Mit GPRS ins Intranet – Optimierungsmaßnahmen für den praktischen Einsatz in Unternehmensnetzen

Der Autor
Dipl.-Ing. Stefanus Römer war mehrere Jahre bei T-Data im Produktmanagement für Datex-P sowie FrameLink Plus tätig und dort insbesondere für die Produktentwicklung von LAN to LAN GPRS Access verantwortlich. Seit April 2001 ist er bei T-Mobil im Produktmanagement Mobile Business Solutions beschäftigt und verantwortet dort das Produkt Mobile IP VPN sowie insbesondere die Produktentwicklung von Optimierungslösungen auf der Basis von GPRS.

Die mobile Datenkommunikation hat inzwischen strategische Bedeutung erlangt. Die ständige Erreichbarkeit sowie die Möglichkeit zum Zugriff auf geschäftskritische Daten an jedem Ort und zu jeder Zeit werden in einem sich verschärfenden Wettbewerb immer wichtiger. Kundennähe und Service sind die entscheidenden Wettbewerbsfaktoren. Der direkte Zugriff auf aktuelle und verdichtete Informationen wird somit unverzichtbar. Im Beitrag „LAN to LAN GPRS Access“ aus Heft 3/2001 wurde eine gemeinsame Lösung von T-Data und T-Mobil für den mobilen Intranet-Zugang über GPRS (General Packet Radio Service) vorgestellt. Die Anschaltung der Unternehmensnetze wird bei dieser Lösung auf der Basis der Produkte „Mobile IP VPN“ und „LAN to LAN Dial In“ realisiert. Ergänzend hierzu behandelt der vorliegende Beitrag die Besonderheiten der mobilen Datenübertragung über GPRS und zielt auf entsprechende Techniken zur Optimierung der Übertragungseigenschaften ab. Diese Maßnahmen haben einen entscheidenden Einfluss auf die Qualität der mobilen Datenverbindung und sind daher von großer Bedeutung für die Nutzerakzeptanz. Die Entwicklung und die Verbreitung von Anwendungen und Diensten auf der Basis von GPRS sowie entsprechender Middleware[1] oder Proxy[2]-Lösungen zur Performance-Steigerung sind zudem wichtige Voraussetzungen für den Erfolg von UMTS (Universal Mobile Telecommunications System).

1 Einleitung und Problemstellung

Die spezifischen Übertragungseigenschaften von Mobilfunknetzen beeinflussen auf vielfältige Weise die Arbeitsweise beziehungsweise die Lauffähigkeit von IP-(Internet Protocol-)basierten Protokollen und Anwendungen. Die verbreiteten Datenkommunikations-Protokolle wie die Protokollfamilie TCP/IP (Transmission Control Protocol) wurden in erster Linie für LAN- beziehungsweise Festnetzumgebungen entwickelt und setzen im Allgemeinen eine vergleichsweise hohe Bandbreite sowie ein geringes Delay[3] voraus. Daher sind fallweise besondere Maßnahmen erforderlich, um die Lauffähigkeit der eingesetzten Protokolle und Anwendungen über eine mobile Netzanbindung zu gewährleisten.

Von entscheidender Bedeutung ist dabei die Performance von TCP. Da die Mehrzahl der Anwendungen dieses Protokoll einsetzt, ist die Lauffähigkeit von TCP kritisch für die Akzeptanz und den Erfolg des genutzten mobilen Datendienstes. Als gesichertes, verbindungsorientiertes Protokoll auf der Basis von IP ist TCP besonders empfindlich gegenüber

[1] Middleware: Allgemein für die Erweiterung einer Rechnerarchitektur um eine Verteilungsplattform zwischen Netz-/Betriebssystemschicht und Anwendungsschicht. Das damit erzeugte Application Program Interface gewährleistet die Kommunikation unter den verschiedenen Hardware-Plattformen, Betriebssystemen und Protokollen. Hier: Eine Client-/Server-Software, die zwischen Anwendungs-Client und Anwendungs-Server transparent eingebunden wird, und somit für den Nutzer nicht in Erscheinung tritt.

[2] Proxy: Allgemein für Server, die einen „Stellvertreterdienst“ wahrnehmen – sie nehmen Anforderungen von einem Client entgegen und geben diese, gegebenenfalls modifiziert, an das ursprüngliche Ziel weiter. Proxy-Server können beispielsweise Sicherheitsfunktionen bieten.

[3] Delay: Verzögerungs- oder Wartezeit. Zeitspanne, um die ein Ereignis verzerrt oder verzögert wird. Beispielsweise die Zeit, die vergeht, bis eine abgesandte Information vom Zielsystem empfangen wird.

- großen und variablen Übertragungsverzögerungen (Delay beziehungsweise Latenz) sowie
- häufigen und gebündelten (bursty) Paketfehlern.

Gerade dies entspricht jedoch den typischen Eigenschaften eines mobilen Datendienstes. Die Internet Engineering Task Force (IETF) hat daher eine eigene Arbeitsgruppe (Workgroup = WG) mit der Bezeichnung „Performance Implications of Link Characteristics (PILC)“ eingerichtet, um Empfehlungen zur Optimierung beziehungsweise Modifikation von IETF-Protokollen (z. B. TCP) in Netzwerkumgebungen mit problematischen Übertragungseigenschaften zu erarbeiten. Einige dieser Empfehlungen werden in diesem Beitrag vorgestellt.

2 Eigenschaften der mobilen Datenübertragung mit GPRS

Auf Grund besonderer Störeinflüsse (Mehrwegeausbreitung[4], Abschattungen, Rauschen, Doppler-Effekt, Interferenzen, Wettereinflüsse, Mobilität) ist die Datenübertragung in Mobilfunknetzen grundsätzlich sehr fehleranfällig. Hieraus ergibt sich in der Regel eine höhere Bitfehlerrate[5] (Bit Error Rate = BER) als in terrestrischen Netzen mit häufigen und abrupten Schwankungen der Übertragungsqualität. Daher kommen auf der Funkverbindung (Radio Link Layer, Schicht 2) in der Regel zusätzliche Fehlerkorrekturverfahren wie

- Forward Error Correction[6] (FEC) und
- Automatic Repeat Request[7] (ARQ) zur Anwendung.

