Transportprotokolle

1. Protocol-Port Konzept
2. Socket Programmierung
3. User Datagram Protocol (UDP)
4. Transmission Control Protocol (TCP)
   1. Verbindungsmanagement
   2. Sicherung
   3. Flußkontrolle
   4. Optimierungen
5. Neuere Entwicklungen

Zum Inhalt

In diesem Abschnitt werden wir uns den Transport-protokollen im Internet widmen. Ihre Wahl und Eigenschaften werden direkt aus der Anwendungs-entwicklung heraus angesprochen. Deshalb spielt die Programmierung des „Socket“-Interfaces hier eine besondere Rolle.

Das zugehörige Kapitel im Tanenbaum ist 6 – aber für die Socket-Programmierung bitte Tanenbaum überspringen. Besser: Herbert Wiese: *Das neue Internetprotokoll IPv6*, Hanser, 2002

1. Protokoll-Port-Konzept

Wie werden Kommunikationsprozesse auf einem Rechner identifiziert?

- Direkte Prozessadressierung ist problematisch:
  - Prozesslogik ist betriebssystemabhängig
  - Ziel der Adressierung ist ein Dienst, nicht der Prozeß
  - Prozesse können mehr als einen Dienst anbieten
- Lösung: Verwendung von abstrakten Protokoll Ports
  - Betriebssystem stellt Spezifikation und Zugriff
  - Protokoll-Software (stack) synchronisiert den Zugriff mittels Warteschlangen und Belegungen

1. Netzwerkports

```
atlantis %>netstat -a (Auszug)

TCP
   Local Address         Remote Address    Swind Send-Q Rwind Recv-Q  State
-------------------- -------------------- ----- ------ ----- ------ -------
      *.*                  *.*                0      0      0      0 IDLE
      *.sunrpc             *.*                0      0      0      0 LISTEN
      *.*                  *.*                0      0      0      0 IDLE
      *.32771              *.*                0      0      0      0 LISTEN
      *.ftp                *.*                0      0      0      0 LISTEN
      *.telnet             *.*                0      0      0      0 LISTEN
      *.shell              *.*                0      0      0      0 LISTEN
      *.login              *.*                0      0      0      0 LISTEN
      *.exec               *.*                0      0      0      0 LISTEN
      *.uucp               *.*                0      0      0      0 LISTEN
      *.finger             *.*                0      0      0      0 LISTEN
      *.time               *.*                0      0      0      0 LISTEN
      *.x-server           *.*                0      0      0      0 LISTEN
   sol01.rz.fhtw-berlin.de.1021 amor.rz.fhtw-berlin.de.nfsd  8760      0  8760
                                                                    0 ESTABLISHED
localhost.32785      localhost.32783      32768      0 32768   0 ESTABLISHED
localhost.32783      localhost.32785      32768      0 32768   0 ESTABLISHED
   sol01.rz.fhtw-berlin.de.1023 merlin.rz.fhtw-berlin.de.login 49152   0  8760
                                                                    0 ESTABLISHED
   sol01.rz.fhtw-berlin.de.x-server merlin.rz.fhtw-berlin.de.25478 49152     0
                                                          8760       0 ESTABLISHED
```

1. Zentrale Portvergabe

Zur Kommunikation miteinander müssen sich zwei Rechner auf Portnummern einigen. Hierfür gibt es eine

- Zentrale Vergabe - universal binding to well known ports festgelegt durch die IANA (port < 1024)

  Bsp:

Service	Portnummer	Protokoll
ftp-data	20	tcp
ftp	21	tcp
telnet	23	tcp
smtp	25	tcp
tftp	69	udp
finger	79	tcp
portmap	111	tcp
portmap	111	udp

- Ports < 1024 sind privilegiert (realisiert in Unix)!

1. Dynamische Portvergabe

- Dynamische Zuordnung - (dynamic binding)
- Wird von Anwendungsprogrammen implementiert
- Portnummern sollten > 1024 sein
- Die Festlegung der Dienste und der zugehörigen Port-Nummern befindet sich in der Datei /etc/services
- Diese kann um eigene Definitionen erweitert werden.

1. Transportprotokolle

- IP adressiert nur Zielrechner, nicht einzelne Programme
- Damit Anwendungsprogramme Datagramme senden und empfangen können, stehen traditionell UDP und TCP als Transportprotokolle bereit (Layer 4)
- Beide verwenden das Prinzip der abstrakten Ports
- Beide können via Sockets programmiert werden

2. Programmierschnittstelle

Als Programmierschnittstelle für Software mit Kommunikations-kanälen haben sich die sog. Berkeley Sockets etabliert:

- Windows: winsock.dll
- Unix: libsocket.so, <sys/socket.h>
- Java: java.net.*

Die beiden Tripel (Internet-Adresse, Protokoll, Port) von Sender und Empfänger bilden die (eindeutigen) Kommunikationseckpunkte

Kommunikationsparameter können mit ‚getsockopt' gelesen und verändert werden

2. Applikationsprogramm-Interface (API)

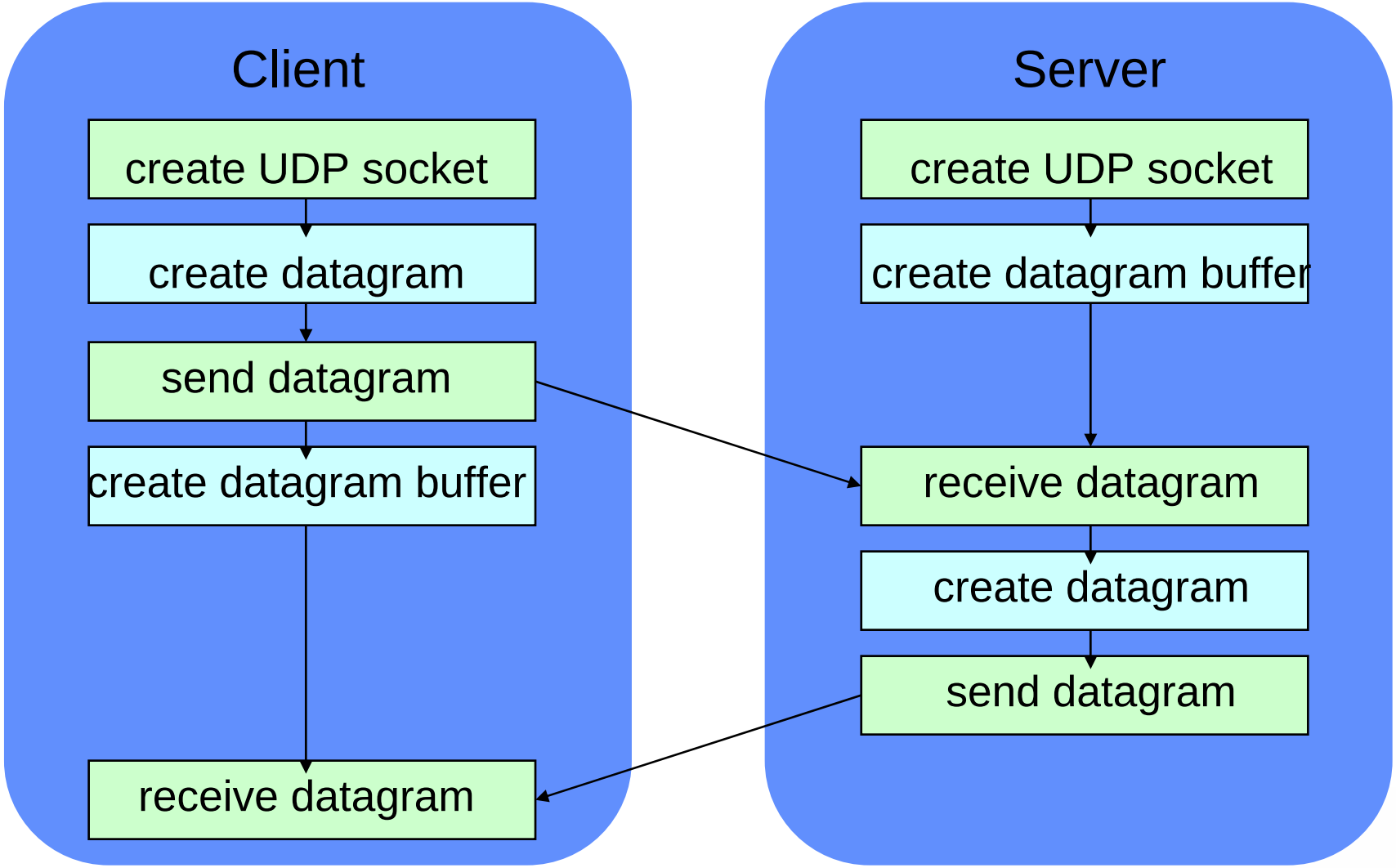
Wie können Anwendungsprogramme einfach in einer Netzwerkschnittstelle lesen und schreiben?

