GABRIEL RÜECK

B6545: TECHNOLOGY FOUNDATIONS FOR E-COMMERCE

INTERNET AND WEB TECHNOLOGIES SUMMARY

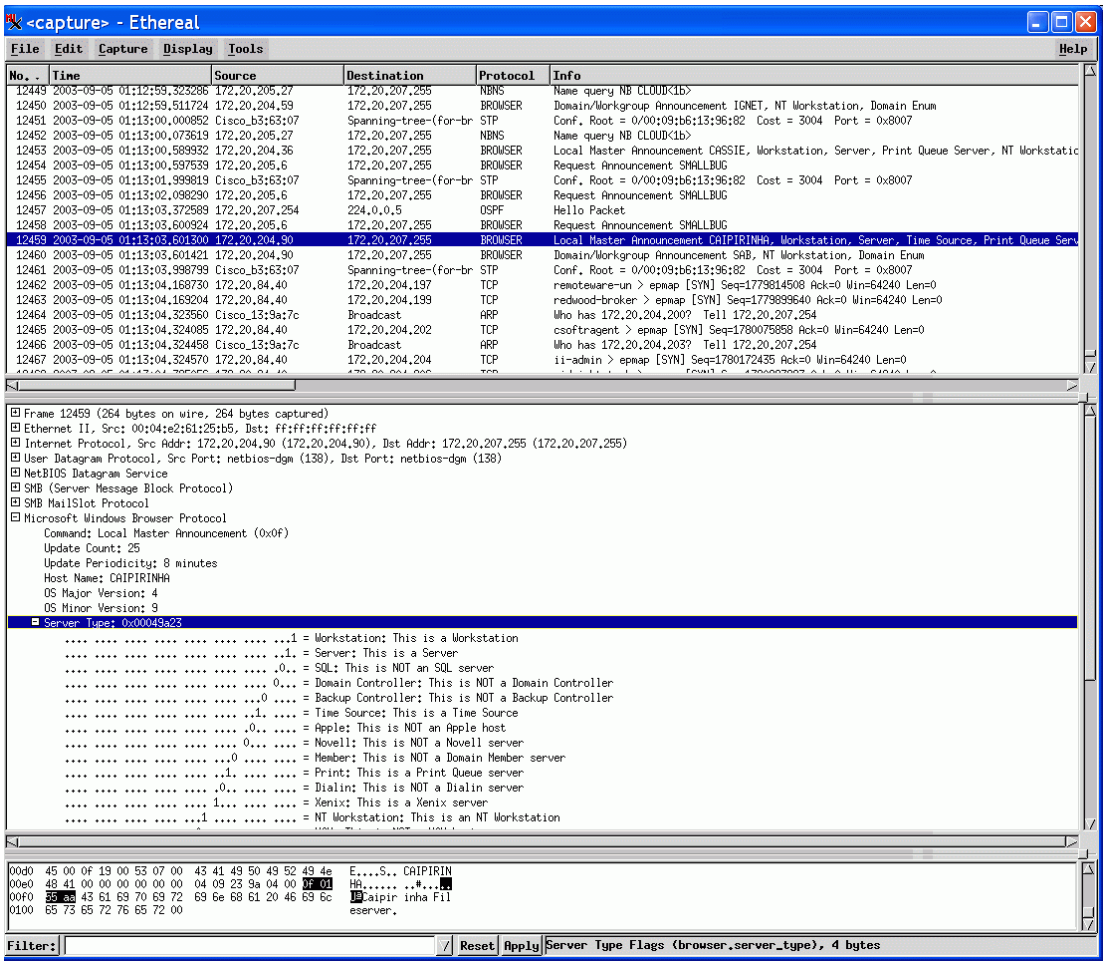


Figure 1: Network Traffic on a Network Segment at Graduate Hall (NTU)

