Institut für Informatik

Vorlesung
Werkzeuge der Informatik

Grundlagen und Werkzeuge des WWW (Teil 1)

Jörg P. Müller

Inhalt

- **Entwicklung von Internet und WWW**
- **WWW-Architektur und Protokolle**
  - **WWW-Architektur (Client-Server)**
  - **Basisprotokoll des Internet: TCP/IP**
  - **Web Ressourcen (oder: Was ist eine URL)**
  - **Das HTTP-Protokoll**
- **Darstellung von WWW-Inhalten**
  - **Das WWW-Dokumentenmodell: HTML**
  - **Cascading Style Sheets (CSS)**
  - **Die eXtensible Markup Language (XML)**

Begriffe: Internet - Intranet - Extranet

- **Internet**
  - **weltweites Netzwerk auf Basis der TCP/IP-Protokollfamilie (s.u.) mit mehreren Millionen Rechnern, für „jedermann“ offen**
- **Intranet**
  - **Kommunikationsnetz auf Basis von Internet-Technologien, das dem Informationsaustausch innerhalb einer begrenzten Interessengemeinschaft dient**
  - **z.B. Mitarbeiterportale großer Unternehmen**
- **Extranet**
  - **Variante des Intranet, bei dem Informationsflüsse aus dem Internet und dem Intranet verknüpft sind**
  - **z.B. Unternehmen erlaubt Geschäftspartnern Zugriff zu Teilen des eigenen Intranets**

Geschichte des Internet

- **Produkt des Kalten Krieges in den 60er Jahren**
- **Überlegungen US-Department of Defense (DoD): auch nach Atom-Angriffen soll Netz funktionieren -> erhebliche finanzielle Unterstützung**
- **„Advanced Research Project Agency“ ==> ARPANET**
- **1964 Vorstellung der Paketvermittlung durch Paul Barran**

Geschichte des Internet

- 1969 Vernetzung von 4 Universitäten
- 1971 Vernetzung von 13 Universitäten
- 1972 Vernetzung von 37 Universitäten
- 1973 Start des Internet durch Verbindung verschiedener paketorientierter Netze
- 1978 Beschluss, bei staatlichen Datenübertragungen nur noch TCP/IP einzusetzen
- 1983 Aufspaltung in Militär und Bildungsnetz
  – Militärischer Teil wird abgelöst:
  Defense ARPA (DARPA); später auch “Milnet”
- Kommerzialisierung ab 1989

Internetdienste

- **Internet bietet Infrastruktur, auf deren Basis für den Anwender nutzbare Dienste zur Verfügung stehen**
- **Beispiele:**
  - **World Wide Web: Zugriff auf Webseiten**
  - **File Transfer Service. Übertragung von Dateien (FTP, File Transfer Protocol)**
  - **Email Service (SMTP, Simple Mail Transfer Protocol)**
  - **Foren, Newsgroup (Usenet)**
  - **Internet Relay Chat**
  - **Instant Messaging**
  - **Internet-Telefonie**
- **Tendenz: Verschmelzen von Internet-Diensten mit dem WWW**

Das WWW

- **Ein Dienst basierend auf dem Internet**
- **Globaler digitaler Informationsraum bestehend aus Millionen Clients und Servern, die auf verknüpfte Informationsobjekte zugreifen**
  - **Server verwalten die Web Ressourcen**
  - **Clients geben Benutzern eine einfache Schnittstelle für Ressourcendarstellung und –zugriff (über Web Browser – Applikation)**
- **Web-Ressourcen : z.B. Texte, Dokumente, Bilder, Multimediadaten, Datenbankinhalte, ausführbare Programme sein**
- **Informationsobjekte sind identifiziert durch kurze, eindeutige Schlüssel, sogenannte Uniform Resource Identifiers (URIs)**
- **Zugriff auf Web Ressourcen über Hyperlinks auf der Basis der URIs**
- **Das WWW unterstützt ein einheitliches Protokoll zur Kommunikation zwischen einem WWW Server und einem WWW Client (HTTP)**
- **Sprache zur Beschreibung von WWW-Inhalten:**
  - **Hypertext Markup Language (HTML)**

Geschichte des WWW

http://www.w3.org/History.html

- **Anfänge des WWW → Geschichte des Internet**
- **1980: Tim Berners-Lee (CERN) schreibt Programm "ENQUIRE", das es erlaubt, Knoten im Internet zu verlinken**
- **1989: Tim Berners-Lee: CERN-Internes Proposal "Hypertext and CERN"**
- **1990: TBL prägt Begriff "World Wide Web" – Beginn eines großen Hypertext-Projekts bei CERN**
- **April 1993: CERN kündigt freie Nutzbarkeit des WWW an**
- **September 1993: Mosaic Browser (NCSA) verfügbar für X, PC/Windows and Macintosh.**
- **Mai 1994: Erste internationale WWW Konferenz**
- **Oktober 1994: Gründung des World Wide Web Consortium**

Geschichte des WWW (2)