Diese Aufgabe wird bei leitungsvermittelten Diensten in GSM-Netzen (Global System for Mobile Communications) vom Radio Link Protocol[8] (RLP) wahrgenommen.

Bei GPRS kann ARQ auf zwei Protokollebenen aktiviert werden, einerseits auf LLC[9]-Ebene (Logical Link Control) zwischen Mobilstation und SGSN (Serving GPRS Support Node) und andererseits auf der Ebene der Luftschnittstelle zwischen Mobilstation und Base Station Controller (BSC) in der RLC[10]-Schicht (Radio Link Control). In beiden Fällen spricht man jeweils von „Acknowledged Mode“. Da ARQ auf LLC-Ebene ähnlich wie die Fehlersicherung im TCP bei Übertragungsfehlern in der Regel ganze IP-Pakete wiederholt – vorausgesetzt die IP-Paketgröße passt in das Informationsfeld eines LLC-Frame – und damit eine Funktion erfüllt, die TCP ohnehin Ende-zu-Ende wahrnimmt, sollte ARQ in diesem Fall lediglich auf RLC-Ebene aktiviert werden.

Das RLC zerlegt (Segmentation) zu sendende LLC-Frames in einzelne Radio-Blocks und setzt die LLC-Frames auf der Empfangsseite wieder korrekt zusammen (Desegmentation). Verloren gegangene Blöcke werden durch ARQ von der Gegenstelle erneut angefordert. Dieses Verfahren wird durch einen Counter für die maximale Anzahl von Wiederholungsanfragen oder durch einen Timer gesteuert, der angibt, wie lange auf die Quittierung eines übertragenen Radio-Blocks maximal gewartet wird, bevor der betreffende Block erneut übertragen wird.

Durch den Einsatz von FEC und ARQ kann somit zwar die Wahrscheinlichkeit unentdeckter Paketfehler reduziert werden, andererseits verursachen diese Verfahren jedoch in einem verrauschten Kanal auf Kosten der knappen Übertragungskapazität zusätzlichen Datenverkehr. Außerdem steigt auf Grund der mehrfachen Übertragung gestörter Radio-Blocks oder LLC-Frames die Übertragungsverzögerung (Delay, Latenz) sowie deren Varianz (Jitter[11]) im gleichen Maße an.

Typisch für die mobile Datenübertragung sind auf Grund der besonderen Störeinflüsse zudem häufige Schwankungen der Funkversorgung. Der Träger kann dabei auf Grund der Mobilität des Nutzers oder plötzlicher Störeinflüsse kurzzeitig für wenige Sekunden verloren gehen. Dies ist besonders für einen File-Transfer kritisch, weil ohne besondere Maßnahmen nach einem Verbindungsabbruch die gesamte Datei neu übertragen werden muss. Eine kurzfristige Verschlechterung der Übertragungseigenschaften des Funkkanals verursacht die für die Funkübertragung typischen Bündelfehler (bursty errors). Diese Bündelfehler können mit Hilfe von ARQ auf RLC-Ebene korrigiert werden. Dabei kommt es jedoch für die Dauer der

Das Thema im Überblick

Die Akzeptanz von GPRS und UMTS im geschäftlichen Umfeld hängt wesentlich von der Performance TCP-basierter Anwendungen ab. Auf Grund der spezifischen Eigenschaften der mobilen Datenübertragung wird die theoretisch maximal zur Verfügung stehende Kanalbandbreite von TCP jedoch nur sehr schlecht genutzt. Das TCP ist besonders empfindlich gegenüber großen und variablen Übertragungsverzögerungen sowie häufigen, gebündelten Paketfehlern. Gerade dies entspricht jedoch den typischen Eigenschaften eines mobilen Datendienstes und führt im Falle eines Paketverlusts dazu, dass TCP den Datendurchsatz plötzlich verringert, um eine vermeintliche Überlast im Netz zu verhindern. Mit Hilfe verschiedener Optimierungsmaßnahmen lassen sich jedoch deutliche Verbesserungen erzielen. Der einfachste Ansatz zur Leistungssteigerung ist, die zu übertragende Datenmenge zu komprimieren. Weitere Maßnahmen sind z. B. Vermeidung von Wiederholungen auf Anwendungsebene, von unnötigem Verkehr, von TCP-Aufbauzeiten durch den Einsatz von „Persistenten TCP-Verbindungen“ oder Vermeidung von Wartezeiten beim HTML-Seitenaufbau durch „Request-Pipelining“.

---

[4] Siehe hierzu auch den Beitrag „Funkdienste und Frequenzmanagement der Deutschen Telekom“, Unterrichtsblätter Nr. 7/1998, S. 308–327.

[5] Bei binären digitalen Übertragungen das Verhältnis der Gesamtzahl der Bitfehler und der Anzahl insgesamt empfangener Bits bei gleicher Zeitbasis.

[6] Dt. Vorwärts-Fehlerkorrektur. Durch Hinzufügen von Redundanz wird ein Datenpaket mit einer Frame Check Sequence (FCS) geschätzt, die sich aus den zu schätzenden Datenbits ergibt.

[7] Normalform eines Sicherungsprotokolls, bei der die gesicherte Datenübertragung auf der blockweisen Fehlerüberwachung beruht.

[8] Allgemein für fehlertolerantes Sicherungsprotokoll der Luftschnittstelle eines Mobilfunksystems. Beispiel: RLP an der Luftschnittstelle der digitalen Mobilfunksysteme GSM und UMTS.

[9] Dt. Logische Verbindungskontrolle. Kommunikationsprotokoll der Sicherungsschicht (LLC-Teilschicht) für IEEE-802-LAN, das der nachfolgenden Vermittlungsschicht Übertragungsdienste für den Austausch von Dateneinheiten zur Verfügung stellt.

[10] In der Protokollarchitektur des Mobilfunksystems der dritten Generation UMTS eine Protokollschicht, die – aufsetzend auf der MAC-Schicht (Media Access Control) – für die Übertragungssteuerung und -sicherung auf den logischen Kanälen zuständig ist.

[11] Jitter: Weitgehend zufallsbestimmte Schwankungen der Flanken eines realen Datensignals um die Sollzeit des Nulldurchganges.

jeweiligen Störung zu einem ebenso sprunghaften Anstieg des Delay. Das Block-Interleaving[12] wie bei leitungsvermittelten GSM-Datendiensten kommt bei GPRS nicht zur Anwendung.