INTERNET PROTOCOLS

The OSI 7 layer model is the underlying framework for the internet protocols. Of the OSI model, layers 1-4 are of special importance. While layer 1 describes the physical network interface (cable type, levels, signal timing), layer 2 deals with the connectivity of network interfaces. Layer 3 is where one would put the 32 bit long internet protocol (IP) addresses. When talking of IP addresses, one must actually know two numbers, the IP address itself and the corresponding network mask. Besides IP, other protocols like IPX exist, however, more and more machines exclusively use IP. An IP address is an address for a networked machine on the internet. While in the early days networks were separated into class A, B, C, D and E networks according to the first bits in the IP address [1], the current approach of subnetting is using a netmask that does not necessarily have to correspond to a class A, B, or C network. Therefore, it is necessary to give the netmask together with the IP address when configuring a machine. An example is the network 172.20.204.0/22 at NTU which indicates that the network contains all IP addresses from 172.20.204.1 until 172.20.207.254 and the netmask is 255.255.252.0 (the first 22 bits are "1"). The IP address 172.20.204.0 is used for the network itself, while 172.20.207.255 is the broadcast address of the network.

A networked machine can offer network services to other machines. In order to be able to serve multiple clients or to host several applications, a networked machine must needs to separate the incoming and outgoing data streams. Therefore, the concept of a port was introduced, which is essentially a 16-bit number used to address a service on a machine.

Networked machines can send IP datagrams from one machine to the other. An IP datagram contains a destination IP address, a source IP address, the protocol type, the source port, the destination port and the payload. Layer 3 does not ensure a reliable transmission of the IP datagramns, and in congested networks, IP datagrams may gone lost or be discarded by routers whose capacity is exhausted. However, in some connections, it is necessary to ensure that all packets arrive. This task is the domain by layer 4 of the OSI 7 layer protocol. In this layer, we find protocols like the User Datagram Protocol (UDP) or the Transmission Control Protocol (TCP). UDP is a connectionless protocol and does not ensure a reliable data transmission. TCP is a protocol that ensures that all IP datagrams reach their destination and forces a re-transmit of individual packages if IP datagrams get lost. Clearly, TCP is the choice for applications like file transfer, web browsing, email, and for many other applications. UDP does not care for a re-transmission, and once lost packages remain lost. Despite this strong disadvantage, UDP is a good choice for multimedia streaming since the overhead of TCP would cause more network traffic than UDP. This additional traffic is better invested in a higher data rate for the multimedia stream, and small disturbances in the transmission are acceptable by the user. Other UDP applications are the Simple Network Management Protocol (SNMP), the Network Time Protocol (NTP) or applications like *talk*. In order to secure a transmission, TCP uses a three-way handshake called SYN, SYN/ACK, ACK handshake which is used to synchronize both parties. As Figure 1 shows, many more protocols in the layer 2 and 3 domain exist on a typical network segment. Some important other protocols are the Dynamic Host Configuration Protocol (DHCP) at layer 2 which assigns a dynamic IP number to computers on a network or the Address Resolution Protocol (ARP) at layer 3 which requests MAC addresses belonging to machines with certain IP addresses.

In the beginning of the internet, the Internet Assigned Numbers Authority (IANA) was tasked to distribute IP addresses to organizations. Later, that task was transferred to the Network Information Centre (NIC), and subsequently to regional NICs which themselves distribute the IP addresses further to national NICs. In Singapore, this is the task of SGNIC [5]. The current version of the internet protocol (v4) comes with severe limitations like the lack of the possibility to prioritize IP traffic. This is necessary for Voice-over-IP connections where a guaranteed bandwidth and low delay of the IP packets is crucial for a good voice quality. These issues fall into the category "Quality of Service" (QoS), and the upcoming IP protocol in version 6 (IP v6) addresses them. IP v6 also deals with authentication and encryption of packets on the IP layer.

The concept of Domain Name Servers were introduced because it is easier to remember names rather than IP addresses for humans. So-called Domain Name Servers (DNS) resolve names to IP addresses, and when a user enters a "web address" into the address line of his browser, the browser in fact contacts a

DNS in order to resolve the name to an IP number. The DNS are structured hierarchically, and the constant availability of the top level domain servers is crucial for the functioning of the internet.