- Zielstellung: Nutzung wie `read` and `write`
- Probleme:
  - Netzwerkschnittstelle komplexer als Filesystem (Protokolle, -arten, Sicherheit)
  - Andersartige Kommunikationsparadigmen (C/S, Message Passing, Broadcast,...)
  - Systemübergreifende Realisierung (Vielfältige OSs, Enkodierungen & Programmiersprachen)
- Lösung: Standardisierte Netzwerk-API zur Bedienung von Standardprotokollen - Berkeley Sockets und SystemV TLI

2. Sockets

- Ursprünglich Netzwerk-API von BSD 4.3 (Unix)
- Seit Jahren am meisten verbreiteter Programmierstandard (C, C++, Java, ...)
- Typen:
  - Stream (`SOCK_STREAM`) für TCP
  - Datagram (`SOCK_DGRAM`) für UDP
  - Raw (`SOCK_RAW`) für IP & ICMP
- Erstellung: `s = socket(domain, type, protocol)`

2. Verbindungslose Kommunikation: Struktur eines UDP-Programms



2. Beispiel: UDP-Client in Java

```
import java.net.*;
import java.io.*;
public class UDPClient{
    public static void main(String args[]){
        // args give message contents and server hostname
          try {
                aSocket = new DatagramSocket();
                byte [] m = args[0].getBytes();
                InetAddress aHost = InetAddress.getByName(args[1]);
                int serverPort = 6789;
                DatagramPacket request = new DatagramPacket(m, args[0].length(),
                                                            aHost, serverPort);
                aSocket.send(request);
                byte[] buffer = new byte[1000];
                DatagramPacket reply = new DatagramPacket(buffer, buffer.length);

                aSocket.receive(reply);
                System.out.println("Reply: " + new String(reply.getData()));
                aSocket.close();
          }catch (SocketException e){System.out.println("Socket: " +
e.getMessage());
          }catch (IOException e){System.out.println("IO: " + e.getMessage());}}
    }
}
```

Sendet Nachricht an Server

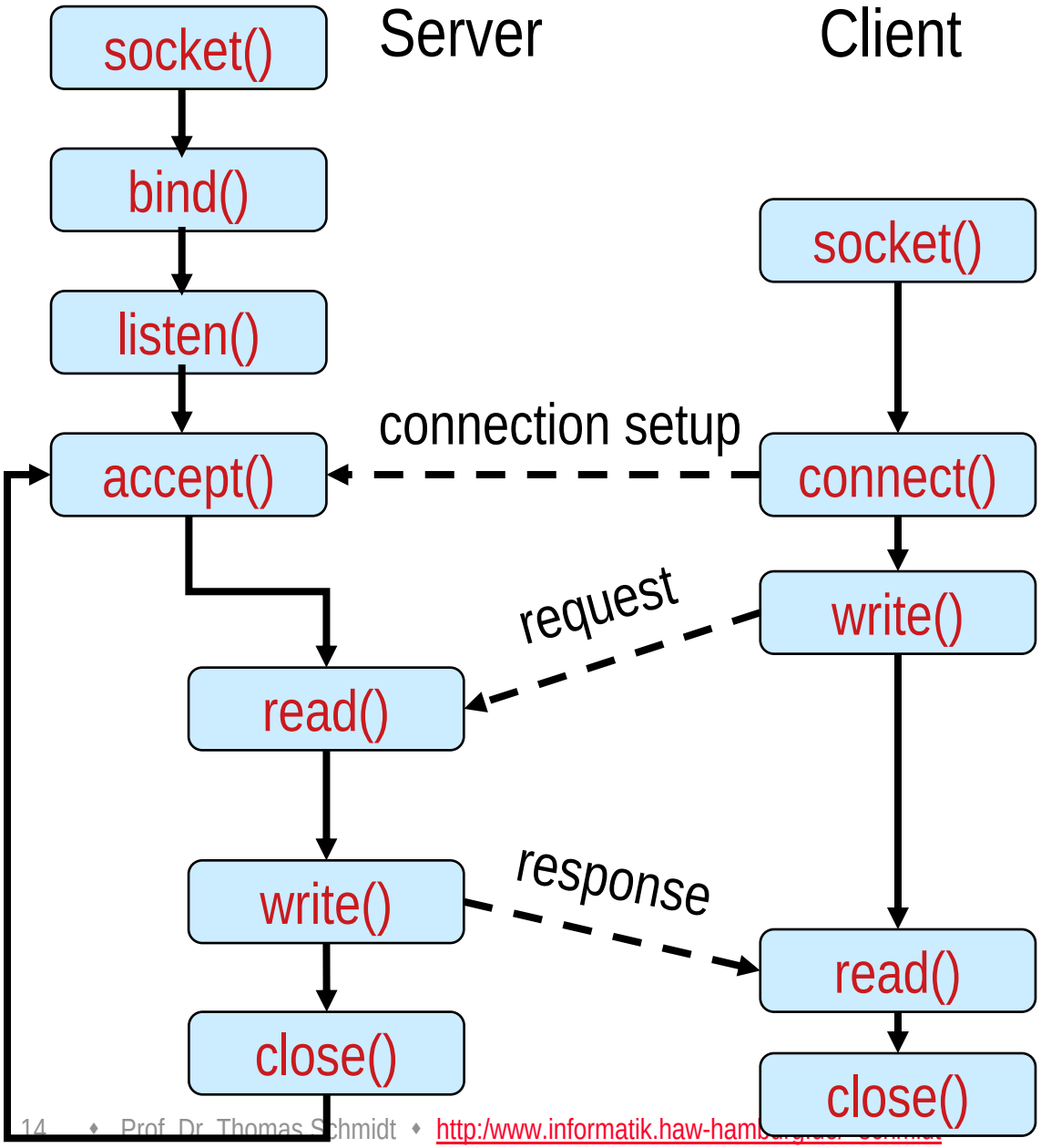
Empfängt Antwort vom Server

2. Beispiel: UDP-Server in Java

```
import java.net.*;
import java.io.*;
public class UDPServer{
        public static void main(String args[]){
            try{
                aSocket = new DatagramSocket(6789);
                byte[] buffer = new byte[1000];
                while(true){
                    DatagramPacket request = new DatagramPacket(buffer, buffer.length);
                    aSocket.receive(request);
                    DatagramPacket reply = new DatagramPacket(request.getData(),
                                                              request.getLength(),
                                                              request.getAddress(),
                                                              request.getPort());
                    aSocket.send(reply);
                }
            }catch (SocketException e){System.out.println("Socket: " + e.getMessage());
            }catch (IOException e) {System.out.println("IO: " + e.getMessage());}
    }
}
```

Empfängt Nachricht von Client

Sendet Antwort an Client



Verbindungs-orientierte Kommunikation: Call-Sequenz für TCP

2. Wichtige Funktionen (C)

- int socket (int af, int type, int protocol);
- int bind (int s, const struct sockaddr *addr, socklen_t addrlen);
- int listen(int s, int backlog);
- int accept(int s, struct sockaddr *addr, socklen_t *addrlen);
- int connect(int s, const struct sockaddr *serv_addr, socklen_t addrlen);
- ssize_t send(int s, const void *buf, size_t len, int flags);
- ssize_t recv(int s, void *buf, size_t len, int flags);
- int close(int s);
- int set/getsockopt(int s, int level, int optname, const void *optval, socklen_t optlen);

2. IPv4 → IPv6 Koexistenz: Versionsübergreifende Adress-API

	IPv4	IPv6	
Data structures	AF_INET	AF_INET6	
	in_addr sockaddr_in	in6_addr sockaddr_in6	
Address conversion functions	inet_aton() inet_addr()	inet_pton()	IPv4 and IPv6 functions
	inet_ntoa()	inet_ntop()	
Name-to-address functions	gethostbyname() gethostbyaddr()	getipnodebyname() getipnodebyaddr getnameinfo() * getaddrinfo() *	

*** POSIX protocol independant functions**

2. Benutzung der Koexistenz-API

Problem: `sockaddr_in` und `sockaddr_in6` sind inkompatible Datenstrukturen ...