Die besondere Eigenschaft von GPRS, nämlich die dynamische, bedarfsorientierte Kanalzuweisung, die eine effizientere Nutzung der knappen Frequenzressourcen ermöglicht, führt bei gebündeltem (bursty) Verkehr auf Grund des Zuteilungsverfahrens zu einem weiteren Delay-Anteil. Zeitschlitze, die eine Mobilstation zum Senden nutzen möchte, müssen zuvor vom GPRS-Netz beziehungsweise BSC durch die Mobilstation ausdrücklich angefordert und der Mobilstation zugeteilt werden. Dabei kann die Mobilstation bei der Anmeldung gleich mehrere Zeitschlitze reservieren (Dynamic Allocation) und damit diesen Delay-Anteil reduzieren.

Ein weiteres Merkmal mobiler Datenübertragung ist neben dem vergleichsweise hohen Delay die geringe Übertragungsbandbreite. Nach RFC 2757 (Request for Comments) werden solche Übertragungsnetze daher als „Long Thin Networks" bezeichnet: „Long" wegen des hohen Delay und „Thin" wegen der Schmalbandigkeit. Obwohl mit der Einführung von GPRS deutliche Verbesserungen erzielt werden konnten, wobei Bandbreiten im Bereich der Übertragungsgeschwindigkeit von ISDN erzielbar sind, bleiben die Übertragungsraten im Vergleich zu einer reinen LAN-Umgebung oder zu einem festnetzbasierten Weitverkehrsnetz (z. B. Asynchronous Transfer Mode = ATM[13], Frame Relay[14]) auch weiterhin deutlich zurück. Dies hat Auswirkungen auf die Lauffähigkeit von Anwendungen, die für reine LAN-Umgebungen beziehungsweise Weitverkehrsnetze mit geringem Delay und höherer Bandbreite erstellt wurden und unverändert über einen mobilen Datendienst wie GPRS genutzt werden sollen.

Als Besonderheit kommen bei GPRS spezielle Kanalcodierungen (CS = Coding Schemes) zum Einsatz, die abhängig von der jeweiligen Übertragungsgüte unterschiedliche Datenraten erlauben. Man unterscheidet vier verschiedene Codierverfahren, zwischen denen je nach Feldstärke dynamisch gewechselt werden kann. Derzeit sind in den GSM-Netzen CS-1 und CS-2 verfügbar:

- CS-1: 9,05 kbit/s,
- CS-2: 13,4 kbit/s,
- CS-3: 15,6 kbit/s,
- CS-4: 21,4 kbit/s.

Bei den Datenraten der CS handelt es sich um Bruttoraten pro Time-Slot auf der Ebene unterhalb des Radio Link Layer einschließlich des Overhead[15] aller höher liegenden Protokolle. Die auf Anwendungsebene effektiv nutzbare Bandbreite hängt bei idealen Bedingungen lediglich von der jeweiligen Kanalkodierung, der IP-Paketgröße sowie der Kanalbündelung des genutzten Endgerätes ab. Der GPRS-Standard definiert eine Kanalbündelung von maximal acht Time-Slots je Endgerät. Derzeit verfügbare Endgeräte erlauben im Downlink eine Kanalbündelung von zwei bis vier Time-Slots.

Bei einer Kanalbündelung von zwei Time-Slots ergibt sich beispielsweise für CS-2 auf Anwendungsebene eine maximale theoretische Bandbreite von etwa 20 kbit/s. Die Übertragung eines IP-Pakets mit einer für LAN-Umgebungen typi-

---

[12] Von engl. interleaving = verschachteln, verwürfeln. Insbesondere bei der Funkübertragung benutztes Verfahren zur digitalen Kanalcodierung, bei dem zu sendende Datenblöcke zunächst gesammelt und dann ineinander in neue Datenblöcke verschachtelt (interleaved) werden.

[13] Siehe hierzu den Beitrag „ATM – Kommunikationstechnologie der Zukunft", Unterrichtsblätter Nr. 10/1998, S. 492 bis 519, Nr. 2/1999, S. 92–115, Nr. 4/1999, S. 230–241.

[14] Frame Relay: von engl. frame = Rahmen und to relay = weiterleiten; Bezeichnung für ein paketorientiertes Übertragungsprotokoll für Punkt-zu-Punkt-Verbindungen. Frame Relay arbeitet auf Schicht 1 und 2 des OSI-(Open Systems Interconnection-)Referenzmodells und ist auch für Breitbandanwendungen geeignet.

[15] Overhead: Bezeichnung für alle Paketteile, die keine Nutzdaten enthalten.

Verwendete Abkürzungen

ACK	Acknowledgement
ARQ	Automatic Repeat Request
ATM	Asynchronous Transfer Mode
BER	Bit Error Rate
BSC	Base Station Controller
BTS	Base Transceiver Station
CS	Coding Schemes
DFÜ	Datenfernübertragung
Dial In	Einwahl
FCS	Frame Check Sequence
FEC	Forward Error Correction
FTP	File Transfer Protocol
GNUZip	Datenkompressionsprogramm
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IP	Internet Protocol
ISDN	Integrated Services Digital Network
ITU-T	International Telecommunications Union/ Telecommunication Sector
LAN	Local Area Network
LLC	Logical Link Control
MAC	Media Access Control
MTU	Maximum Transfer Unit
PILC	Performance Implications of Link Characteristics
PDP	Packet Data Protocol
PKZip	Datenkompressionsprogramm
PLC	Packet Loss Concealment
PPP	Point-to-Point Protocol
RFC	Request for Comments
RLC	Radio Link Control
RLP	Radio Link Protocol
RTO	Retransmission Timeout
RTT	Roundtrip Time
SACK	Selective Acknowledgement
SCM	Supply Chain Management
SGSN	Serving GPRS Support Node
TCP	Transmission Control Protocol
UMTS	Universal Mobile Telecommunications System
VAD	Voice Activity Detection
VoIP	Voice over Internet Protocol
WG	Workgroup

schen Paketlänge von 1500 Byte verursacht bei der Übertragung über GPRS (2 Slot, CS-2) alleine auf Grund der Bandbreite ein Queueing-Delay von etwa 500 ms bis 600 ms (Millisekunden). Dieser Wert ist für Echtzeit- oder bestimmte Dialog-Anwendungen bereits zu hoch.