HARDWARE ISSUES

According to the OSI 7 layer model, several hardware components are involved on transmitting IP traffic. On the terminations of network devices, we find network interface cards (NICs). Each NIC has a unique media access control (MAC) address which identifies the NIC on the network even if the NIC has not yet been assigned an IP address. Different network segments, especially with different network technologies can be connected by a bridge which still operates in the layer 2 level. A DSL modem acts as a bridge, for example, when it is used to connect a PC to the remote end of the DSL line card in the telephone exchange. Hubs also operate in layer 2 and transmit every packet that is sent by one of the computers which is connected to the hub, to every other computer. However, hubs are increasingly replaced by switches which forward the packet right away to its destination without sending it to every attached computer. This can increase the capacity of a network segment. In order to do so, switches hold a table with MAC addresses and the respective interface that these machines are connected to. Some switches also process IP addresses, and consequently, they already operate on layer 3 of the OSI 7 layer model. A component which always operates on layer 3, is a router. A router is distinguished from a switch in that its interfaces have been assigned an IP address. Routers connect subnets with different network addresses together and determine which route IP packets should take when they want to reach a computer that is not in the same network. Routers are not used to connect a few computers in the same network, that is the task of a switch. Routers are only used to connect different networks. Switches and routers range from the low end of consumer devices to expensive components that can be administered by telnet, web interfaces or via SNMP. The switch market which was threatened into entering a commodity business has seen a revival by the introduction of Gigabit Ethernet which is increasingly being employed in organisations. Routers, however, are facing top become a commodity market which is solely decided upon on the price. The largest manufacturers of routers are Cisco, Avici, Lucent, Juniper and Huawei. Therefore, router manufacturers are more and more seeking to implement value-added services into their routers that explore the higher layers of the OSI 7 layer model and offer features like QoS for certain types of traffic, intelligent routing algorithms, load balancing router concepts, redundant routing techniques, resource reservation for critical IP packet transfers, integrated firewalls, etc.

Enterprises and big organizations use Network Address Translation (NAT) in order to connect a large amount of clients to the internet by only holding a few “real” internet addresses. However, clients behind a NAT system cannot offer services to the internet actively unless static NAT is implemented in the respective device. Nevertheless, this concept is ideal for clients that only use passive applications like email or an internet browser in order to communicate with the internet. The advantage is that the clients are protected by attacks from the internet and that the organization does not have to hold a “real” IP address for every machine since IP addresses are a valuable good. Mobile devices like telephones usually are also connected via NAT to the internet.

THE WWW AND RELATED TECHNOLOGIES

Two protocols are essential in order to understand the World Wide Web (WWW). The Hypertext Transfer Protocol (HTTP) which was developed at the *Centre Européen des Reseaux Nucléaires* (CERN) in 1998 and is now maintained by the Wild Wide Web Consortium (W3C) and the IETF, is currently used in version 1.1. This protocol specified how web browsers communicate with web servers and how the involved information is transferred. Web servers and their resources are typically identified by a Uniform Resource Locator (URL), which start with http:// indicating the HTTP protocol. Web pages are typically stored in a language named Hypertext Markup Language (HTML) which is a subset of the lesser known general form of the Standard Generalized Markup Language (SGML) that had been specified long before the web became popular. While SGML focuses on the logical content rather than on the presentation of a document, HTML defines ways in order to structure a document. A sample HTML page is given at attachment to this report, and some important parts of a HTML document are highlighted and explained. HTTP and HTML must not be confused. A web server can, for example, deliver a plain text document which has nothing to do with HTML, but which is still transferred from the web server to the web

browser using HTTP. Besides the http:// header, users frequently encounter the header https:// which means that the HTTP transfer is encrypted using the Secure Socket Layer (SSL). This is effectively a server-to-browser end-to-end encryption that is mostly used to transfer sensitive data like credit card information.

Web servers usually are internet applications that are usually awaiting network requests at port 80 (http) or at port 443 (https). The most widely used web servers were Apache and Microsoft's Internet Information Server, according to [3]. An overview of the market shares is given in the annex. Apache is an open-source project which has been released in version 2 meanwhile and which has been very successful since the software is available for many different platforms (Windows, Linux, Unix, Solaris, etc.). Microsoft's Internet Information Server (IIS) is bundled with Windows 2000 and XP and is the company's attempt to get market share in the server business. Since the IIS is integrated into Windows, it makes more use of Windows features than Apache does. However, many administrators still favour Unix and Apache over Microsoft and IIS.