Lösung:

- Eine Indirektionsstruktur `addrinfo` erlaubt den transparenten Zugriff auf die (versionsabhängigen) `sockaddr*` Strukturen
- Diese werden automatisch gefüllt durch `getaddrinfo`: liefert im `result` einen Pointer auf eine verkettete Liste von `addrinfo` Adressstrukturen.
- Um einen erfolgreichen Kommunikationsweg zu finden, müssen die Adressen der Liste ausprobiert werden.

2. Programmbeispiele

In der Vorlesung

client.c **server.c**

2. Wichtige Funktionen (JAVA)

Socket (Hostname/InetAddress, Port) /
ServerSocket (Port) / DatagramSocket (Port)

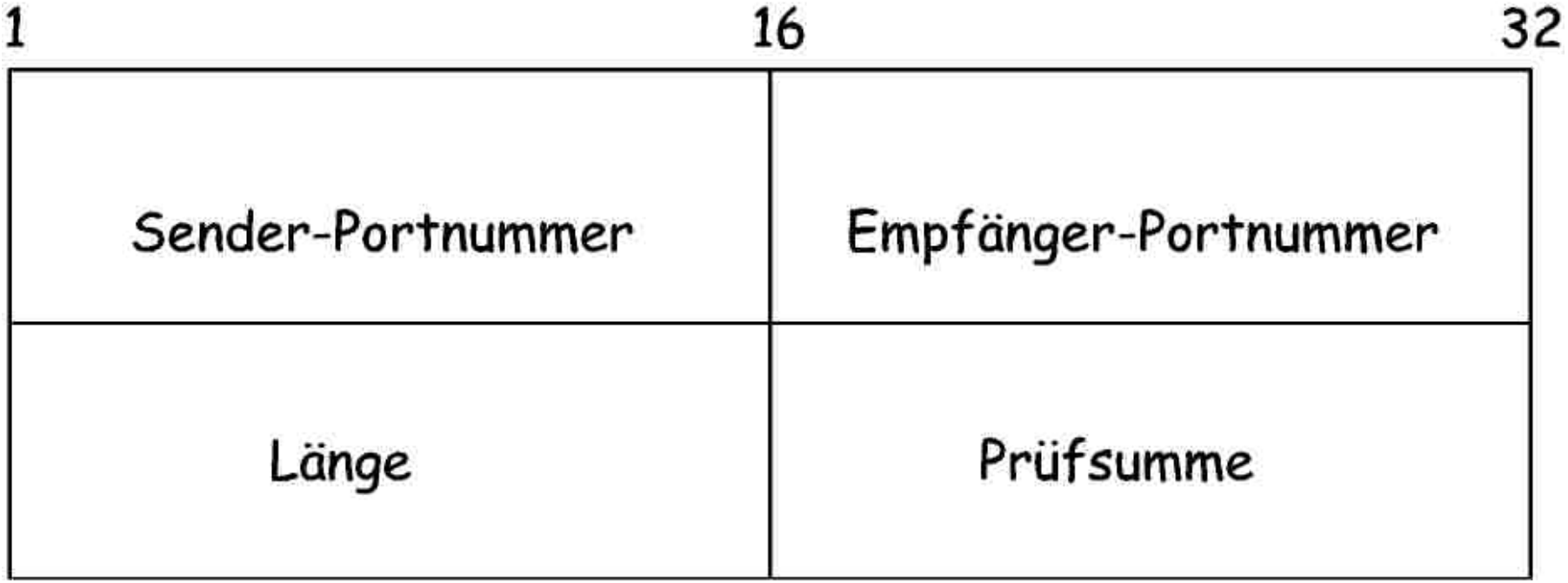
mit den Methoden:

- bind
- connect/accept
- close
- getInputStream/getOutputStream

3. User Datagram Protocol

- UDP (RFC 768) ist ein ungesicherter, verbindungsloser Transportdienst
- Es besitzt eine optionale Checksumme für transferierte Daten
- UDP unterstützt das Multiplexing zwischen verschiedenen Anwendungsprogrammen auf einem Rechner
- UDP besitzt minimalen Overhead
- UDP veranlaßt selbst keine Paketwiederholungen
- UDP kennt keine Mechanismen der Flusskontrolle

3. UDP Datagramm



Länge	Anzahl der Bytes des gesamten Datagramms
Prüfsumme	über Header und Datenteil (optional)

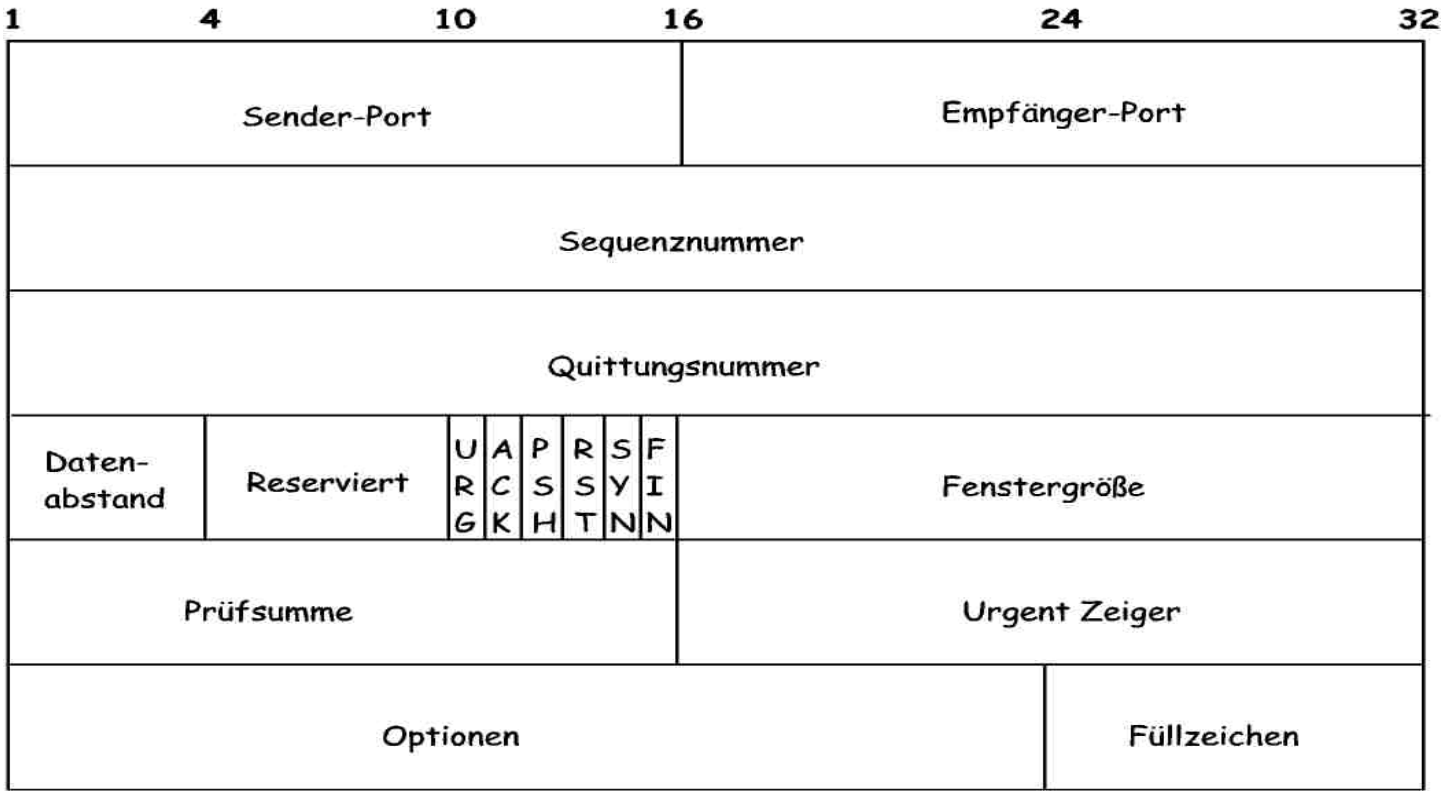
4. Transmission Control Protocol - TCP

- TCP ist der zentrale Transportdienst im Internet
- Spezifiziert in RFC 793
- TCP stellt einen verbindungsorientierten, gesicherten Transferdienst zur Verfügung
- TCP-Pakete heißen Segmente
- TCP sendet Daten als Byte-Strom (stream oriented)
- TCP unterstützt Full Duplex