3 Anforderungen an ein Übertragungsnetz aus Anwendungssicht

Nicht alle Anwendungen stellen die gleichen Anforderungen an die Qualität eines Übertragungsnetzes. Unternehmenskritische Anwendungen wie beispielsweise Supply Chain Management[16] (SCM) oder Anwendungen im Finanzbereich erfordern eine hohe Verfügbarkeit sowie ein hohes Maß an Sicherheit in Bezug auf Authentizität, Vertraulichkeit und Manipulationsschutz. Demgegenüber ist für eine E-Mail-Anwendung eine „Best-Effort Performance[17]“ wie im Internet bereits zu tolerieren. Das Bild 1 zeigt qualitative Anforderungen beispielhafter Anwendungen.

	Anforderungen					
	Bandbreite			Delay-Sensitiv		
	hoch	mittel	gering	hoch	mittel	gering
SAP R/3			■	■		
E-Mail		■				■
WWW-Zugang		■			■	
File Transfer	■					■
Voice over IP			■	■		

Unterrichtsblätter

IP Internet Protocol WWW World Wide Web

Bild 1: Qualitative Anforderungen beispielhafter Anwendungen

Um bei der Vielzahl unterschiedlichster Anwendungen dennoch einheitliche Anforderungen herausarbeiten zu können, ist es zunächst erforderlich, weitestgehend homogene Anwendungsklassen zu definieren und dann für diese Klassen die jeweiligen Anforderungen zusammenzustellen.

3.1 Netz-Services

Dies sind Dienste, die zur Organisation des Netzes benötigt werden. Sie werden typischerweise bei der Netzanmeldung beziehungsweise zu Beginn einer Session einmalig in Anspruch genommen (beispielsweise IP-Adresszuweisung, Netzanmeldung oder Namensauflösung).

3.2 Dialog-Anwendungen

Diese Anwendungen sind auf Grund der Interaktion Mensch-Maschine durch häufige kleinere Transaktionen charakterisiert. Typische Beispiele sind Datenbankabfragen, SAP- oder Terminal-Emulationen von Großrechnern, aber auch das Web-Browsing (Hypertext Transfer Protocol = HTTP), das von seiner Übertragungscharakteristik zwischen Dialog-Anwendung und File-Transfer einzuordnen ist. Kritisch sind bei dieser Anwendungsklasse insbesondere das Antwortzeitverhalten und die Netzstabilität. Größere Datenmengen werden seltener übertragen.

3.3 Groupware/Messaging/Workflow

Typisch für diese Anwendungsklasse ist das asynchrone Kommunikationsmuster. Die Nachrichten werden auf dem Weg vom Sender zum Empfänger zwischengespeichert. Die zeitliche Verfügbarkeit der Netzverbindung sowie das Delay sind aus Anwendungssicht in der Regel eher unkritisch. Die typischen Vertreter dieser Klasse sind Groupware[18]-Anwendungen wie z. B. E-Mail (ohne Attachments).

3.4 File-Transfer

Beim File-Transfer werden Dateien zwischen Server und Client übertragen. Er stellt hohe Anforderungen in Bezug auf Bandbreite und Netzstabilität. Die Übertragungsverzögerung ist eher unkritisch.

3.5 Steuerung (Logistik/Telemetrie)

Diese Anwendungsklasse ist durch ein sehr geringes, sporadisches Verkehrsaufkommen gekennzeichnet. Beispiele sind bestimmte Telemetrie-Anwendungen oder Alarmüberwachung. Die Anforderungen an Bandbreite und Delay sind gering. Die Mobilität, Zuverlässigkeit und Verfügbarkeit stehen dabei im Vordergrund.

3.6 Echtzeit-Anwendungen

Diese Anwendungsklasse erfordert ein geringes Delay und eine geringe Delay-Varianz (Jitter). Für Video-Anwendungen werden zudem höhere Bandbreiten benötigt. Echtzeit-Anwendungen sind tolerant gegenüber Paketverlust. Die Netzservices und Steuerungsanwendungen stellen die geringsten Anforderungen an die Übertragungsbandbreite und das Delay, sind jedoch empfindlich gegenüber einer schlechten Dienstverfügbarkeit.

Die höchsten Anforderungen in Bezug auf Übertragungsverzögerung und Paketverlust stellen Echtzeit-Anwendungen wie Video-Conferencing oder Voice over IP[19] (VoIP). Nach ITU-T G.144 (International Telecommunications Union) muss die Mund-zu-Ohr-Verzögerung bei VoIP im Bereich von 150 ms bis 400 ms liegen, um die Qualität des (digitalisierten) Telefonnetzes zu erzielen. Der Anhang zu ITU-T G.113 gibt je nach eingesetztem Sprachcodec[20] gemessen an der Qualität des Telekommunikationsnetzes Grenzwerte für die Paketverlustrate an:

- 10 Prozent (ITU-T G.711 mit PLC[21]),

16 Supply Chain Management: In den Unternehmen der Wirtschaft das Management der Liefer- und Logistikkette (Supply Chain).

17 Best-Effort Performance: Dt. „Leistungsfähigkeit nach bestem Bemühen“. Bezeichnung für Datenübertragungsverfahren, bei denen im Netz mit Hilfe von sporadischen oder periodischen Messungen Leistungs-Reports erstellt werden, wodurch die aktuelle Nutzung und Auslastung des Netzes sich darstellen lässt und zyklisch in der Datenbank abgespeichert werden kann.

18 Groupware: Relativ unspezifische Bezeichnung für Software, die von Arbeitsgruppen (Teamworking) eingesetzt wird und oft mehrere Programme und Kommunikationsdienste umfasst. Sie unterstützt den standortunabhängigen mehrfachen Zugriff auf Objekte und fördert so die innerbetriebliche Kooperation.

19 Siehe hierzu auch den Beitrag „Voice over IP – Standards und Technik“, Unterrichtsblätter Nr. 7/2001, S. 398–408.

20 Codec: Abk. Coder/Decoder; Komponente zur Komprimierung und Dekomprimierung von Daten als Hardware oder Software.

21 PLC: Packet Loss Concealment: Ein Verfahren zur Minimierung des Einflusses von Paketverlust auf die Sprachqualität.

- 3,4 Prozent (ITU-T G.729 (A) + VAD[22]),
- 2,1 Prozent (ITU-T G.723.1 6,3 kbit/s + VAD).