In the market for browsers, it is Microsoft's Internet Explorer (IE) who is the dominant browser because it comes along with the Windows operating system that is the working base of most web users [4]. A graph showing the market shares of browsers in Germany is attached in the annex. Besides IE, a small community of Linux users mostly browse with Mozilla, Netscape or Konqueror [4]. With his latest version 6, IE has reached significant stability and comfort, especially under Windows 2000 or XP.

Web pages can be located on a proprietary server of an enterprise or on a shared or dedicated web server of a web hoster. While individuals, small enterprises or enterprises in countries which lack a fast internet connection usually prefer to get their pages hosted by a professional and well-connected web hoster, large enterprises mostly opt to have the servers in-house. However, if a company is engaged in a business that leads to large file downloads, it may consider to seek help from a Content Distribution Network (CDN). Akamai and Inktomi are two of the major players who have server capacity distributed all over the world. Users who download large files from an enterprise are then redirected to a server of the CDN which is located in their geographical vicinity so that they experience faster and more reliable downloads [1]. In order to cope with the big load of all the downloads, CDNs and large enterprises make use of load balancing techniques that distribute the numerous requests to several servers that have mirrored content.

HTML was originally not designed for multimedia content, but rather as a document description language. In order to cope with the upcoming desire to include multimedia content and to connect the originally static HTML pages with dynamic contents and databases or other applications, a number of techniques have evolved. We can distinguish server-side programming languages and browser-side languages. Server-side languages are Java, Perl, Python, PHP Hypertext Processor (PHP), Microsoft's Active Server Pages (ASP), and Allaire ColdFusion . These languages usually run on the machine which acts as a web server and create dynamic content of the requested web page. A Perl script may, for example, search for key word in a number of documents that are stored on a server and hand the result back to the user. Most of the times, it is not clear upfront which action the user will request, and therefore, an interaction between the users requested pages and the called scripts or programs has to be established. This is the task of an application programming interface (API). Although many companies have come up with proprietary solutions, the most common interface is still is Common Gateway Interface (CGI) which specified how HTTP requests hand over arguments to and retrieve results from a program on a server. These programs could range from a simple shell script to a compiled C program. On the browser side, a series of languages have appeared, too. Among them are Netscape's JavaScript, Java Applets, Microsoft's Jscript, and Microsoft's ActiveX. The task of these browser-side languages is either to run small programs on the user machine or even to interact with files on the user machine (ActiveX).

HTML is a stateless language, and hence, a web server cannot store information about which pages a user has visited before. This poses a problem to eCommerce application where users may first put items into a shopping cart or to other sites where users first have to authorize themselves before being able to browse a certain amount of web pages. In order to overcome this problem, techniques like Cookies, Hidden Fields in HTML documents, and URL rewriting have been developed that can transmit the previous state of the user's browsing activity back to the web server. However, security concerns are attached to this [1].

The Extensible Markup Language (XML) has recently created some hype in the business world. XML is not a markup language on its own, but a language for describing other languages and contains a definition

part for the content and the content itself. It is intended to facilitate the display of contents on different devices and browsers and to facilitate the data exchange between applications [1]. Whether XML will live up to its expectations, however, will depend on whether companies will be able to join forces and to agree on a common proceeding or whether we will individual "extensions" to the language by certain companies like in the case of HTML or Java.

MULTIMEDIA

Increasingly, the web is used for the transfer of multimedia content, and video streaming and audio streaming are the major applications of this development. At the moment, several formats are used for audio and video streaming. Proprietary formats are Microsoft's Windows Media Format which can be reproduced with Mircrosoft's Media Player, Real Networks' Real Media Format which can be reproduced by their software RealPlayer and Apple's QuickTime format which can be replayed with the respective Apple QuickTime player. However, the most popular format for audio files that has erupted in the recent past is the MPEG Audio Layer 3 format which originally had been developed by the *Fraunhofer-Institut für Integrierte Schaltungen*. The reason is that the implementation of the decoder does not require licensing, and that a number of players are available for different platforms and devices. While this format is liked among users because of its lack of copyright protection, commercial audio and video broadcasters prefer either Windows Media Format or Real Media format and usually do not allow storing the stream on the user's computer. A video format which has recently come up and which is based on MPEG 4, is the DivX format which yields a high video compression at moderate file sizes.