4. Eigenschaften von TCP

- Virtuelle Verbindung (virtual circuit connection)
  - verdeckt Details zu Verbindungsaufbau und -sicherung
  - erscheint dadurch wie eine direkte Hardware-Verbindung
- Jedes Segment wird (initial) in einem IP-Datagram transportiert. Daraus ergibt sich für die (konfigurierbare) TCP Maximum Segment Size $MSS \leq MTU - 40$ Bytes
- Das Receive-TCP quittiert den Empfang von Segmenten
- Das Send-TCP veranlaßt ein erneutes Versenden, wenn die Quittung ausbleibt

4. TCP Segment



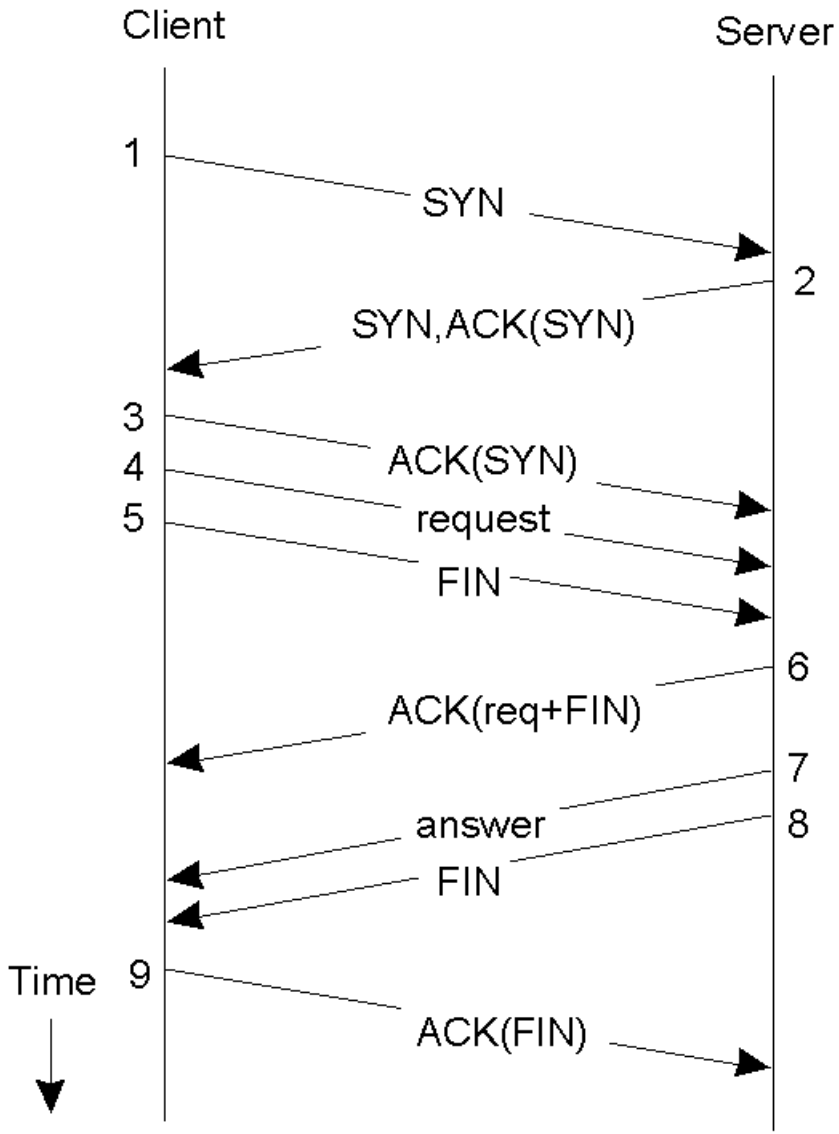
Sequenznummer	Nummer des ersten Bytes in der Sequenz
Quittungsnummer	Nummer des letzten quittierten Segment-Bytes
Datenabstand	Länge des Headers in 32-bit Worten
Fenstergröße	Anzahl der Bytes, die der Empfänger abnimmt
Prüfsumme	über Header und Datenteil (obligatorisch)

4. Coded Bits

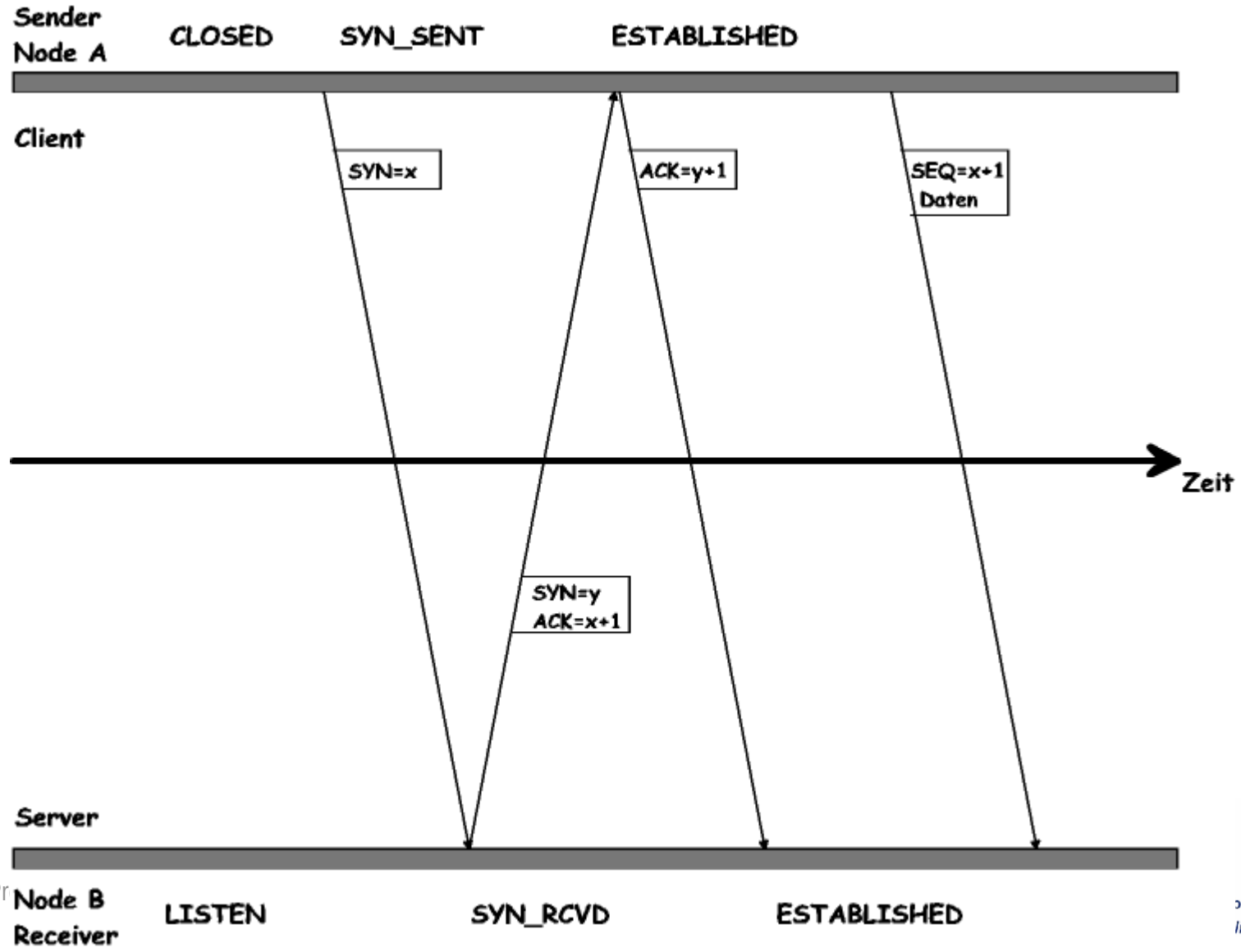
- URG - Urgent Pointer ist gültig: Setzt den Empfänger in ‚Urgent Mode' bis der Urgent Pointer passiert ist.
- ACK - Acknowledgment ist gültig
- PSH - Push Flag: ‚verarbeite Daten sofort'
- RST - Reset: Setze Verbindung zurück
- SYN - Synchronize zur Verbindungsaufnahme
- FIN - Finish zum Verbindungsabbau

