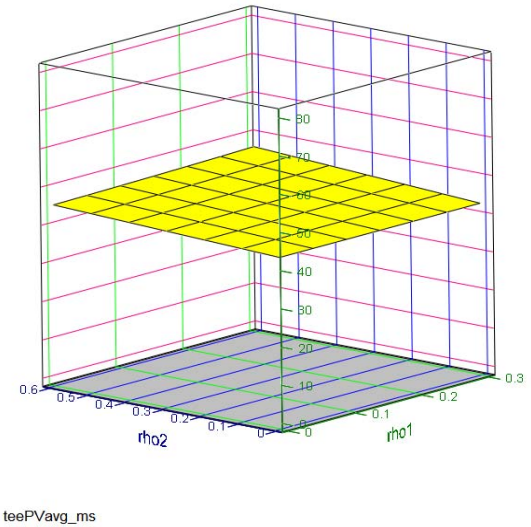


Bild E.1:
Modell E (Mittelwert-Analyse, zwei Verkehrsklassen mit Priorisierung, Q = 40).
Mittlere Ende-zu-Ende-Durchlaufverzögerung in ms für Verkehrsklasse 1.
Eingangs-Parameter siehe Tabelle 5.2.

Bild E.1 zeigt die mittlere Ende-zu-Ende-Durchlaufverzögerung in ms für die Priorisierte Verkehrsklasse 1 in Abhängigkeit von den Verkehrsbelastungen ρ_1 (durch Verkehrsklasse 1) und ρ_2 (durch Verkehrsklasse 2) jeder Netzkante. Die Ende-zu-Ende-Durchlaufverzögerung liegt im Wertebereich [42.7 ms, 46.3 ms]. Zur Berechnung der Wartezeiten wird auf Kapitel 4 verwiesen.

F Modell F:
Worst-Case-Analyse, zwei Verkehrsklassen, mit Priorisierung, $0 \le Q \le 40$

Verkehrsklasse 1 (Sprache) und Verkehrsklasse 2 (Daten) liegen vor. In den Paketvermittlungsknoten werden die Sprachpakete mit höchster Priorität (nicht unterbrechende Priorität) abgefertigt.

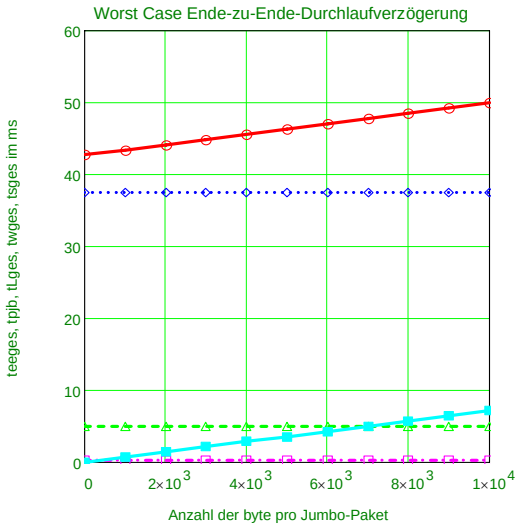


Bild F.1:
Modell F (Worst-Case-Analyse, zwei Verkehrsklassen mit Priorisierung, $0 \le Q \le 38.8$).
Maximale Ende-zu-Ende-Durchlaufverzögerung in ms für Verkehrsklasse 1.
Eingangs-Parameter siehe Tabelle 5.2.

Bild F.1 zeigt von oben nach unten folgende Verzögerungen in ms in Abhängigkeit von der Jumbo-Paketlänge in byte:

- Maximale Ende-zu-Ende-Durchlaufverzögerung der hoch priorisierten Verkehrsklasse (Summe der nachfolgenden Anteile);
- Verzögerung durch Paketisierung und Jitterbuffer;
- Physikalische Gesamt-Laufzeit;
- Maximale Gesamt-Wartezeit (ansteigender Graph);
- Gesamt-Serialisierungsdauer.

Die maximale Ende-zu-Ende-Durchlaufverzögerung (Worst Case Analyse) liegt im Wertebereich [42.7 ms, 49.9 ms]. Dem Wertebereich [0, 10000 byte] der Jumbo-Paketlänge entspricht der Wertebereich [0, 38.8] für den Parameter Q.

Hinweis:
Für die Netz-Parameter nach Tabelle 5.2 ist die maximale Wartezeit pro Paketvermittlungsknoten bei einer „Jumbo-Paketlänge“ von 10 000 byte genau (10 000·8 bit) / (100 Mbit/s) = 0.8 ms. Bei $N_{PVK} = 10$ Paketvermittlungsknoten folgt dann eine maximale Gesamt-Wartezeit von $t_{w1,max,ges} = (10-1)\cdot 0.8$ ms = 7.2 ms, siehe hierzu die Hinweise in Kapitel 3.8. Die minimale Ende-zu-Ende-Durchlaufverzögerung ist durch den deterministischen Anteil von 42.7 ms gegeben. Als Gesamt-Maximalwert folgt somit wie angegeben (42.7 + 7.2) ms = 49.9 ms.

Dieser Maximalwert gilt nur bei geringer Auslastung durch die Verkehrsklasse 1 (beispielsweise $\rho_1 \leq 0.1$ Erl), weil das vereinfachte Berechnungsverfahren für das Modell F die zusätzliche Wartesituation innerhalb der Verkehrsklasse 1 bei hoher Verkehrsauslastung ($0.1 << \rho_1$ / Erl <1) nicht berücksichtigt.

Literaturverzeichnis

BADA07
Badach, A.
Voice over IP – Die Technik
Hanser-Verlag, München / Wien 2007

CISCO
Understanding Delay in Packet Voice Networks
Document ID: 5125
www.cisco.com

HAZO95
Haas, M.; Zorn, W.
Methodische Leistungsanalyse von Rechensystemen
Oldenbourg-Verlag, München Wien 1995

IMHO06
Imhoff, F.
Quality of Service
Der Netzwerk Insider, Juni 2006, S. 8-17
www.comconsult-research.de

ITUG
ITU-T Recommendation G.114
One-way transmission time (05/2000)

ITUP
ITU-T Recommendation P.800
Methods for subjective determination of transmission quality (08/96)

KADE95
Kaderali, F.
Digitale Kommunikationstechnik 2
Vieweg-Verlag, Braunschweig / Wiesbaden 1995

LANG92
Langendörfer, H.
Leistungsanalyse von Rechensystemen
Hanser-Verlag, München Wien 1992

NOCK05
Nocker, R.
Digitale Kommunikationssysteme 2
Grundlagen der Vermittlungstechnik
Vieweg-Verlag, Wiesbaden 2005

NOEL03
Nölle, Jochen
Voice over IP
VDE-Verlag, Berlin / Offenbach 2003

ROBE94
Robertazzi, T. G.
Computer Networks and Systems: Queuing Theory and Performance Evaluation
Springer Verlag, Berlin / Heidelberg / New York 1994

SCHM03
Schmitt, Jens
On Average and Worst Case Behaviour in Non-Preemptive Priority Queuing
Darmstadt University of Technology, Multimedia Communications Lab
http://disco.informatik.uni-kl.de/publications/Schmitt03-2.pdf

SIEG02
Siegmund, G.
Technik der Netze
Hüthig-Verlag, Heidelberg 2002

SIEG02b
Next Generation Networks
Hüthig Verlag, Heidelberg 2002

TRAN96
Tran-Gia, P.
Analytische Leistungsbewertung verteilter Systeme
Springer-Verlag, Berlin / Heidelberg / New York 1996