The market for audio and video telephony and conferencing is still evolving. An insufficient support of QoS issues by the current IP protocol v4 which has to be overcome by workarounds, is still responsible for a limited quality in the connections. Nevertheless, VoIP is already being commercialized ranging from clients that individual users can install on their PCs to offers from professional telecommunication providers like Singtel. Cisco is pushing VoIP very much and offers a range of products. Experts are predicting a gradual move from connection-oriented phone calls to VoIP in future, due to the lower costs. Apart from videotelephony applications like Microsoft's Messenger under Windows XP or Microsoft's Netmeeting, professional videoconferences are still done in connection-oriented calls via ISDN. The major supplier of videoconferencing equipment is PictureTel [1].

LITERATURE

[1] PriceWaterhouseCoopers, "Technology Forecast: 2002-2004 – Volume 2: Emerging Patterns of Internet Computing", PriceWaterhouseCoopers, Menlo Park (CA), October 2002.

[2] Münz, Stefan, "Self-HTML", 8th Edition, 27 October 2001, http://selfhtml.teamone.de/.

[3] Keßler, Astrid and André Malo, "Der neue Indianer", c't Magazin für Computertechnik, Vol. 1 (2003), Verlag Heinz Heise, Zorneding, January 2003.

[4] Heise news, "Internet Explorer setzt sich bei deutschsprachigen Surfern durch", Heise Newsticker, 6 March 2003, http://www.heise.de/newsticker/data/anw-06.03.03-001/.

[5] SGNIC, Singapore Network Information Centre, http://www.nic.net.sg/.

AN EXAMPLE HTML PAGE

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd">
<html>
<head>
<style type="text/css">
<!--
 a:link { text-decoration:underline; font-weight:bold; color:#0000FF; }
 a:visited { text-decoration:underline; font-weight:bold; color:#800080; }
 a:hover { text-decoration:underline; font-weight:bold; color:#FFFFFF; }
 a:active { text-decoration:blink; font-weight:bold; color:#008080; }
-->
</style>
  <title>Navigationsframe</title>
  <meta http-equiv="content-type" content="text/html;charset=iso-8859-1">
  <meta http-equiv="expires" content="86400">
  <meta http-equiv="content-language" content="de">
  <meta name="author" content="Gabriel Rüeck">
  <meta name="date" content="2003-08-29T018:30:00+08:00">
  <meta name="robots" content="noindex">
  <meta name="description" content="Navigationsrahmen der Homepage von Gabriel Rüeck.">
  <base target="content">
</head>