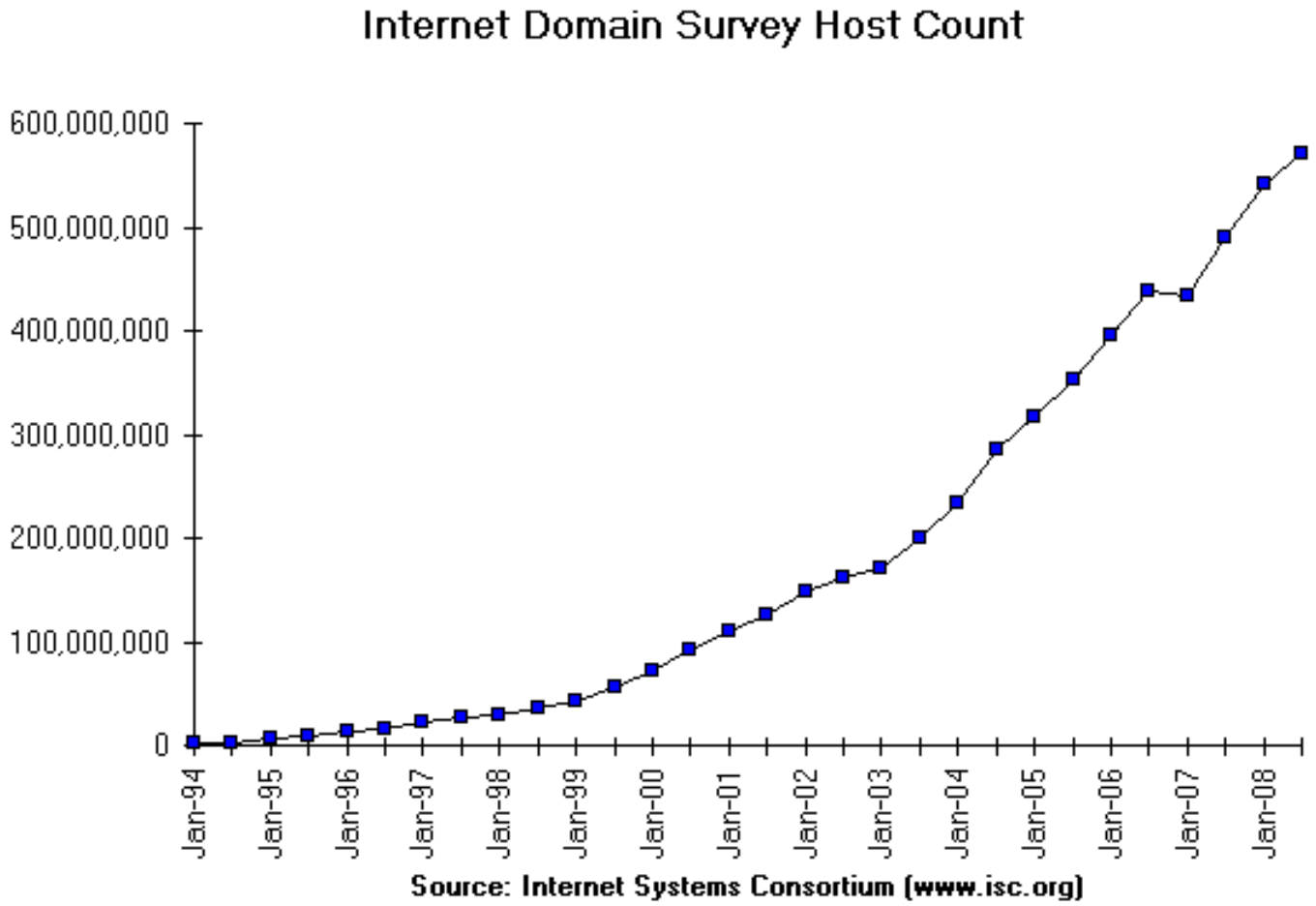
- 1993: Erste Web Search Engine (Wandex, MIT)
- 1995: Sun bringt Java Programmiersprache heraus mit Unterstützung für WWW; wenig später kündigen Netscape und Microsoft an, dass ihre Browser Java unterstützen werden
- Google
  - 1996 beginnt als Forschungsprojekt
  - 1998 Firmengründung
- seit ca. 1998:
  - Verfügbarkeit sicherer Kommunikations-protokolle für das WWW (https)
  - Entstehen der ersten webbasierten Electronic Commerce Systeme (eShops)
- 1999: Tim Berners-Lee prägt Vision des "Semantic Web"
- 2004: Medienunternehmen O'Reilly Media prägt den Begriff des "Web 2.0"

Weltweite Nutzung – Das Netz für alle?

Region	Anz. Nutzer in Mio.	% der Be-völkerung	Wachstum in % 2000-2008
Afrika	51,1 (32,8)*	5,3	1.031
Asien	578,5 (394,9)	15,3	406
Europa	384,6 (308,7)	48,1	266
Naher /Mittlerer Osten	41,9 (19,0)	21,3	1.177
Nordamerika	248,2 (229,1)	73,6	130
Mittel- u. Südamerika	139,0 (83,4)	24,1	669
Ozeanien / Australien	20,2 (18,4)	59,5	165
GESAMT	*1.463,6*	*21,9*	*306*

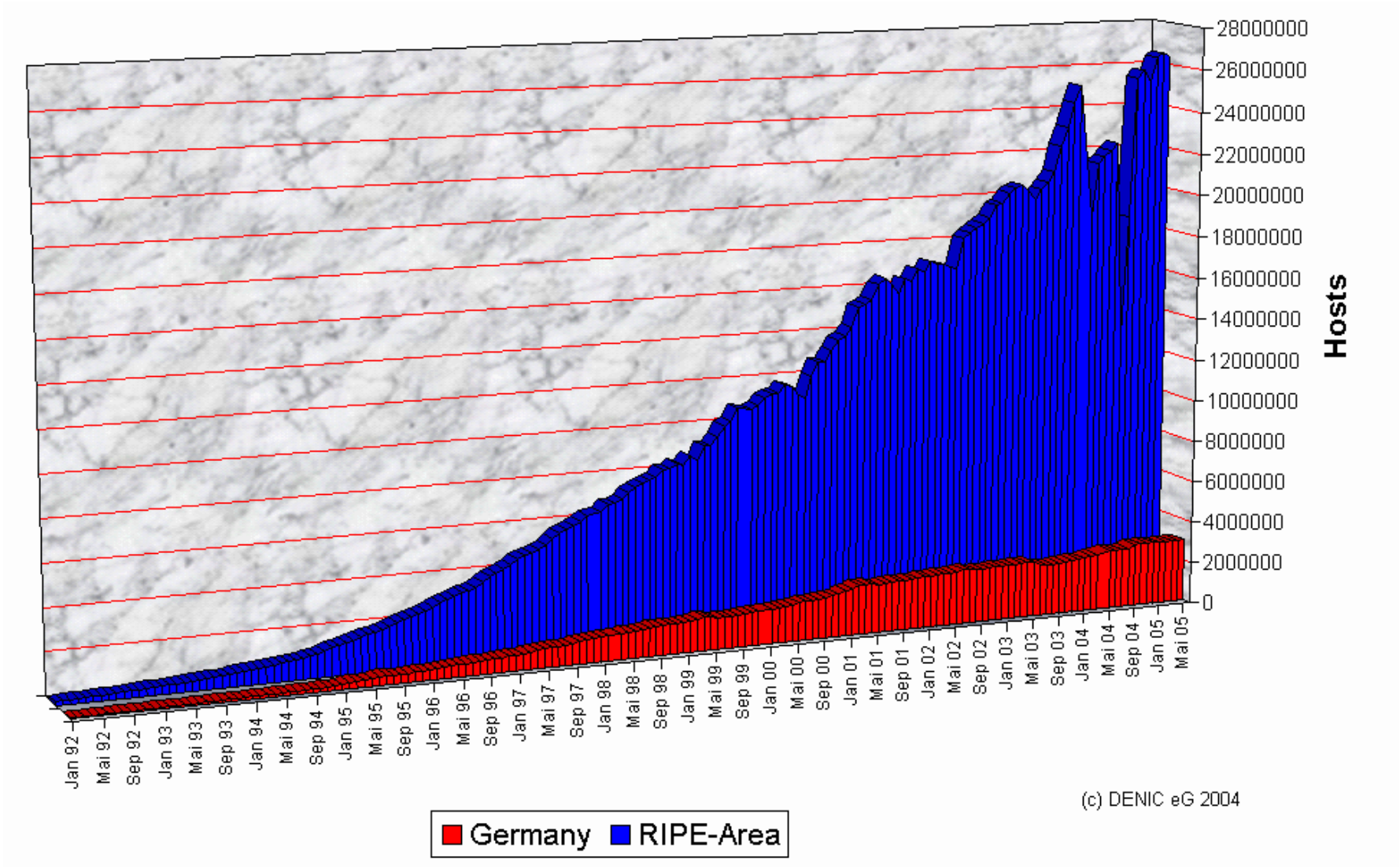
Quelle: http://www.internetworldstats.com/stats.htm von Nov 10, 2008
* zum Vergleich in Klammern: Werte vom 18.9. 2006)