Allgemein wird eine Verlustrate von drei Prozent als Obergrenze akzeptiert.

Ähnlich hohe Anforderungen an das Übertragungs-Delay stellen Dialoganwendungen. Der RFC 1144 definiert als kleinste durch den Menschen wahrnehmbare Reaktionszeit einen Wert von 200 ms. Die Obergrenze ergibt sich für diesen Wert individuell als die Reaktionszeit, die der jeweilige Nutzer beziehungsweise die jeweilige Nutzergruppe maximal bereit ist zu akzeptieren. Delay-Werte von mehreren Sekunden sind bei diesen Anwendungen problematisch.

Die Anwendungsklasse des File-Transfer erfordert in erster Linie eine hohe, weitgehend konstante Übertragungsrate und eine stabile Verbindung, sie ist jedoch andererseits sehr unempfindlich gegenüber Delay und Jitter.

4 Grundlagen und Eigenschaften von TCP/IP

Die Protokollfamilie TCP/IP wurde ursprünglich für reine LAN-Umgebungen und terrestrische Weitverkehrsnetze mit vergleichsweise hoher Übertragungsbandbreite und geringem Delay entwickelt. Leitungsgebundene Netze zeichnen sich im Gegensatz zu Mobilfunknetzen zudem durch eine hohe Zuverlässigkeit und Stabilität in Bezug auf Paketverlust und Delay aus. Gerade diese Eigenschaft war eine grundlegende Design-Voraussetzung für den Fluss-Steuerungs-Mechanismus von TCP.

Die Annahme einer stabilen Netzverbindung führt in einer Mobilfunkumgebung dazu, dass die Übertragungsbandbreite auf Grund der Fluss-Steuerung bisweilen nur sehr schlecht genutzt wird. Gerade wenn es auf Grund der Mobilität des Nutzers zu häufigen Störeinflüssen kommt, ist dieser Effekt spürbar. In einem rein stationären Umfeld des Nutzers arbeitet TCP in der Regel zufrieden stellend.

Das TCP ist ein verbindungsorientiertes Ende-zu-Ende-Übertragungsprotokoll, das einen gesicherten Übertragungsdienst (Reliable Stream Delivery[23]) über verschiedene Übertragungsnetze zur Verfügung stellt. Verbindungen über TCP werden mit Hilfe des „Three-Way-Handshake"-Algorithmus zwischen zwei als Port bezeichneten Endpunkten initiiert. Ein TCP-Port ist dabei eine Nummer auf dem initiierenden Rechner beziehungsweise dem Zielrechner und bezeichnet dort jeweils eindeutig eine Anwendung.

Der „Three-Way-Handshake"-Algorithmus umfasst drei Nachrichten, mit denen Sender und Empfänger die Startparameter einer Verbindung wie

- Port Number und
- Initial Sequence Number

vereinbaren. Daher dauert ein TCP-Verbindungsaufbau ungefähr dreimal so lange wie das augenblickliche Delay. Dies kann z. B. bei Dialog-Anwendungen in einer Mobilfunkumgebung ebenfalls bereits kritisch sein.

Zur Übertragungssicherung sind in TCP im Wesentlichen folgende Mechanismen definiert:

- Paketfehlererkennung anhand einer Checksum (16 bit),
- Positive Quittierung empfangener Pakete (Positive Acknowledgement),
- Erkennen von Duplikaten anhand der Sequence-Number,
- Wiederholungen (Retransmission) bei Zeitüberschreitung (Timeout) der jeweiligen Quittung (ACK = Acknowledgement) oder beim Erkennen von Duplicate Acknowledgements,
- Sliding Window Technique,
- Congestion Avoidance and Slow Start,
- Selective Acknowledgement (SACK) Option.

Die Quittierung empfangener Segmente wird von TCP dadurch realisiert, dass der Empfänger in seinem Acknowledgement mit Hilfe der Acknowledgement Number stets anzeigt, welches Segment er als nächstes erwartet. Hierdurch werden implizit alle vorangegangenen Segmente quittiert. Gleichzeitig kann der Sender anhand von duplizierten Acknowledgements mit identischer ACK-Number erkennen, ob Übertragungsfehler aufgetreten sind.

Da ein einfacher Quittierungs-Mechanismus in den meisten Fällen dazu führt, dass die zur Verfügung stehende Übertragungsbandbreite nur sehr schlecht genutzt würde, arbeitet TCP mit der so genannten „Sliding Window Technique". Dabei werden hintereinander gleich mehrere Segmente übertragen, ohne dass eine Quittung empfangen werden muss. Die Größe des Senderfensters gibt dabei an, wie viele dieser Segmente maximal übertragen werden, bevor die nächste Quittung durch den Empfänger vorliegt.

Da TCP als Ende-zu-Ende-Protokoll über viele verschiedene Übertragungsnetzwerke definiert ist, wird die Fenstergröße des Senders von TCP dynamisch an die augenblickliche Übertragungsqualität angepasst. Dieser Mechanismus dient dazu, schnell Überlast-Situationen zu erkennen und entsprechend zu reagieren und wird daher als „Congestion Avoidance and Slow Start" bezeichnet. Das TCP verwaltet hierzu zwei Fenster, das

- Congestion Window, das vom Sender zur Fluss-Steuerung genutzt wird, sowie
- Advertised Receiving Window, mit dessen Hilfe der Empfänger angibt, wie viele Segmente er maximal hintereinander verarbeiten kann, bevor der Sender warten muss, bis er das nächste ACK empfangen kann.

Zu Beginn einer Verbindung sowie nach einer Überlast-Situation beträgt die Größe des Congestion Window des Senders (Initial Window Size) in der Regel ein Segment. Danach befindet sich TCP zunächst in der Phase des Slow Start. In diesem Modus wird das Congestion Window für jedes implizit quittierte Segment um ein Segment erhöht und wächst somit exponentiell an. Sobald die Größe des Congestion Window einen bestimmten Schwellwert, den so genannten „Slow Start Threshold", erreicht hat, wechselt TCP in den Modus des Congestion Avoidance. In dieser Phase wird das Congestion Window mit jedem ACK nur noch um eins erhöht. Das Congestion Window kann jedoch niemals über das vom Empfänger vorgeschlagene (Advertised Receiving Window) hinaus anwachsen. Eine

[22] VAD: Abk. Voice Activity Detection. Oberbegriff für Verfahren zur Erkennung und effizienten Ausnutzung von Sprachpausen bei der Übertragung von Sprachsignalen. Mit dem Verfahren können die „ungenutzten" Übertragungsleistungen während der Sprachpausen anderen Kommunikationsbeziehungen oder anderen Anwendungen zugewiesen werden.