<body text="black" background="../Grafiken/Granit.gif">

<font face="VERDANA,ARIAL,HELVETICA">
  <table align="right">
  <tr>
    <td align="right"><b><a href="Persoenlich/Ich/index.html">Persönliche Daten</a><br><font
       color="red">(privat)</font></b></td>
    <td><a href="Persoenlich/Ich/index.html"><img src="../Grafiken/Ich.jpg" width="100" height="75" border="3"
       alt="Persönliches"></a></td>
  </tr>
  </table><br clear=all>
<hr noshade size="2">
  <table align="left">
  <tr>
    <td><a href="Foto/index.html"><img src="../Grafiken/Island.jpg" width="100" height="75" border="3"
       alt="Fotografien"></a></td>
    <td><b><a href="Foto/index.html">Fotografien</a><br><font color="yellow">(teilöffentlich)</font></b></td>
  </tr>
  </table><br clear=all>
<hr noshade size="2">
  <table align="right">
  <tr>
    <td align="right"><b><a href="Reise/index.html">Reiseberichte</a><br><font color="red">(privat)</font></b></td>
    <td><a href="Reise/index.html"><img src="../Grafiken/Travel.jpg" width="100" height="75" border="3"
       alt="Reiseberichte"></a></td>
  </tr>
  </table><br clear=all>
<hr noshade size="2">
  <table align="left">
  <tr>
    <td><a href="Persoenlich/freunde.html"><img src="../Grafiken/Freunde.jpg" width="100" height="75" border="3" alt="Freunde
       und Kollegen im Web"></a></td>
    <td><b><a href="Persoenlich/freunde.html">Freunde und Kollegen im Web</a><br><font
       color="green">(öffentlich)</font></b></td>
  </tr>
  </table><br clear=all>
<hr noshade size="2">
  <table align="right">
  <tr>
    <td align="right"><b><a href="Spass/unworte.html">Unworte</a><br><font color="green">(öffentlich)</font></b></td>
    <td><a href="Spass/unworte.html"><img src="../Grafiken/m.jpg" width="100" height="75" border="3" alt="Unworte"></a></td>
  </tr>
  </table><br clear=all>
<hr noshade size="2">
  <table align="left">
  <tr>
    <td><a href="Spass/saetze.html"><img src="../Grafiken/Kraha.jpg" width="100" height="75" border="3"
       alt="Firmenmitteilungen"></a></td>
    <td><b><a href="Spass/saetze.html">Firmen-<br>Mitteilungen</a><br><font color="green">(öffentlich)</font></b></td>
  </tr>
  </table><br clear=all>
<hr noshade size="2">
  <table align="right">
  <tr>
    <td align="right"><b><a href="MBA/index.html">MBA-Studium (englisch)</a><br><font
       color="green">(öffentlich)</font></b></td>
    <td><a href="MBA/index.html"><img src="../Grafiken/MBA.jpg" width="100" height="37" border="3" alt="MBA-
       Studium"></a></td>
  </tr>
  </table><br clear=all>
<hr noshade size="2">
  <table align="left">
  <tr>
    <td><a href="Krypto/index.html"><img src="../Grafiken/Schluessel.jpg" width="100" height="75" border="3" alt="Krypto-
       Schlüssel"></a></td>
    <td><b><a href="Krypto/index.html">Krypto-Schlüssel</a><br><font color="green">(öffentlich)</font></b></td>
  </tr>
  </table><br clear=all>
<hr noshade size="2">
<p align="center"><a href="http://www.banrap.com" target="_blank"><img src="../Grafiken/banrap.gif" width="260" height="29"
       border="0" alt="banana rappresentanza &ndash; Photographen Agentur"></a></p>
<p align="center"><a href="http://www.bilderwelt.net" target="_blank"><img src="../Grafiken/Bilderwelt.gif" width="146"
       height="75" border="0" alt="Annekes Bilderwelt"></a></p>
<p align="center"><a href="http://www.dilbert.com/" target="_blank"><img src="../Grafiken/dilbert.gif" alt="Get Dilbert!"
       border="0" width="120" height="90"></a></p>
```

```
<p align="center"><a href="http://petition.eurolinux.org" target="_blank"><img src="http://aful.org/images/patent_square.gif"
       border="0"></a></p>
<p align="center"><a href="http://www.eff.org/blueribbon" target="_blank"><img src="http://br.eff.org/br.gif" alt="[Blue
       Ribbon Campaign icon]" height="76" width="112" border="1" align="middle"><br>
Join the Blue Ribbon Online Free Speech Campaign!</a></p>
</font>