Statistiken: Anzahl der Server im Internet



Quelle: Internet Systems Consortium, http://www.isc.org/ds/ (10.11.2008)

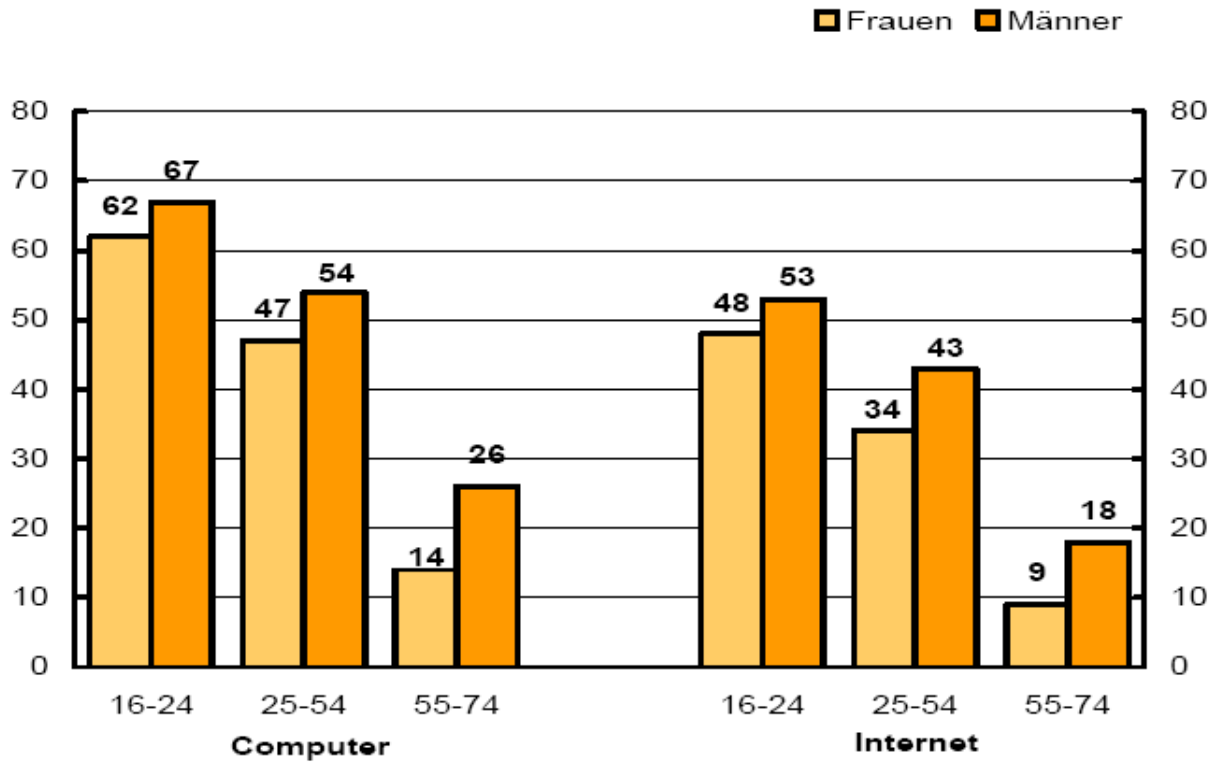
Hosts in Deutschland und Europa



Quelle: http://www.denic.de/de/domains/statistiken/hostentwicklung/hostcount.html

Internet-Nutzung in Deutschland

Abbildung 1: Frauen und Männer in EU-25, die in den letzten drei Monaten (2006) durchschnittlich einmal täglich oder fast einmal täglich einen Computer oder das Internet nutzten (% der Frauen/Männer in jeder Altersgruppe)

