Course Overview

Web technologies

- Client - Server Architecture
- HTML and beyond

The structure of the web

- WWW as a graph.
- Web search.
- Ranking webpages.

Beyond the web

- Social networks.
- Advertising and Auctions.
- Elements of text analysis.

1 The web architecture

Web architecture

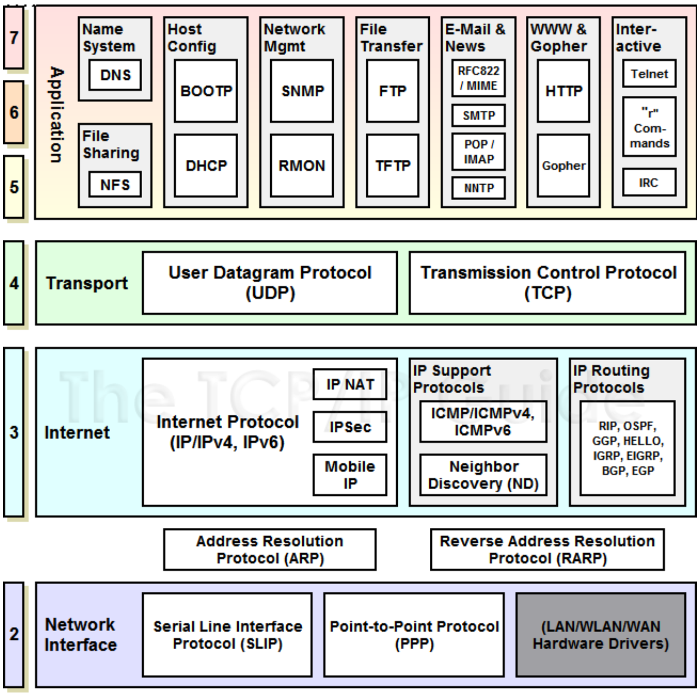
Point-to-point communication

- When browsing the web, we obtain data from a remote machine.
- `TCP/IP`: Provides *point-to-point* communication functionality.
- `HTTP`: The web *application* level protocol.
- `HTML`: The web-page *data format*.

The Layers

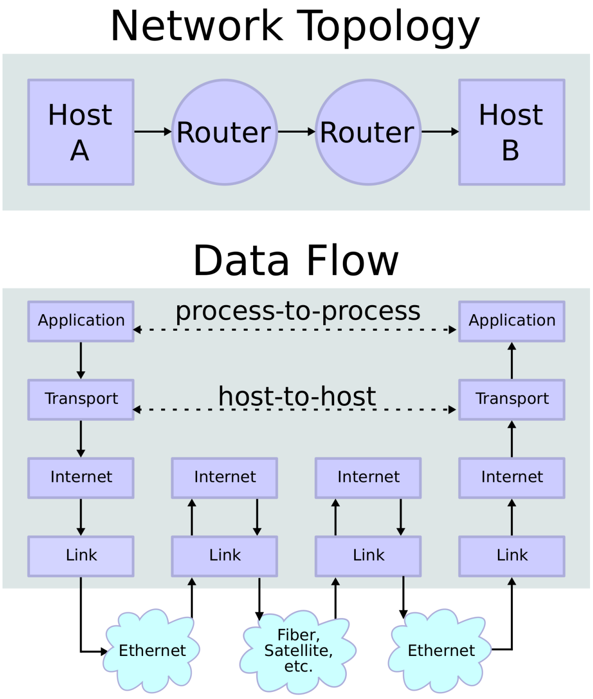
- Physical layer.
- Transport/Network layer.
- *Application layer.*

Common Internet protocols



The above is an overview of the different protocols employed on the internet to establish communication between different applications. The lower levels protocols deal with physical connectivity and network management. In this course, we are only concerned with the HTTP protocol for serving web pages.

Connections through the TCP/IP stack



A simplified view of the process that occurs in any given protocol is above. Any two computers communicate through intermediary nodes. The actual data flows through the nodes as shown in the figure below. The application creates the top-level data packet, and then sends it to the layer below. This layer then wraps the original data into another layer. So, the lowest level layer has the largest amount of data to send.

The HTTP protocol and the HTML format

Open a shell and type the following

- `telnet www.cse.chalmers.se 80`
- `GET`

This should return you the top-level web page, with a redirect   Now type

- `telnet www.cse.chalmers.se 80`
- `GET http://www.cse.chalmers.se/ chrdimi/`

This should return you a webpage in HTML format. A specialised client, such as Chrome, IE, Lynx, Mozilla (Firefox/Iceweasel), Opera, Safari, then has the job of rendering the HTML nicely.

The page is a mixture of text and HTML tags. The most important tag is

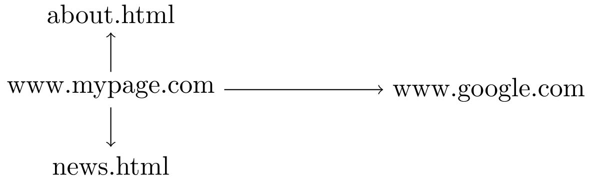
```
<A HREF=’’link’’>text of link</A>
```

This shows links to other web-pages (other methods are possible)

In order to obtain a webpage, the client connects to the target computer via TCP. When it connects, the system wakes up the web-server and notifies it of a new connection. Now the client can talk directly with the server. This communication is governed by the HTTP protocol, and most importantly, the (HTTP) GET command. The webpage usually refers to other elements, such as images (which most clients obtain automatically) and other webpages (which are normally only downloaded after user input).

The link structure of HTML pages

An HTML page is composed of links, text, images, and other elements. In this course we are mainly interested in the *structure* of the web itself. This is expressed through links between webpages. We can visualise links graphically


about.html
www.mypage.com
www.google.com
news.html