4.1 TCP Verbindungsmanagement

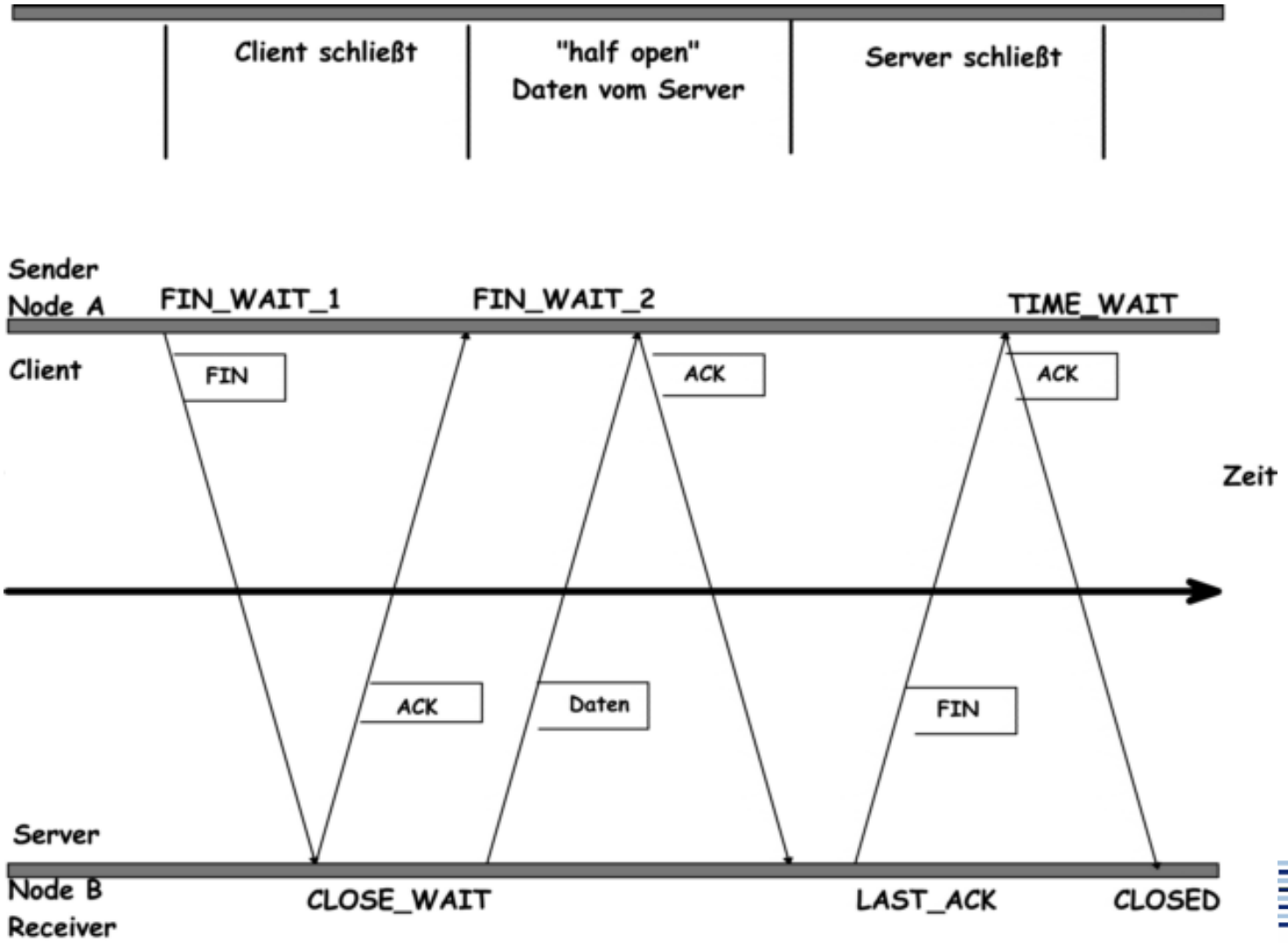
- TCP beinhaltet ein automatisches Verbindungsmanagement
- Realisiert im Client-Server Modell: Client initiiert, Server akzeptiert Verbindung
- Eine Verbindung besteht, wenn Client und Server den zugehörigen Verbindungszustand etabliert haben: Hierzu ist (mindestens) ein Drei-Wege-Handshake erforderlich



4.1 TCP Verbindungsaufbau



4.1 TCP Verbindungsabbau



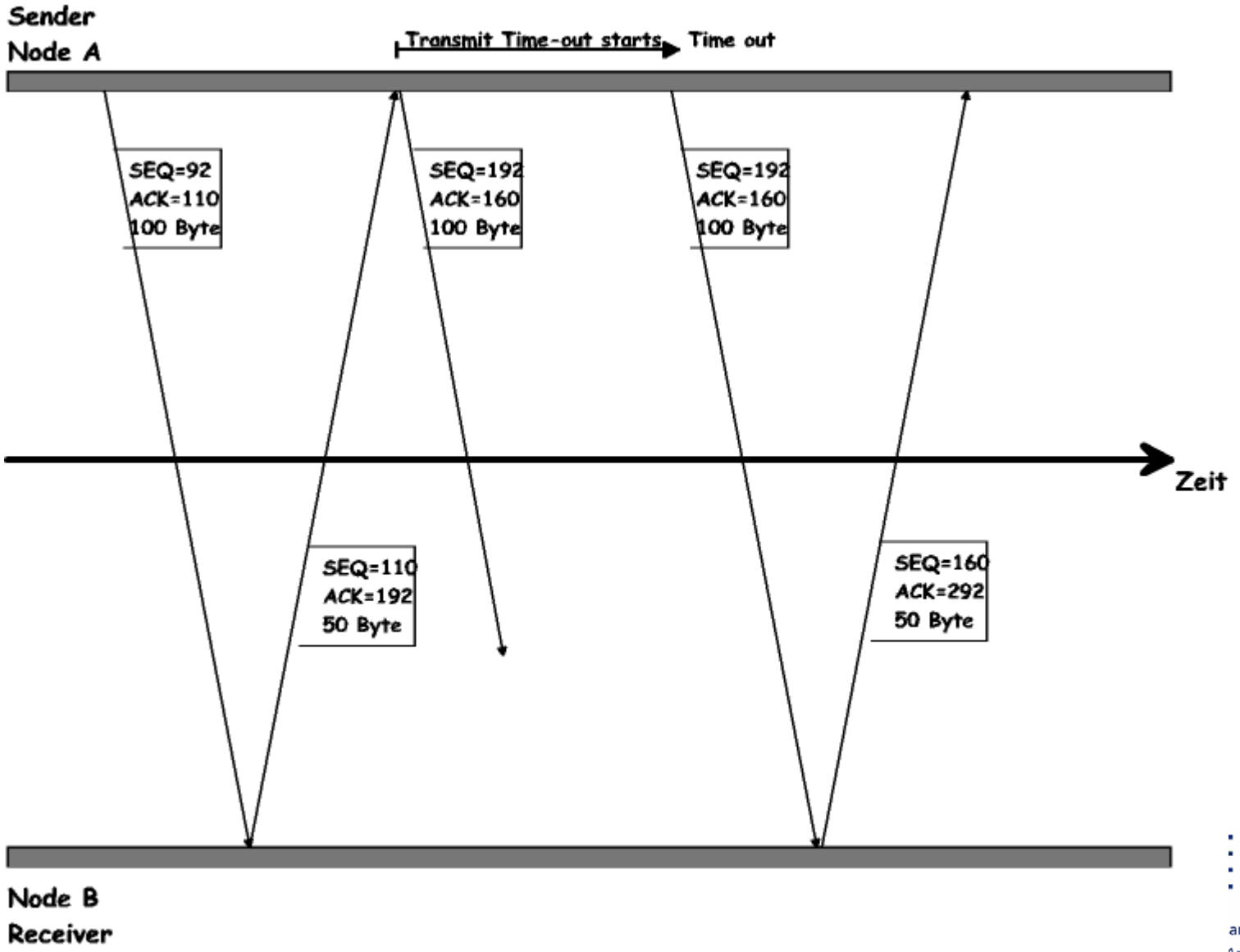
4.2 TCP Sicherung

- TCP sichert den Transport seiner Segmente derart ab, dass beim Empfänger ein vollständiger, geordneter Datenstrom erhalten wird
- Fehlt ein Segment beim Empfänger, wird der Strom angehalten und auf das fehlende Paket gewartet - „Head of Line Blocking“
- Datenverluste werden an der Sequenznummer erkannt
- TCP meldet jedoch keine Verluste, sondern versendet Empfangsquittungen (ACKs)
- Dabei quittiert TCP (ursprünglich) das letzte zusammenhängende Segment – „Cumulative acknowledgement“
- Nach einem Verlust beginnt TCP (ursprünglich), vom verlorenen Segment an den Strom neu zu senden – „go back N“