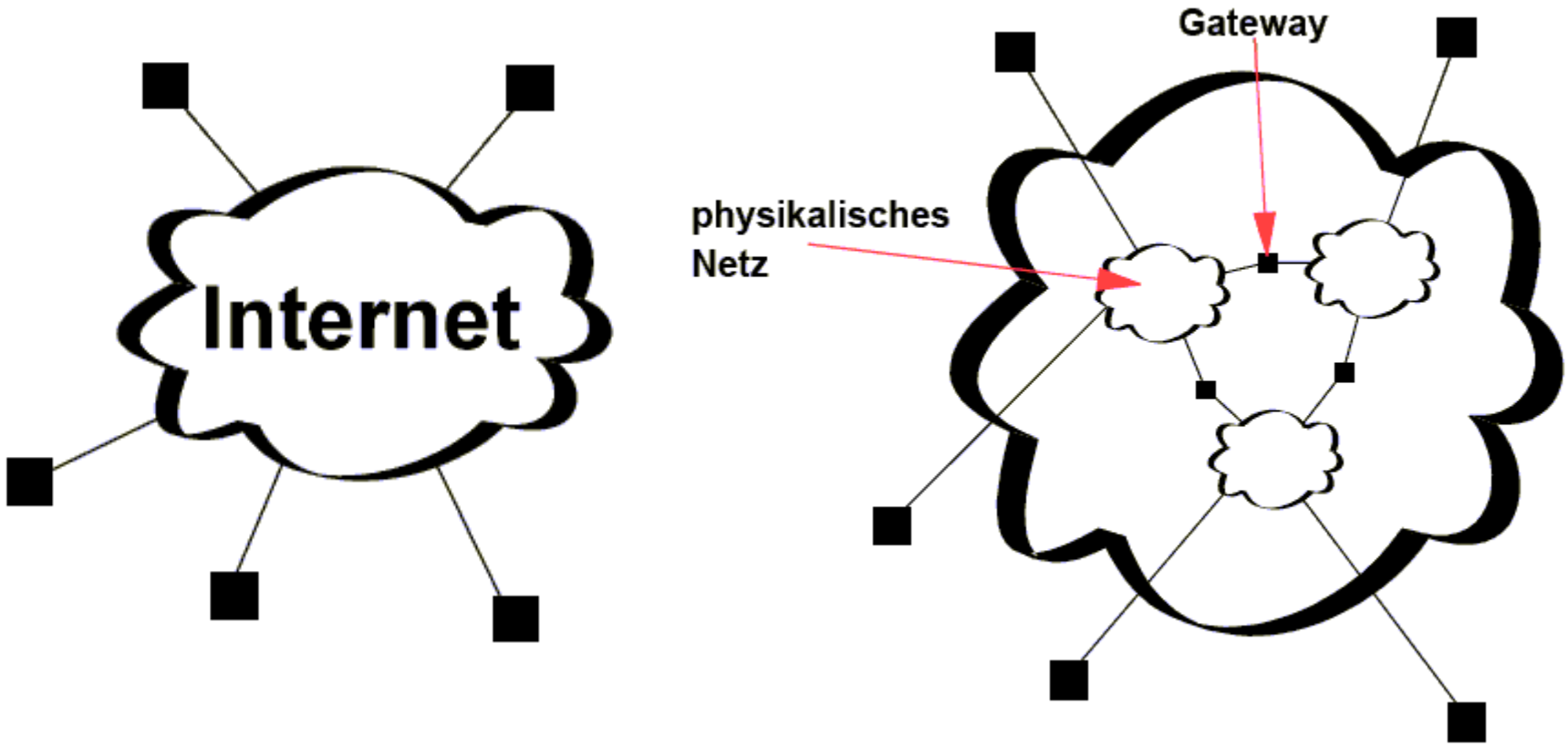


Quelle: Eurostat, Gemeinschaftserhebung über den IKT-Einsatz durch Haushalte und Einzelpersonen

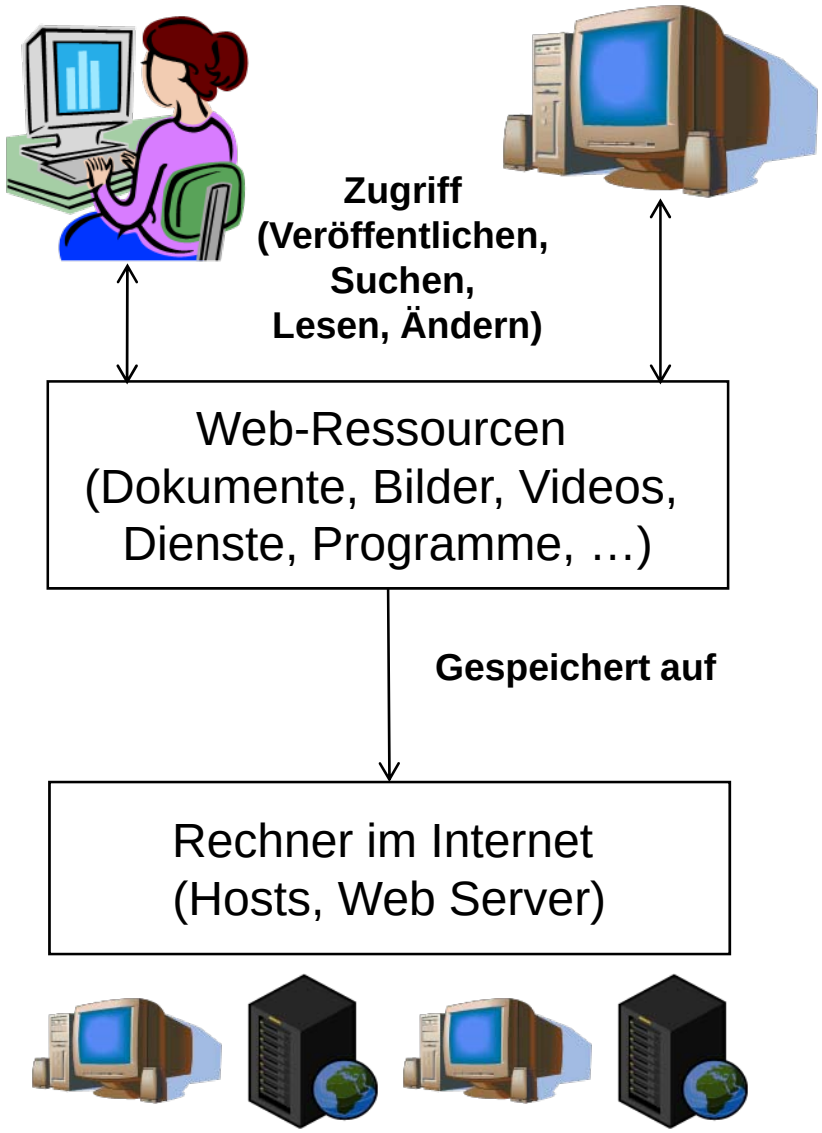
Organisationen

- **Internet Activities Board (IAB)**
  - **betreut den Standardisierungsprozess**
  - **Verwaltung der RFCs (Request for Comments)**
  - **Unterorganisationen:**
    - **IETF (Internet Engineering Task Force): Funktion des Internet sowie für die Lösung aller Protokoll- und Architekturfragen**
    - **IRTF (Internet Research Task Force): Entwicklung neuer Technologien**
- **World Wide Web Consortium (W3C)**
  - **Zusammenschluss der Industrie**