[23] Reliable Stream Delivery: Dt. zuverlässige Übertragung eines Datenstroms mit Hilfe der Fehlersicherung des Transportprotokolls.

Überlast (Congestion) wird auf zwei Arten detektiert (festgestellt), durch

- den Empfang von drei Duplicate Acknowledgements oder
- das Retransmission Timeout (RTO-Timer).

In beiden Fällen wird der Slow Start Threshold auf die Hälfte seines aktuellen Werts eingestellt. Im zweiten Fall wird zusätzlich das Congestion Window auf den Wert 1 Segment zurückgesetzt. Mit diesem Verfahren kann sehr schnell auf eine Überlast reagiert werden. Dies ist entscheidend, um einen Zusammenbruch des gesamten Netzes zu vermeiden. Im ersten Fall, das heißt in der Regel nach dem Empfang von drei Duplicate Acknowledgements, wird nicht in den Slow-Start-Modus, sondern in den Congestion-Avoidance-Modus umgeschaltet. Dieses Verfahren nennt man „Fast Retransmit with Fast Recovery“. Das Congestion Window wird dann also auch nicht auf den Wert von einem Segment, sondern auf den Wert des Slow Start Threshold + 3 Segmentgrößen eingestellt. Das verlorene Segment wird anschließend erneut übertragen.

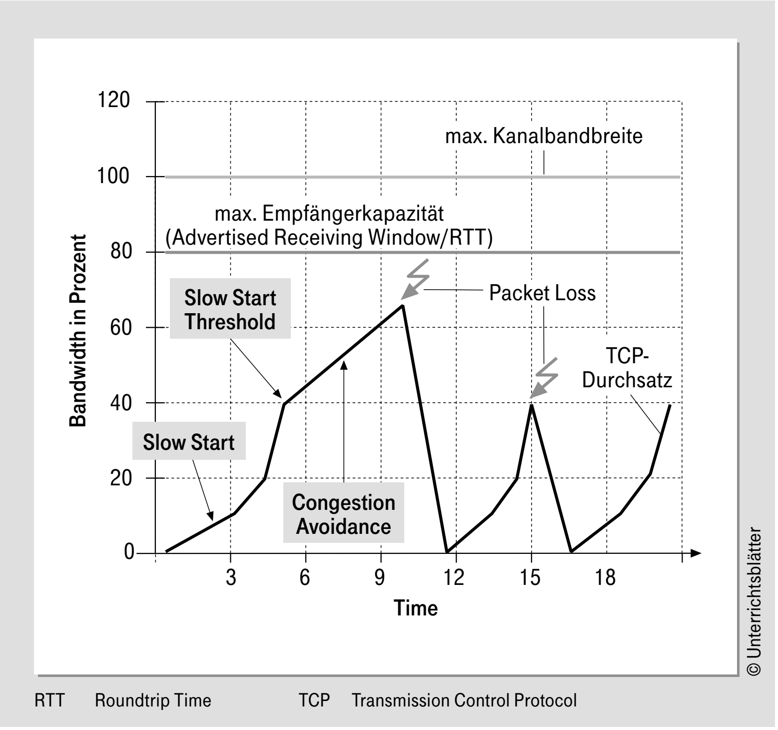

Bild 2: Ausnutzung der Bandbreite auf Grund der Fluss-Steuerung von TCP

Für jedes weitere empfangene Duplicate ACK wird das Congestion Window um die Größe eines weiteren Segments erhöht. Trifft nun wieder ein reguläres ACK ein, wird das Congestion Window genau auf den Wert des Slow Start Thresholds eingestellt. Dieses empfangene ACK sollte nun die Quittung des erneut übertragenen Segments sein. Dieses ACK wird zusätzlich alle empfangenen Datenpakete zwischen dem verlorenen Datenpaket und den drei Duplicate ACK quittieren.

Die Zeit zur Ermittlung dieses Timeout wird anhand der fortlaufend gemessenen Roundtrip Time (RTT) über eine gewichtete Mittelwertbildung bestimmt. Dabei wird jeweils die Zeit zwischen dem Versenden eines Segments und dem Empfang der zugehörigen Empfangsbestätigung gemessen und zur Anpassung des Retransmission Timer herangezogen.

In einer Mobilfunkumgebung, bei der sich die Übertragungsgüte sowie das Delay auf Grund kurzfristiger Störeinflüsse auch mehrmals während einer TCP-Verbindung plötzlich ändern kann, führen diese adaptiven[24] Mechanismen zur Fluss-Steuerung beziehungsweise zur Steuerung von Wiederholungen (Retransmission) häufig zu einer sehr schlechten Ausnutzung der zur Verfügung stehenden Bandbreite. Paketverluste werden im Mobilfunk hauptsächlich durch Bitfehler verursacht. Das TCP kann jedoch nicht die Ursache des Paketverlusts – Übertragungsfehler oder Netzwerküberlastung – unterscheiden. Es reagiert immer mit Überlastabwehr und reduziert sein Sendefenster. Große Laufzeitvariationen auf Grund von Paketwiederholungen auf LLC- oder RLC-Protokollebene (ARQ-Mechanismen) verursachen bei der Timeout-basierten TCP-Fluss-Steuerung (der Timeout-Wert wird geschätzt) große Leerlaufzeiten beim Sender (Bild 2).

Zu einer weiteren Verschlechterung des TCP-Durchsatzes kann die Strategie der exponentiellen Vergrößerung des RTO (exponentieller Backoff) führen. Wenn eine Paketwiederholung auf Grund eines abgelaufenen Timeout durchgeführt wird, wird der RTO bei TCP jeweils auf das Doppelte des bisherigen Timeout gesetzt. Gehen während einer temporären Störung auf der Funkstrecke mehrere aufeinander folgende Wiederholungen verloren, so wächst der Timeout schnell an. Wenn dann die Störung beendet ist, vergeht so viel Zeit, bis die Übertragung wieder aufgenommen wird.