4.2 Empfangsquittungen (ACK)

- TCP ordnet die empfangenen Segmente in ihre ursprüngliche Reihenfolge (gem. Sequenznummer)
- Sobald Segmente in zusammenhängender Reihenfolge eingetroffen sind, ist TCP ‚quittungsbereit'
- TCP versucht, ACK gemeinsam mit Daten zu senden (Piggybacking)
  - Sind keine Daten ‚versandfertig', wird ACK verzögert
  - Nach (typisch) 100 – 200 ms wird ACK auch alleine versandt

4.2 Retransmission



4.2 Retransmission

Wann eine Sendung von TCP-Messages wiederholen?

- Feste Timeouts problematisch bei variabler Verzögerung
  - zu groß: Performanceverlust
  - zu klein: unnütze Netzlast durch Wiederholungen
- Lösung: Round Trip Time = durchschnittl. Verzögerung zwischen Aussenden und Quittungsempfang
- TCP ermittelt RTT für jede Verbindung
  - Retransmit Timer basiert auf RTT
  - dennoch: Probleme bei schnell veränderlicher RTT!

4.2 Retransmit Timing, Slow start

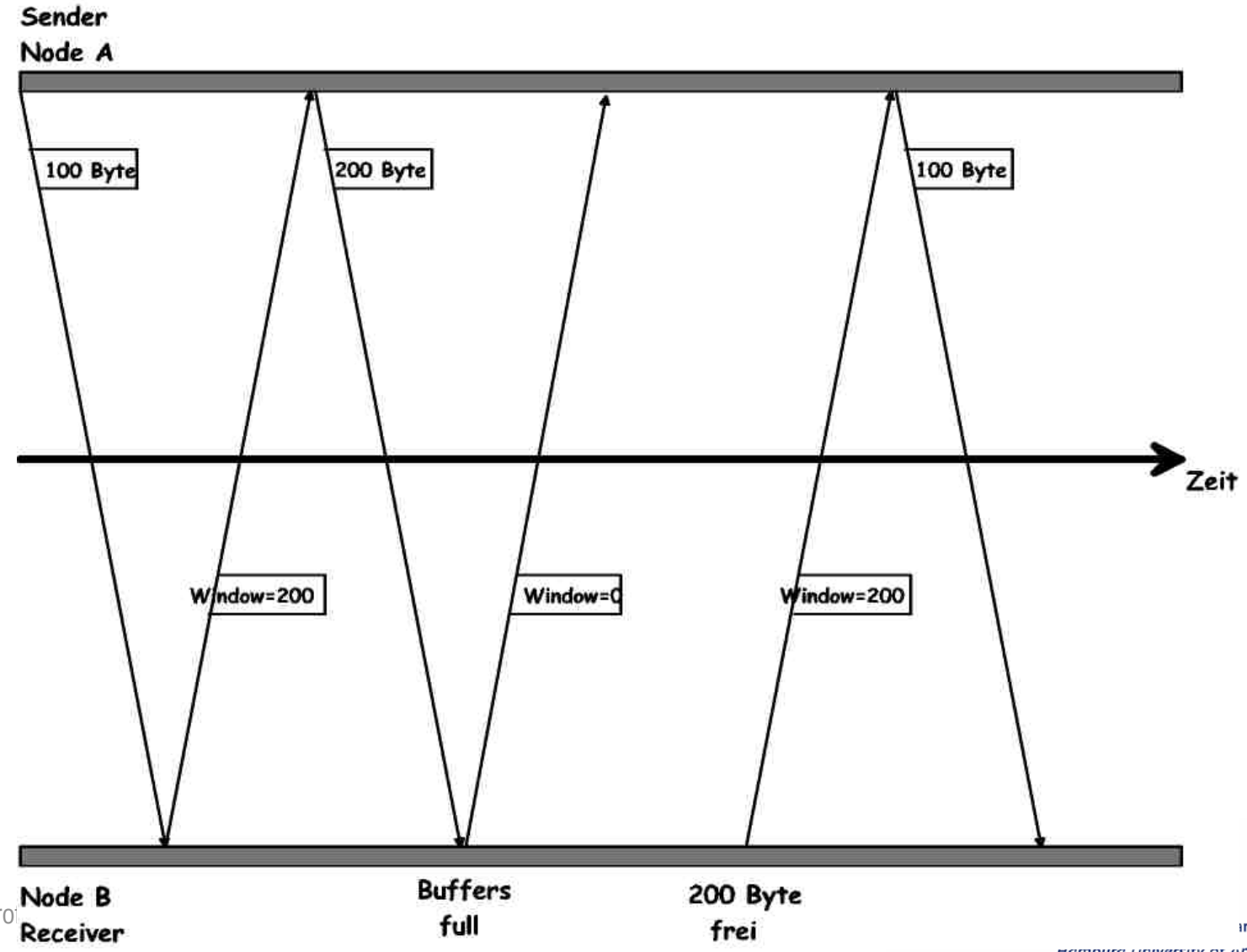
Adaptive Zeitsteuerung ist komplex:

- TCP ermittelt erwartete RTT, Delay-Variation (Jitter) und Timeout
  - Timeout = estRTT + 4 * D-V (Initialisierung)
  - Timeout-Verdopplung bei Retransmit (Exponential Backoff)
- Nach Retransmit (und zu Beginn) operiert TCP zur Vermeidung von Netzwerkstaus (congestion avoidance) einen Slow Start:
  - Congestion Window wird beim Sender geführt
  - Es startet mit dem Wert ‚1 Segment'
  - Per ACK erhöht sich die Fenstergröße um 1 Segement

4.3 Flußkontrolle

- Prinzip der dynamischen Flußkontrolle:
- TCP teilt den Datenfluß zur Übertragung in Segmente ein
- Der Empfänger steuert den Datenfluß durch Mitteilen der verfügbaren (Empfangs-) Puffergrößen: windowsize
- Ein Fenster der Größe 0 stoppt den Fluß
- Der Empfänger kann zusätzliche ACKs schicken, um den Fluß wieder in Gang zu setzen