Das Internet :
Zusammenschluss vieler Teil-Netze



Nutzerorientierte Sicht auf WWW/Internet



Fragestellungen

- Identifikation/Benennung von Web Ressourcen
- Architekturen v. Web-Anwendungen
- Datenkommunikation im WWW
- Menschenlesbare Darstellung von Inhalten
- Maschinenlesbare Darstellung von Inhalten

Identifikation / Benennung von Web-Ressourcen: Uniform Resource Identifier (URI)

- **Zeichenfolge, die zur Identifikation einer Ressource dient**
- **Allgemeiner Aufbau:**
  - `<Schema>:<Schemaspezifischer Teil>`
- `<Schema>` **gibt Typ der URI an, z.B. http, ftp, mailto**
- `<Schema>` **legt Interpretation des Schemaspezifischen Teils fest**
- **URI-Schemata wie ftp und http sind hierarchisch aufgebaut:**

```
<Schema>://[<Benutzer>[:<Passwort>]@]<Server>[:<Port>]/[<Pfad>]
            [?<Anfrage>][#<Fragment>]
```

- **Zwei Formen von URIs**
  - **Uniform Resource Locators (URLs): Identifikation der Ressource durch Beschreibung des Zugriffs (positionsabhängige Referenz) auf die Ressource**
  - **Uniform Resource Names (URNs): Logische Identifikation i. S. einer global eindeutigen, positionsabhängigen und persistenten Referenz**

URL: Häufig verwendete Strukturen

- **Verwendung nur eines DNS-Namens**

Schema	Host-Name	Pfadname
http ://	www.in.tu-clausthal.de	/home/mueller/mbox

- **Kombination DNS-Name mit Portnumber**

Schema	Host-Name	Schema	Pfadname
http ://	www.in.tu-clausthal.de :	80	/home/mueller/mbox

- **Kombination IP-Adresse mit Portnumber**

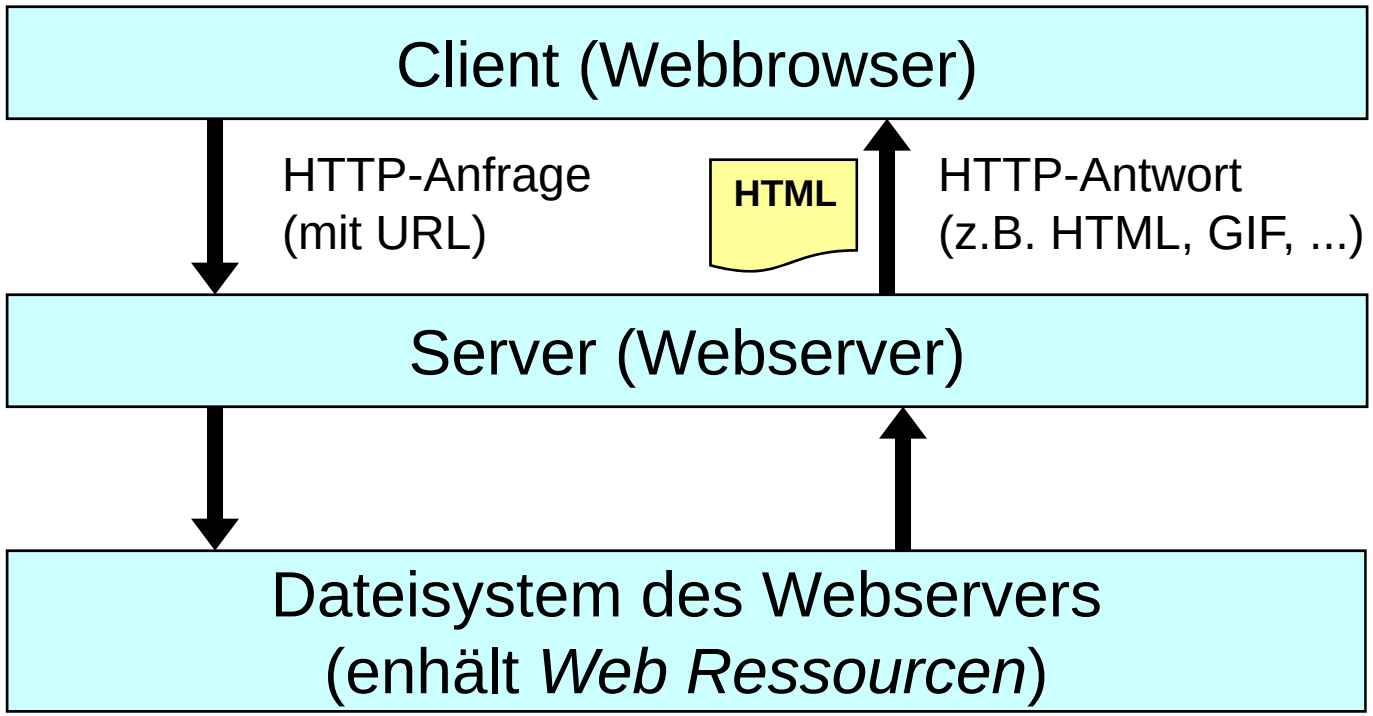
Schema	IP-Adresse	Schema	Pfadname
http ://	139. 174. 2. 135 :	80	/home/mueller/mbox