Der RTO-Wert wird bei vielen Implementierungen mit einem Wert von drei Sekunden initialisiert. Treten auf dem Übertragungsweg regelmäßig längere RTT-Zeiten als drei Sekunden auf, empfiehlt es sich, den Initial-RTO-Wert entsprechend größer zu initialisieren.

Lösungen der genannten Probleme werden derzeit im Rahmen der IETF erarbeitet. Durch den Einsatz entsprechend proprietärer[25] Middleware lassen sich bereits heute einige dieser Probleme entschärfen.

5 Optimierungsmaßnahmen und Empfehlungen

Um den Anforderungen an ein Übertragungsnetz weitestgehend gerecht zu werden und die Performance von TCP in mobilen Datennetzen – wie beispielsweise GPRS – zu verbessern, gibt es eine Reihe von Empfehlungen und Optimierungsmöglichkeiten.

Auf Grund der vergleichsweise geringen Bandbreite eines mobilen Übertragungsdienstes besteht der einfachste Ansatz zur Performance-Steigerung darin, die zu übertragende Datenmenge zu komprimieren. Hier unterscheidet man Verfahren zur Komprimierung der Anwendungsdaten (z. B. HTML-[Hypertext Markup Language-]Seiten) sowie

[24] adaptiv: angepasst.

[25] proprietär: Engl. proprietary. Als Adjektiv verwendet, bezeichnet „proprietär“ in der Informations- und Telekommunikationstechnik in der Regel herstellergeprägte Entwicklungen, die keine Rücksicht auf Standardisierungen nehmen.

Header[26]-Komprimierungsverfahren, die den TCP- beziehungsweise den IP-Header von insgesamt 40 Byte auf 3 Byte bis 5 Byte verringern und somit den Overhead reduzieren. Die IETF empfiehlt hierzu folgende Verfahren:

- Header Compression nach van Jacobsen RFC 1144,
- Payload[27] Compression entsprechend RFC 2393 (Framework).

Eine Header-Kompression nach RFC 1144 kann beispielsweise optional im Windows-DFÜ[28]-Netzwerk aktiviert werden. Da die PPP-Verbindung (Point-to-Point-Protocol) bei GPRS bereits vom GSM-Endgerät terminiert wird, muss die Header Compression beim PDP[29]-Kontext-Aufbau durch das GSM-Endgerät explizit vom GPRS-Netz angefordert werden. Dabei kommt ebenfalls das Verfahren nach RFC 1144 zur Anwendung.

Als Payload-Komprimierungsverfahren lassen sich bewährte verlustlose Datenkompressions-Algorithmen verwenden wie z. B. in PKZip oder GNUZip[30]. Hier ist stets abzuwägen zwischen Kompressionsgüte und Kompressionsgeschwindigkeit, weil es sich bei diesem Einsatzbereich um eine Echtzeitumgebung handelt. Der Algorithmus GNUZip wird beispielsweise von allen Internet-Browsern ab HTTP Version 1.1 unterstützt, wobei die Steuerung durch einen Eintrag im HTTP-Entity-Header (Content-Encoding) vorgenommen wird. Darüber hinaus gibt es die Möglichkeit, verlustbehaftete Algorithmen oder Filter einzusetzen. Beispielsweise kann man die Farbauflösung von Bildern in HTML-Seiten herabsetzen oder unnötige Banner-Werbung oder Kommentare herausfiltern.

Um den Protocol-Overhead weiter zu reduzieren, sollte die maximale Segmentgröße (Maximum Transfer Unit = MTU) möglichst groß gewählt werden. Hier ist jedoch andererseits zu beachten, dass die Paketfehler-Wahrscheinlichkeit sowie das Queueing[31]-Delay mit der Segmentgröße steigt. Durch den Einsatz von ARQ auf RLC-Ebene lässt sich die Paketfehler-Wahrscheinlichkeit zwar reduzieren, hierdurch verschlechtern sich jedoch die Delay- beziehungsweise die Jitter-Werte ebenfalls proportional zur Segmentgröße. Gerade in Mobilfunknetzen mit kurzfristig hoher Bitfehler-Wahrscheinlichkeit sowie geringer Bandbreite ergibt sich hieraus eine Begrenzung für den Wert der MTU. Die IETF WG PILC empfiehlt, für die MTU einen Wert zu wählen, der einem Übertragungs-Delay von maximal 200 ms entspricht.

Beispiel:
GPRS, 2 Slot, CS-2: 20 kbit/s · 200 ms = 512 Byte

[26] Header: Kopfteil vorzugsweise asynchroner Übertragungsformate (Blöcke, Zellen, Pakete) mit administrativen Einträgen, wie Adressen, Prioritäten, Nachrichtentyp.

[27] Payload: In Kommunikationsprotokollen allgemein für den Inhalt (z. B. Nutz-/Anwenderdaten) von Übertragungsrahmen (Block, Zelle, Paket).

[28] DFÜ: Datenfernübertragung.

[29] PDP: Abk. Packet Data Protocol.

[30] Komprimierprogramme für Dateien.

[31] Queueing: Allgemein für zeitliche Verzögerung in Wartespeichern. In ATM-Systemen beispielsweise ist damit die Verzögerung gemeint, die entsteht, wenn eine Zelle im System zwischengespeichert werden muss, weil die Ressourcen für den Weitertransport der Zelle nicht vorhanden sind. Die Ursachen dafür können Überlastung des Links oder Zellen einer Verbindung höherer Priorität sein.

Glossar

Verbindungsaufbauzeit
Bei verbindungsorientierten Diensten die Zeit zwischen der Verbindungsanforderung und dem Zustandekommen der Verbindung.