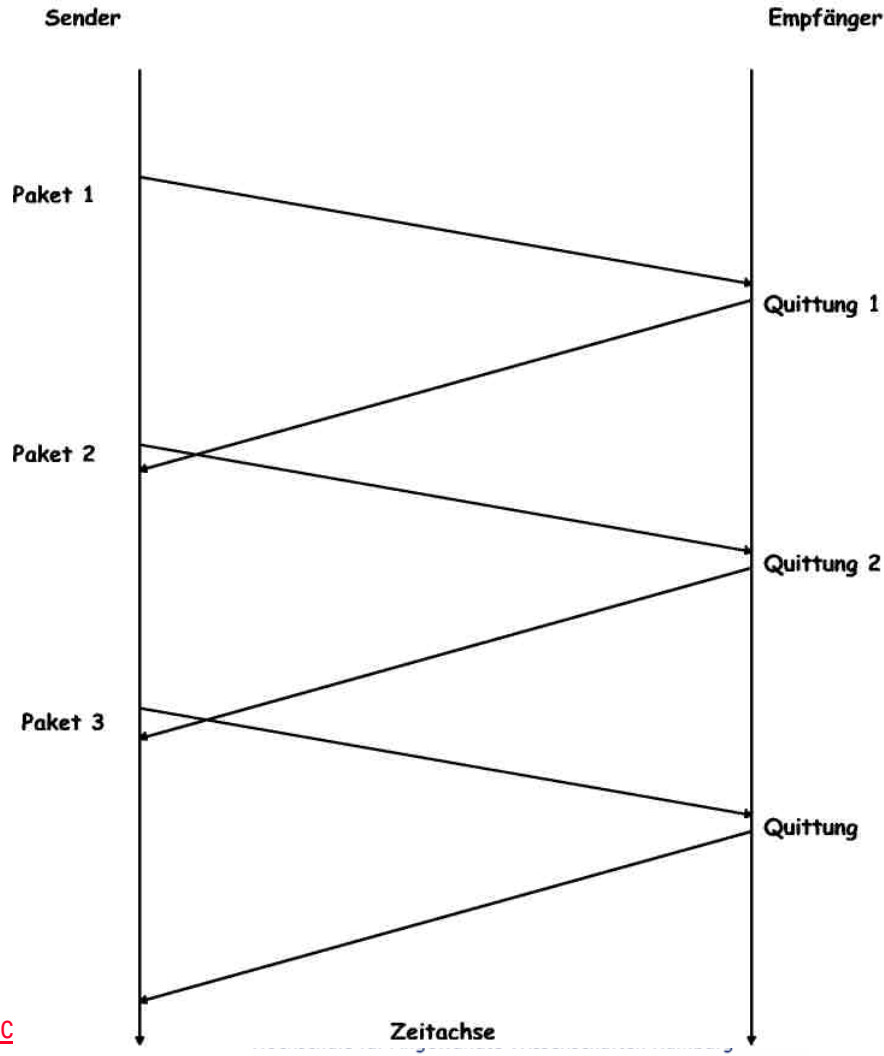
4.3 Window Advertisement



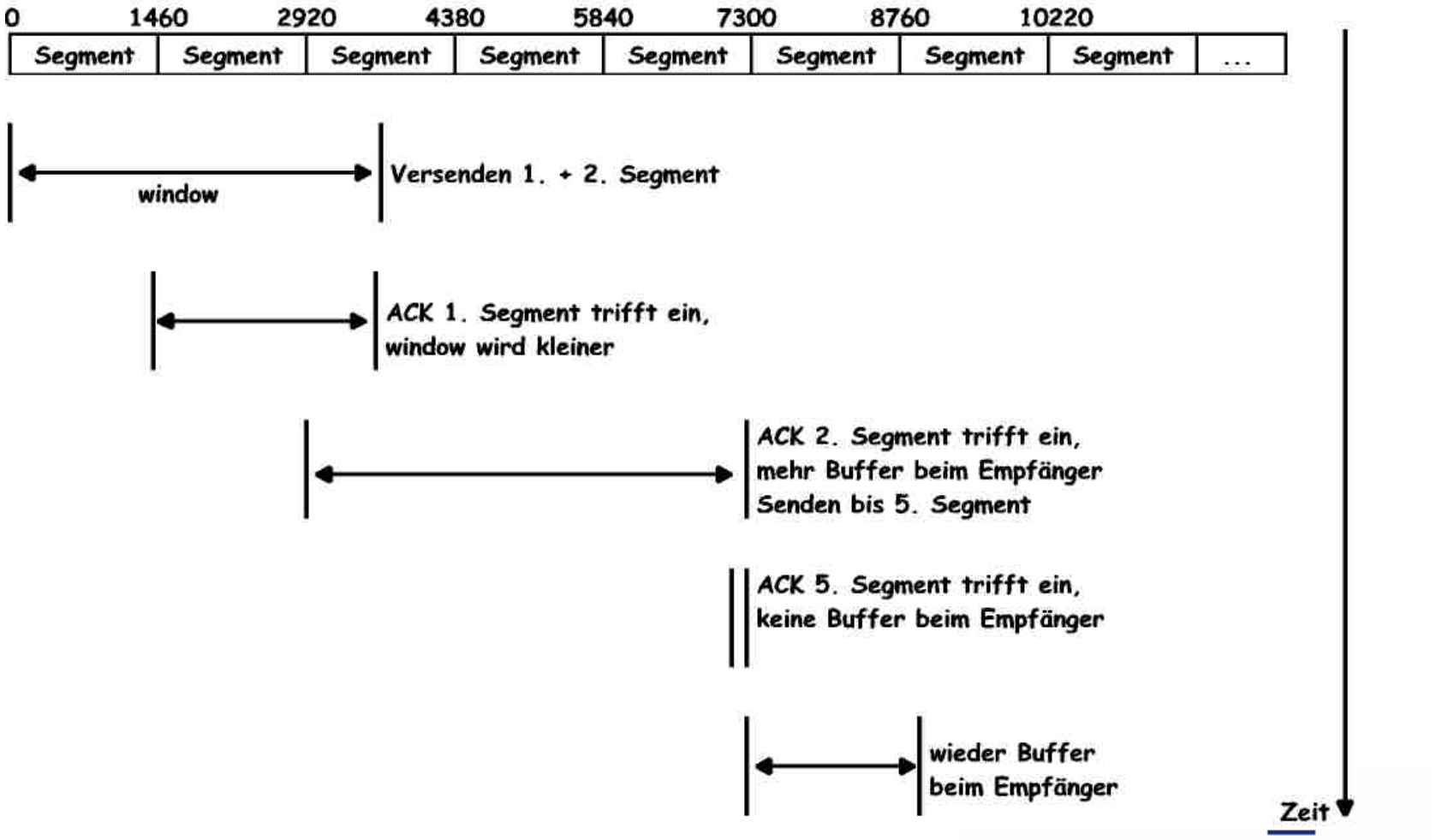
4.3 Sliding Window

Optimierung der Flußkontrolle:

- Jeder darf die Anzahl Bytes im Window senden, ohne auf eine Quittung zu warten
- Typische default Fenstergrößen: 4096 - 16384 Bytes
- Höhere Werte insbesondere bei ‚long fat pipes‘
- Max. Fenstergröße: Receive Buffer Size
- Das Empfangsfenster wandert mit jedem ACK entsprechend weiter



4.3 Datenstrom und Datenfluss



4.4 TCP Optimierungen

- TCP besteht seit seiner Erfindung unverändert ‚on the wire' – TCP kennt keine Versionsnummer
- Dennoch hat TCP kontinuierlich Optimierungen und Erweiterungen erfahren
  - Um ‚besser', d.h. effizienter und leistungsfähiger zu werden
  - Um sich den veränderten Übertragungsanforderungen anzupassen (z.B. Wireless)
  - Um die gestiegenen Endgeräte-Kapazitäten zu nutzen
- Dabei liegt die Herausforderung darin, gleichmäßig kompatibel zu bleiben

4.4 Vermeidung von Tinygrams

Einzeln versendete Datenbytes haben (auch) 40 Byte Header. Zur Vermeidung unnötigen Overheads:

- Nagle Algorithmus:
  - Sende unvollständige Segmente erst nach vollständigen ACKs
  - Während der Wartezeit werden Daten gesammelt
- Problem: Graphische Interaktion (X11) und verzögertes ACK
  - Nagle Algorithmus kann mit Socket-Option TCP_NODELAY ausgeschaltet werden.

4.4 Jacobson Fast Retransmit

TCP quittiert empfangene Pakete kontinuierlich

- Wird innerhalb eines sliding Window ein Segment verloren, muß das ganze Fenster erneut gesendet werden.

Einfache Idee zur Verbesserung:

- Wenn ein ‚out-of-order' Segment eintrifft, sendet TCP eine erneute Quittung des kumulativ erhaltenen Stroms (duplicate ACK).
- Annahme: wenn mehrere gleiche ACKs eintreffen, ist wahrscheinlich nur ein Zwischensegment verloren worden.
- Fast Retransmit:
  - Wird das dritte duplicate ACK erhalten, schickt der Sender das erste nicht bestätigte Segment unverzüglich erneut.
  - Der Sender wechselt nicht in die Slow Start Prozedur.