Architekturen von WWW- Anwendungen

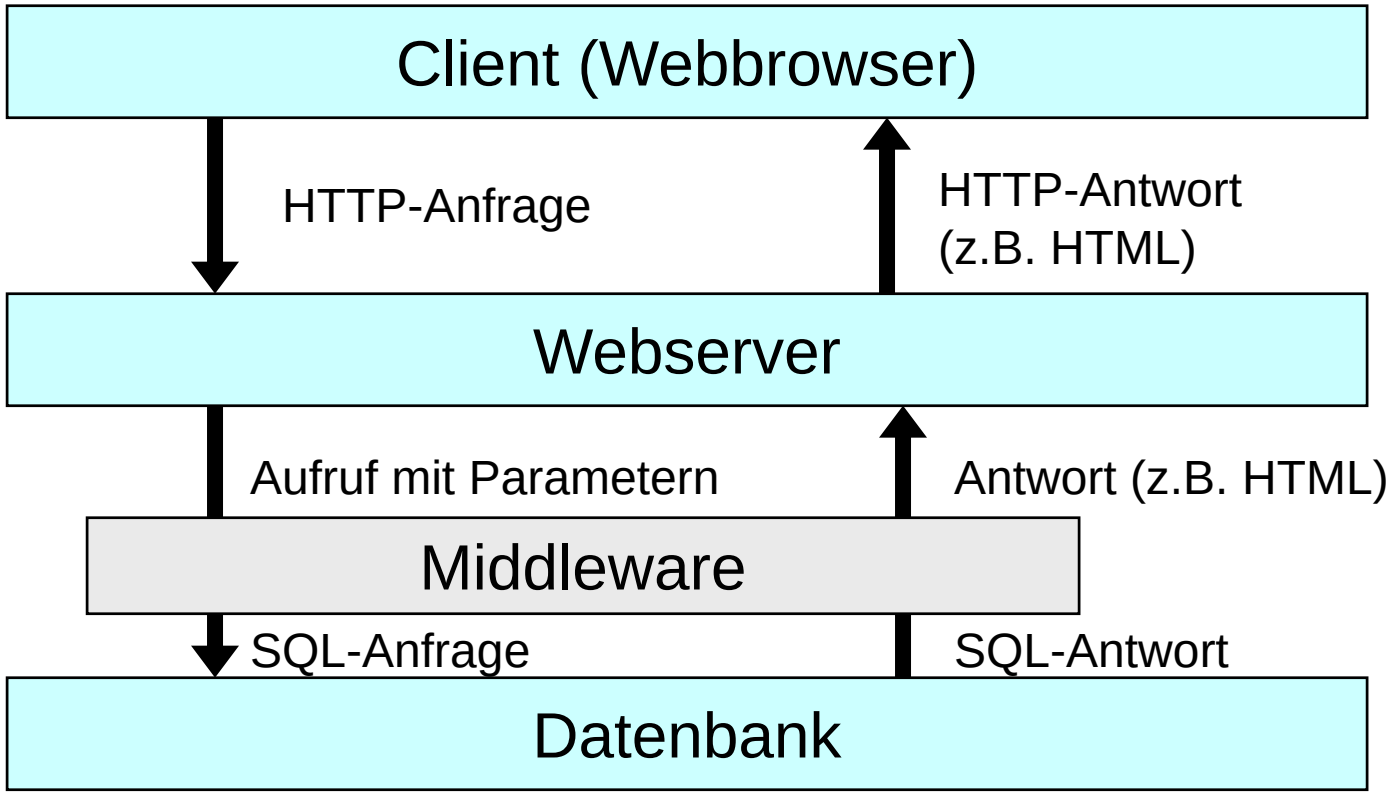
Grundsätzlicher Aufbau webbasierter Anwendungen