Ende-zu-Ende-Verzögerung (Delay, Latenz)
Die Laufzeit für die Übertragung eines Datenpakets vom Netzzugangspunkt bis zum Ziel. Das Delay hängt von der Größe des Datenpakets und von der aktuellen Netzlast ab. Das Ende-zu-Ende-Delay setzt sich aus folgenden Anteilen zusammen:

- Queueing-Delay: Am Netzzugangspunkt sowie in jedem Netzknoten wird das Datenpaket in einer Warteschlange zwischengespeichert und danach bitweise übertragen.
- Processing-Delay: In jedem Netzknoten wird anhand einer Zieladresse beziehungsweise eines Identifiers in einer Forwarding-Tabelle nach dem Next-Hob gesucht.
- Link-Delay: Das Link-Delay bezeichnet die Verzögerung bei der Übertragung eines Datenpaketes zwischen zwei Netzknoten. Dieses Delay hängt unmittelbar von der augenblicklichen Qualität des physikalischen Kanals (Schicht 1) und den Fehlerkorrektur-Mechanismen der Sicherungsschicht (Schicht 2) ab.
- Signallaufzeit: Dieses Delay ergibt sich aus der Ausbreitungsgeschwindigkeit elektromagnetischer Wellen und der Entfernung zwischen Sender und Empfänger. Die Signallaufzeit ist bei terrestrischen Verbindungen in der Regel zu vernachlässigen. Bei Satellitenverbindungen verursacht die Signallaufzeit allerdings den größten Delay-Anteil. Die Signallaufzeit bei einer Entfernung von beispielsweise 300 km beträgt nur eine Millisekunde.

Jitter
Streuung des Delay: Das Delay ist keine feste Größe, sondern unterliegt zufälligen, das heißt nichtdeterministischen, Einflüssen.

Bandbreite
Die Bandbreite gibt an, wie viele Datenbits pro Zeiteinheit vom jeweiligen Übertragungsdienst (Link) übertragen werden können.

Stabilität
Die Stabilität eines Übertragungsdienstes gibt an, ob und wie die Bandbreite zeitlichen Schwankungen unterliegt.

Verfügbarkeit
Die Verfügbarkeit eines Übertragungsdienstes gibt an, wie häufig über einen längeren Zeitraum (z. B. ein Jahr) Dienstanforderungen nicht erfüllt werden können.

Paketfehlerrate (Packet Loss)
Die Paketfehlerrate gibt an, wie häufig Datenpakete fehlerhaft übertragen werden oder auf Grund von Überlast verloren gehen.

Für die Wahl der MTU gilt jedoch:
Je kleiner die IP-Segmentgröße gewählt wird, desto größer fällt der durch die IP- und TCP-Header verursachte Protocol-Overhead aus. Je kleiner die IP-Segmentgröße gewählt wird, desto mehr RTT-Zeiten vergehen nach der Auslösung des TCP-Slow-Start-Algorithmus, bis die volle Kanalkapazität auf dem Übertragungsweg wieder erreicht werden kann. Daher empfiehlt T-Mobil für die MTU in den meisten Anwendungsfällen einen Wert von 1500 Byte und damit einen höheren Wert als die WG PILC. Da Paketverluste auf Funkstrecken überwiegend durch Bitfehler verursacht werden, wächst die Wahrscheinlichkeit eines Paketverlustes proportional zu der Länge der Pakete. Im Falle eines Verlustes muss das gesamte Paket über den langsamen Funkkanal wiederholt werden. Mit Hilfe des RLC Acknowledged Mode kann dieser Effekt jedoch weitestgehend kompensiert werden, so dass sich hieraus für die Wahl einer höheren MTU tatsächlich keine Begrenzung ergibt.

Die adaptive Anpassung der Größe des Congestion Window sowie die Reaktion auf Paketverlust sind von großer Bedeutung für die Performance von TCP. Hier sind modifizierte (abgewandelte) Retransmission- und Congestion-Avoidance-Mechanismen erforderlich, um eine spürbare Verbesserung zu erzielen. Hieran wird derzeit im Rahmen der IETF gearbeitet. Proprietäre Middleware-Lösungen sind bereits verfügbar.

Weitere Maßnahmen sind:

- Vermeidung von Retransmission auf Anwendungsebene (z. B. File Transfer Protocol = FTP) auf Grund von plötzlichen Unterbrechungen durch Einsatz einer Middleware.
- Vermeidung von unnötigem Verkehr (z. B. „Keep-Alive-Meldungen“ zwischen Client und Server) durch „Spoofing“ und „Caching“ in der Middleware. Viele Standardanwendungen wurden für eine LAN- beziehungsweise Festnetzumgebung entwickelt und zeichnen sich daher durch ein ausgedehntes Kommunikationsverhalten (z. B. Polling, Übertragung redundanter Daten, ineffiziente Dialoge) aus, das in einer Mobilfunkumgebung zu einer schlechten Performance führt.
- Vermeidung von TCP-Aufbauzeiten durch Einsatz von „Persistenten TCP-Verbindungen“ (beispielsweise ab HTTP Version 1.1),
- Vermeidung von Wartezeiten beim HTML-Seitenaufbau durch „Request-Pipelining“ (beispielsweise ab HTTP Version 1.1).

6 Zusammenfassung und Ausblick

Auf Grund der spezifischen Übertragungseigenschaften der mobilen Datenübertragung ergeben sich deutliche Unterschiede zwischen Mobilfunknetzen und dem Festnetz. Anwendungen und Protokolle wie z. B. TCP, die für eine reine Festnetzumgebung entwickelt wurden, müssen unter Berücksichtigung geringerer Bandbreiten, höherer Delay- und Jitter-Werte sowie einer höheren Paketverlustrate fallweise angepasst werden, um eine zufrieden stellende Performance sicher zu stellen.

Durch verschiedene Maßnahmen lassen sich die genannten Übertragungseigenschaften und deren Auswirkungen auf die Lauffähigkeit von Protokollen und Anwendungen kompensieren. Hierzu sind auf dem Markt bereits eine Vielzahl verschiedener Optimierungslösungen verfügbar, die gerade für Web-basierte Applikationen deutliche Performance-Gewinne ermöglichen.

Die Akzeptanz mobiler Datendienste hängt wesentlich von der Performance und Lauffähigkeit von TCP-basierten Standardanwendungen ab. Die Entwicklung und der Einsatz geeigneter Optimierungs-Software ist daher ein wichtiger Baustein für den Erfolg von GPRS und damit auch für UMTS.

Die Qualitätskenngrößen eines Übertragungsdienstes sind in einem Glossar zusammengestellt.

(Ge)

Weitere Infos zu Mobile IP VPN und GPRS finden Sie unter:
http://www.t-mobile-business-solutions.com/