4.4 Selective Acknowledgment

TCP quittiert empfangene Pakete nur zusammenhängend

- Werden innerhalb eines sliding Window Segmente verloren, muß das ganze Fenster erneut gesendet werden.

Komplexerer Lösungsansatz:

- Segmentbereiche (innerhalb eines sliding Window) können diskontinuierlich quittiert werden (SACK)
- Der Sender hat dann die Option, zunächst die unquittierten Segmente erneut zu senden
- Erfolgt ein erneutes TIMEOUT, wird von der letzten (kumulativen) Standardquittung an erneut gesendet
- Erfolgt ein kumulatives ACK, wird der Vorgang abgebrochen

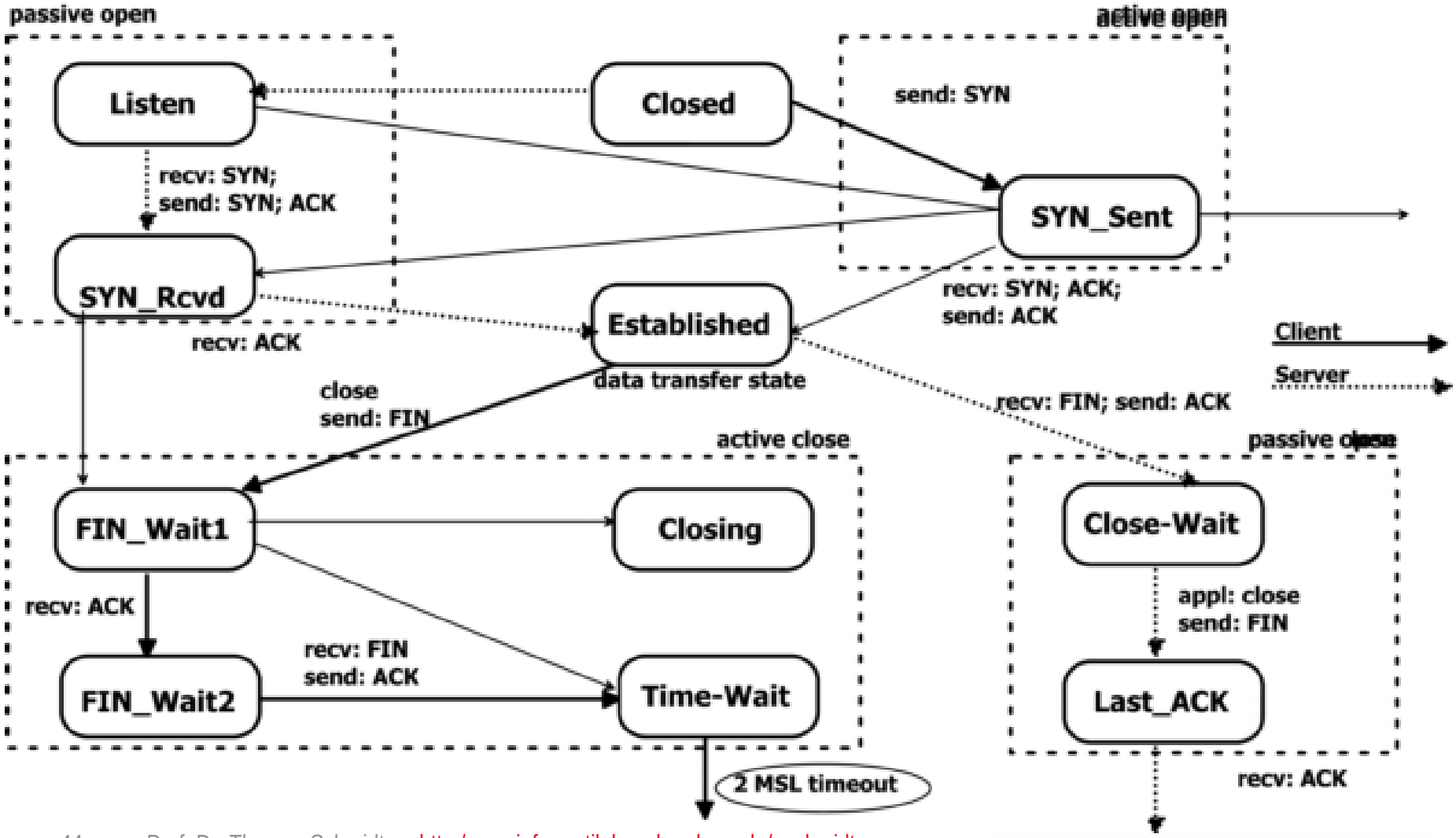
4.4 SACK

- Spezifiziert in RFC 2018
- Wird initial verhandelt bei dem Verbindungsaufbau (SYN)
- Selektive Quittungen (des Empfängers) erfolgen in den TCP-Header Options
- Der Sender muß eine separate ‚Sack'-Tabelle führen
- Der Sender ist berechtigt, auf selektive Quittung nicht zu reagieren

4. TCP Connection Timer

- Retransmission: Dauer, in welcher ein ACK erwartet wird.
- Persist: Dauer, nach welcher das Fortbestehen eines geschlossenen Empfangspuffers überprüft wird.
- Keepalive: Dauer, nach welcher die Verbindungsgegenstelle um ein Lebenszeichen gebeten wird.
- 2MSL: Dauer, in welcher TCP-Segmente im Netz gültig sind (Maximum Segment Lifetime).

4. TCP Zustandsdiagramm

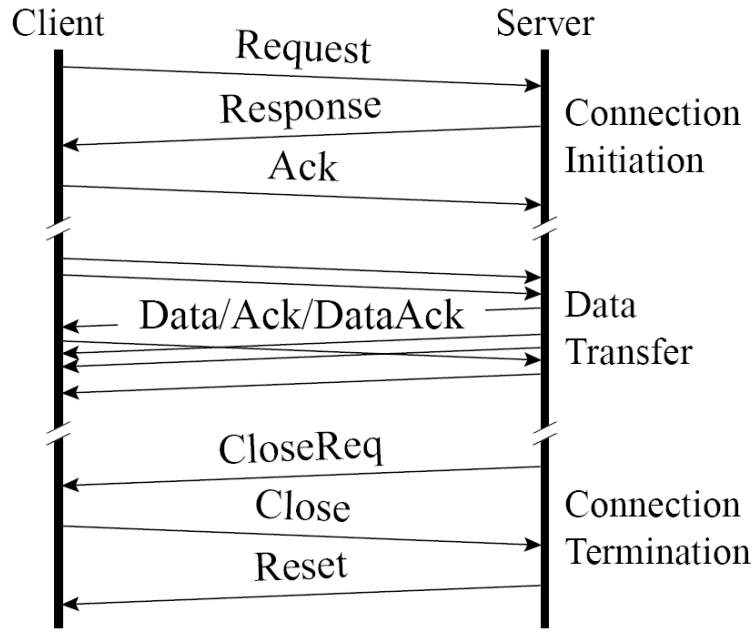


5. Streaming Control Transmission Protocol (SCTP)

- Spezifiziert in RFC 2960
- Verbindungsorientiertes Transportprotokoll
- Ermöglicht mehrere ‚Streams' pro Verbindung (analog SS7)
- Stream-Eigenschaften separat definierbar
- Unterstützt Multi-Homing
- Erweiterungen für Mobility und skalierbare Retransmission
- Implementiert SACK

5. Datagram Congestion Control Protocol (DCCP - RFC 4340)

- Protokoll für ungesicherten Unicast Transport mit Staukontrolle
- Entworfen für Echtzeitanwendungen
- Verbindungsorientiert, entdeckt Packetverlust, ohne Pakete zu wiederholen
- Bietet den Rahmen für verschiedene Staukontrollmechanismen (Window-/Raten-basiert)
- Implementierungen Linux & BSD



Selbsteinschätzungsfragen

1. Welches Transportprotokoll eignet sich zur Übertragung von Dateien, welches zur Gruppenkommunikation?
2. Wozu dient bei Verbindungsbeendigung der TCP Zustand CLOSE_WAIT?
3. Wie unterstützt TCP eine Flussteuerung im Netz? Warum wird TCP auch als ‚höflich' bezeichnet?
4. Wie entscheidet TCP, ein Segment wiederholt zu versenden? Welche Erweiterungen gibt es?