Basiert auf einer Client-Server-Architektur

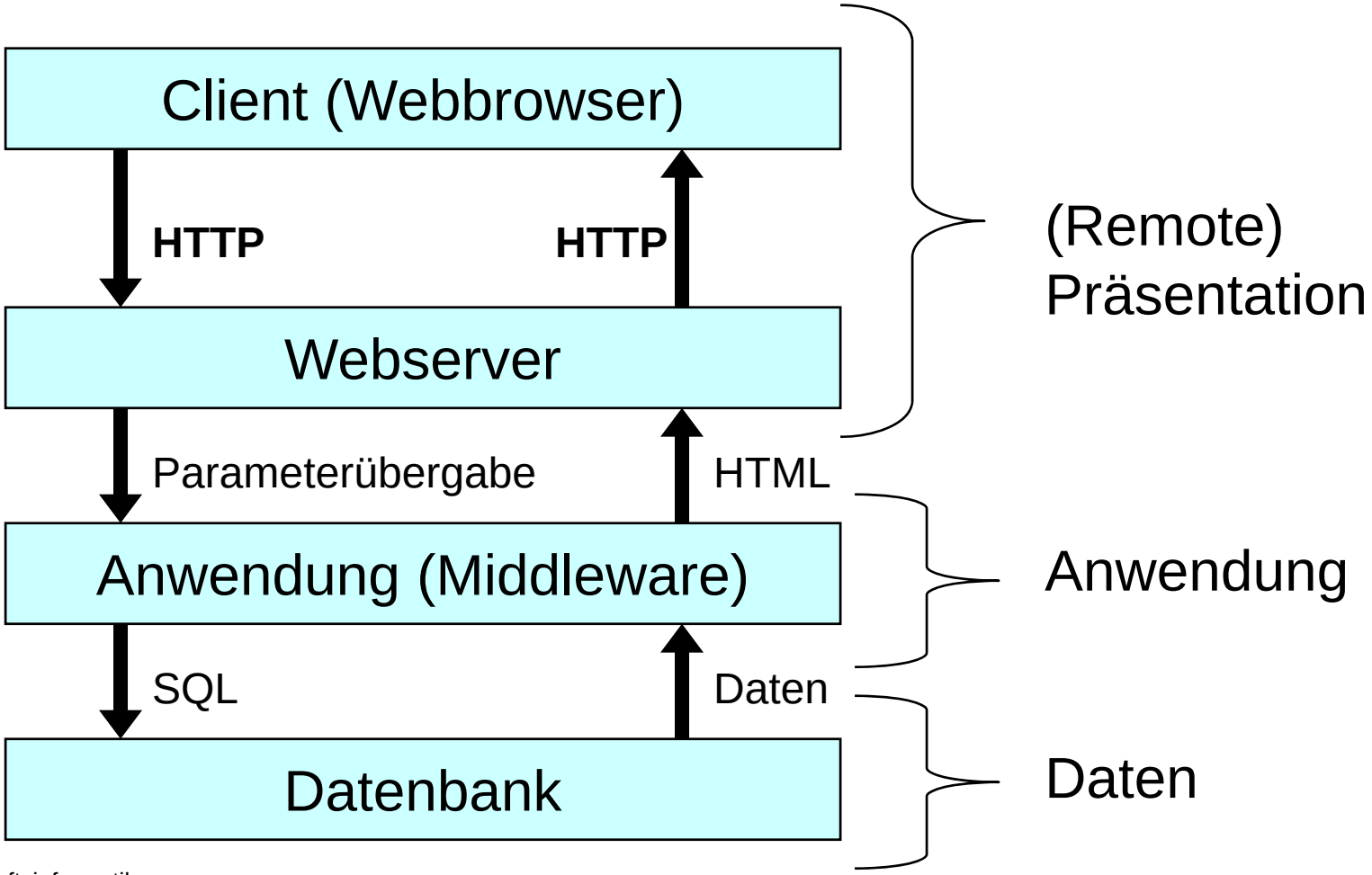
(Remote) Presentation



Architektur datenbankbasierter Web-Anwendungen



Drei-Schichten-Architektur Modell



Datenkommunikation im WWW

- **Frage: Was passiert bei einer Google-Suche?**
- **Der Web-Browser sendet Daten an den Web-Server**
- **Der Web-Server bearbeitet die Anfrage**
- **Der Web-Server schickt eine Antwort zurück**
- **Der Web-Browser zeigt die Antwort an**
- **Rechner kommunizieren mittels <u>Protokollen</u>:** *Vereinbarung über den organisatorischen Ablauf einer Datenübertragung*

Paketorientierte Datenübertragung im Internet

- **Robuste Übertragung von Nachrichten zwischen Sender und Empfänger**
  - **Aufteilen der Nachricht in Datenpakete**
  - **Senden der einzelnen Datenpakete durch das Internet**
  - **Rekonstruktion der Nachricht aus den einzelnen Datenpaketen**
- **Standardprotokolle des Internet: TCP/IP Protokollfamilie**
- **Bestehend aus**
  - **TCP (Transmission Control Protocol)**
  - **IP (Internet Protocol)**
- **Weitere Protokolle, z.B.**
  - **UDP (User Datagram Protocol): alternativ zu TCP, verbindungslos**
  - **ARP (Address Resolution Protocol): Kontrollprotokoll zur Zuordnung von Adressen zu Rechnernamen**

TCP/IP Protokollfamilie

TCP

- Definition und Übertragung von Nachrichten durch das Internet
- Sender und Empfänger durch (IP-Adresse, Port) beschrieben
- Zerlegung der Nachrichten in IP-Pakete, sowie Rekonstruktion
- Verbindungsorientiert und zuverlässig (alle Daten kommen genau einmal und in richtiger Reihenfolge an)

IP

- Definition und Übertragung (Routing) von Datenpaketen durch das Internet (von Ausgangshost zum Zielhost)
- Verbindungslos und unzuverlässig
- Netzübergreifend

HTTP – das Hypertext-Transfer-Protokoll

- Kommunikation zwischen Web Server und Web Clients
- Einfaches Client-Server-Protokoll
  - Client schickt Anforderungsnachricht an Server (sog. HTTP-Request)
  - Server verarbeitet Anforderung und sendet Antwort (sog. HTTP-Response)
- HTTP ist ein zustandsloses Protokoll, d.h.
  - Server verwaltet keine Information über Clients
  - Im Prinzip: Für jeden Request-Response-Vorgang wird eine neue TCP-Verbindung zwischen Client und Server aufgebaut
- HTTP-Request und Response-Nachrichten basieren auf TCP/IP
- HTTP enthält darüber hinaus z.B. Kontrollinformation

Parameterübergabe an den Webserver

- **Typischer Anwendungsfall: Formulardaten verarbeiten**

Serverseitiges Programm z.B. Email versenden

Zwei Methoden:

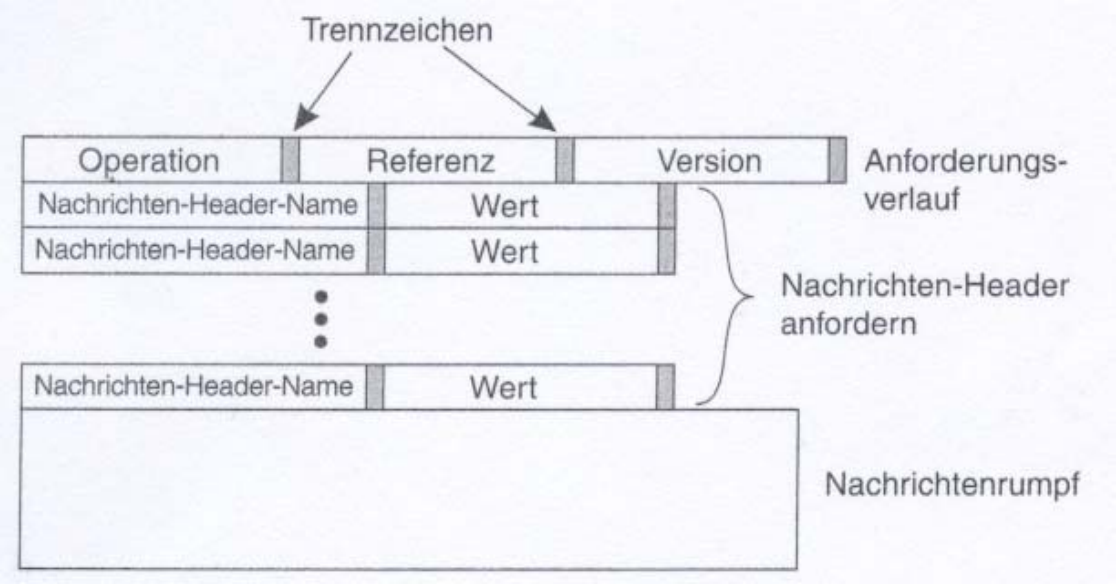
- Get
- Post

Parameterübergabe GET vs. POST

- GET
  - Anhängen an URL
  - für Benutzer sichtbar (auch Passwort-Felder)
  - begrenzte Länge
  - in Formularen und in Links verwendbar
- POST
  - Übertragung im Request
  - unbegrenzte Länge
  - vor allem für Formulare verwendet