</body>
</html>
```

For demonstration purposes, some text sections is this HTML code have been highlighted:

Red underlined text refers to the document declaration which is necessary to identify this page as a proper HTML page. This declaration is unfortunately missing often. It refers to a parser of the W3C which the browser could theoretically take in order to verify the syntactical validity of the HTML code. Few HTML pages are error-free, and so the browsers usually omit this step in the processing [2].

Yellow underlined text refers to a CSS declaration for the links and indicates how links shall be displayed. Four possibilities are indicated, depending on whether the link has not yet been visited, whether it has been visited, whether the mouse pointer is over a link or whether the user clicks on a link. This declaration should normally be held outside the HTML document and be referred to, in order to harmonize the appearance of all web pages.

Grey underlined text indicates the title which is displayed in the browser. However, since this page is a subpage in a frameset, this has no influence in this document.

Green underlined text refers to the meta data of the web page [2]. This information is also often missing, although it contains important information for the browser and search robots: In this document, we can find:

- The character set which is necessary in order to display all the umlaute correctly (unless they are coded in their HTML proper syntax like 'ä' for 'ä'), in this example: ISO8859-1 (Western Europe). The language descriptor is very important if a page uses a non-Latin chcracter set like Thai, Chinese, Vietnamese, Japanese, etc. so that the browser can display the site correctly.
- The expiry date which indicates after how many seconds a cache server shall declare the cached context for invalid and reload the page from the server, in this example: 1 day.
- The content language of the document, in this example: German.
- The author of the document.
- The date of the last modification, the time, and the time zone in UMT+Offset. Time is given in the 24h time format at the local time, and the times zone in this page is GMT+8h (Kuala Lumpur, Singapore, or Beijing).
- An indication for search robots that this page shall not be processed by them.
- A description of the content of the page.

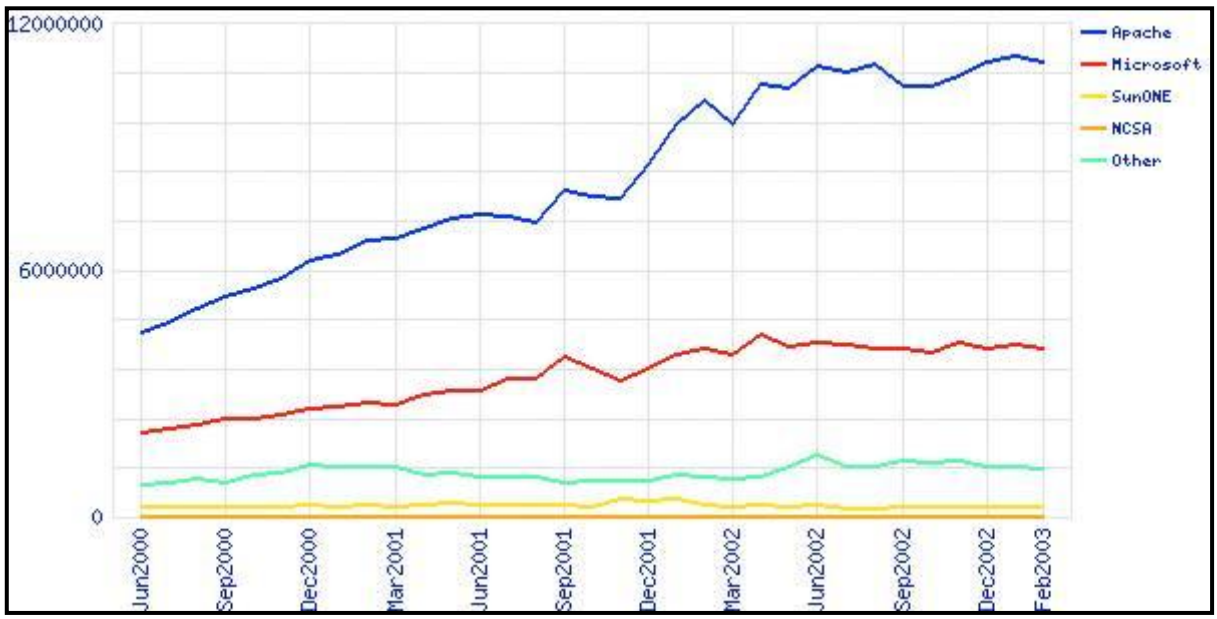
Cyan underlined text refers to a font description. Different fonts are mentioned, and the browser looks up which fonts the underlying operational system offers. The first matching font is used for the display. Windows machines usually contain Verdana, Linux machines at least Arial or Helvetica. If no matching font is found, a substitute font is used.

Pink underlined text refers to an example of a hyperlink. The highlighted hyperlink is relative to the current directory of the webpage in this example.

Ochre highlighted text refers to a link to an image that is loaded when the page is displayed. In order to help the browser in building up the layout of the webpage, the link contains the size of the image and a text declaration that is displayed as long as the image is not yet loaded.

Grey highlighted text refers to a table declaration which is aligned to the left side of the page in this example.

MARKET SHARES OF WEB SERVERS



Source: *www.netcraft.com/survey/*

MARKET SHARES OF WEB BROWSERS