Aufbau eines HTTP-Request

[Quelle: Tanenbaum (2003), p.737]



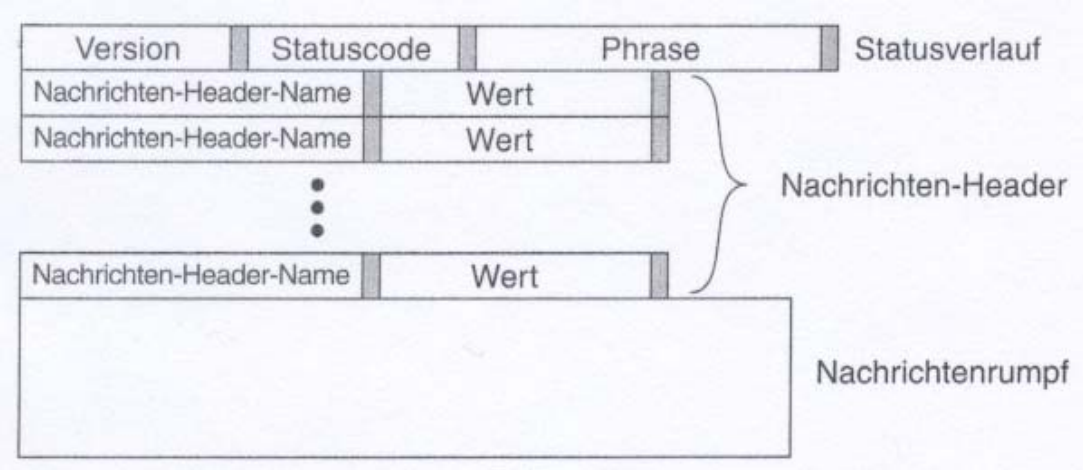
- Enthält IP-Adresse des Servers und die Bezeichnung der angeforderten Ressource (Seite)

```
GET /verzeichnis/seite.pl HTTP/1.0
Host: 100.101.102.103
```

- Möglichkeiten zur Übermittlung weiterer Informationen:
  - QueryStrings: an die URL angehängte Informationen
  - im Request-Body (bei Methode POST)
  - Cookies

Aufbau einer HTTP-Response

[Quelle: Tanenbaum (2003), p.737]



- **Dreistelliger Status-Code mit textueller Beschreibung**
  - **Z.B. 200 = “OK”, 405 = “Method not allowed”**
- **Weitere Information im Response-Header,**
  - **z.B. “Allow head, get”, “LastModified 11.11.2005”**
- **Nachrichtenrumpf enthält in der Regel das HTML-Dokument**

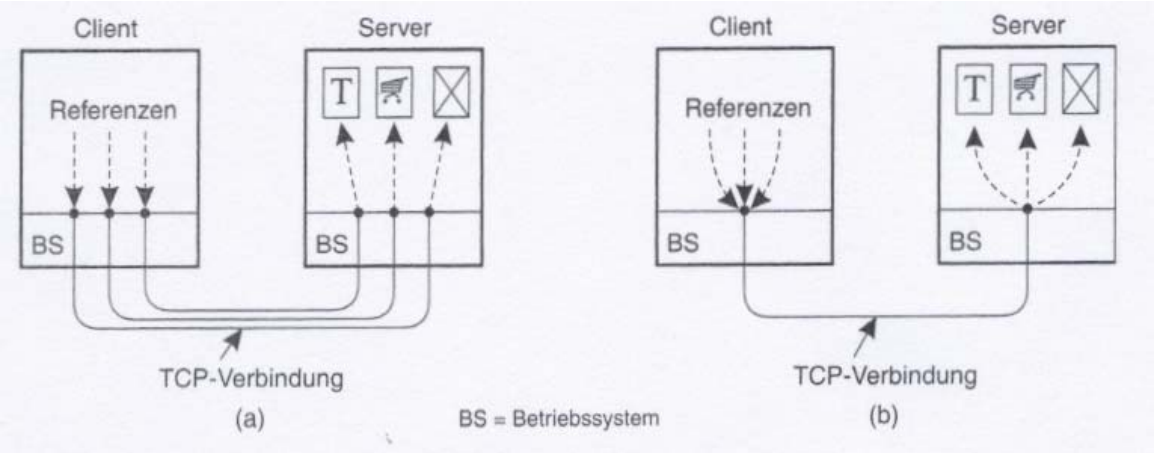
Beispiel HTTP Post Request-Nachricht

```
POST /send.php HTTP/1.1
  Host: meinserver.de
  User-Agent: Mozilla/4.0
  Accept: image/gif, image/jpeg, */*
  Content-type: application/x-www-form-urlencoded
  Content-length: 51
  Connection: close

  Vorname=Max&name=Mustermann&mail=max%40muster%2Ede
```

TCP-Verbindungsarten in HTTP

[Quelle: Tanenbaum (2003), p.735]



(a) nicht-persistente Verbindung (b) persistente Verbindung

- **Problem: Der Zugriff auf ein „logisches Webdokument“ erfordert mehr als einen physischen HTTP-Request (z.B. im Web-Dokument referenzierte Bilder)**
- **Abhilfe: Ab HTTP Version 1.1: Unterstützung für persistente TCP-Verbindungen, d.h. mehr als ein Request-Response-Paar pro aufgebauter TCP-Verbindung**
  - **Kostspieliger Aufbau der TCP-Verbindung bei jedem Request entfällt.**
- **Pipelining: Client kann mehrere Requests absetzen, ohne auf die Antwort auf die erste zu warten**

Ressourcen

- **Internet / WWW:**
  - **http://www.w3.org/WWW**
  - **Internet Society: http://www.isoc.org**
  - **History of the Internet: http://www.isoc.org/internet/history/**
  - **History of the WWW: www.w3history.org/**
- **HTML / CSS:**
  - **http://selfhtml.org**
- **XML:**
  - **http://www.w3.org/XML/**
  - **Elliotte R. Harold und W. Scott Means. *XML in a Nutshell*, O'Reilly, 2005.**
- **XML Schema:**
  - **http://www.w3.org/XML/Schema**