Techniques de Programmation Internet

Année Spéciale Informatique

ENSIMAG 2008-2009

James L. Crowley

Séance 1 19 janvier 2009

Introduction à l'Internet et ces Protocoles

Plan :

Organisation :

Les notes et les exemples sont disponibles surs :
http://www-prima.imag.fr/Prima/Homepages/jlc/Courses/

Objectifs du cours TPI

1. Aborder les différents problèmes techniques liés à la gestion d'applications dédiées au Web.
2. Présenter HTML et son usages.
3. Présenter les principes de programmation CGI
4. Introduire les langages Perl, PHP, MySQL

Introduction au World-Wide Web

World-Wide Web : système d'information hypermédia.

Un peu d'histoire :

Hypertexte

Le concept d'hypermédia est né dans les années 1930 dans les idées de Vanevar Bush.

Bush etait Président de MIT 1929-39. Ensuite il a créé un agence gouvernementale pour financer la recherche. Pendant la guerre de 1939 - 1945, cette agence a lancé les projets technologique comme l'invention de RADAR et la bombe atomique.

Dans l'après-guerre, Bush a proposé la création du National Science Foundation Américaine, pour financer la recherche de base.

En 1945 Bush a publié un article populaire dans lequel il a proposé un bureaux mécaniques qui s'appellent la "Memex". Memex fut une sorte de ordinateur personnel avec les références croisés des documents.

En 1960, Ted Nelson a démontré un système de Hypermédia dénommé *Xanadu*. Mais le système etait MonoUtilisateur, et l'interaction etait fait par texte (le souris n'etait pas encore inventé, et le mémoire pour un écran bitmap etait hors de prix.).

Il a fallu l'Internet et l'interaction "GUI" pour son réussit.

Hypertexte & Hypermédia

hypertexte : document-texte contenant des liens vers d'autres parties du document ou vers d'autres documents

lien hypertexte ("hyper-lien") : formé d'une ancre, mot, groupe de mots, images, ... mis en évidence et d'une adresse vers le document cible

document hypermédia : document hypertexte contenant en plus des images, du son, de la vidéo.

L'Internet et ces protocoles

En 1960, J.C.R. Licklider a publié un article populaire "Man Machine Symbiosis" dans lequel il a argumenté pour un réseau mondial d'ordinateurs.

En 1662-1964, Licklider a dirigé l'ARPA. Un de ces premier projets etait un réseau d'ordinateurs : L'ARPA net.

En 1969 L'ARPA Net a lié Univ. California at Berkley, Univ. California at Los Angeles, Stanford Research Institute et Univ of Utah. par 1971 l'Arpanet été composé de 13 ordinateurs avec les nœuds aux USA, Angleterre et France.
Sa croissance etait très rapide, avec 57 nœuds en 1977.

Le premier Courrier Électronique (E-Mail) fut testé en 1971.
En 1975 il y avait 1000 utilisateurs du E-Mail sur une centaine d'ordinateurs.

En 1985, l'Arpanet est devenu NSFnet. En 1988, une loi proposée par le sénateur Al Gore a ouvert le NSFnet au monde et aux organismes Commerciales. Le NSF est devenu l'"Internet".

L'Internet est un réseau mondial d'ordinateurs communicants par les messages codés en "packet" selon le protocole "IP". (Internet protocole).

Les communications sont gérées par un protocole : "Transfer Control Protocol".

Les concepts et protocoles de l'Internet sont définis dans les RFC ("Request for Comment"). Ceci est une tradition datant de l'année 1969 au début de l'ARPANET. Les RFC sont gérés par la "Internet Architecture Board" (IAB).

Le IAB publie les RFC de spécification intitulé "Internet Official Protocol Standards" (IOPS). L'Internet est défini par les IOPS. Les plus récents est RFC 2300 (May 1998).

Le Web

Le "World Wide Web" (WWW) fut né en mars 89 à l'initiative de Tim Berners-Lee pour la communication au sein de la communauté scientifique du CERN.
Son nom etait le "Mesh".

Décembre 1990, un premier prototype fut réalisé.
Novembre 1992, 26 serveurs etait disponibles
Février 1993 le navigateur "Mosaic" etait distribué gratuitement par NCSA
Mars 1994, Netscape fut fondée (au nom "Mosaic Communications Corp).
L'été 1994 il y avait 1500 serveurs. Le W3C (World Wide Web Consortium) été fondé par CERN et MIT à fin de gérer WWW. En 1995, CERN a donnée le contrôle du W3C au INRIA.

1995 - Microsoft a sorti "Internet Explorer", Netscape a sorti Netscape 2.0.

Définition officielle du WWW :

"wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents "

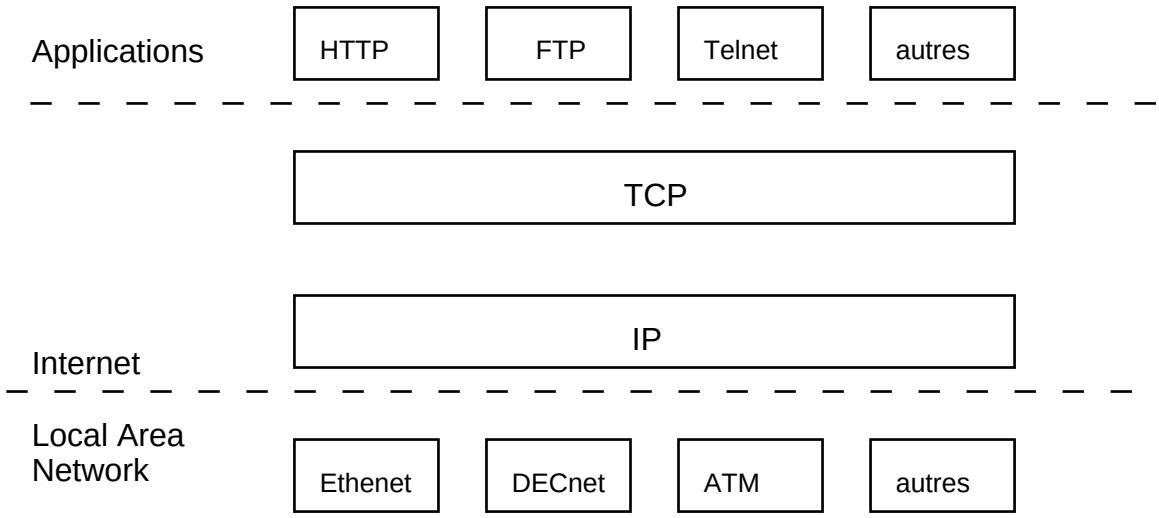
- Accès universel aux bases d'informations
- Accessibles à tous et partout
- Liens entre les informations (hypermédia)
- Informations sous différents formats (multimédia)
- Navigation aisée et rapide
- Supporté par internet "Networks of Networks" 15M d'utilisateurs, 50 pays

Les Concepts Clefs :

HTTP : HyperText Transfer Protocol - le protocole de communication de documents de Hypermédia

HTML : Hyper Text Markup Language - Language de composition de la présentation d'information.

Les Protocoles IP et TCP



IP : Internet Protocol - "A Connectionless, best-effort, Packet Switching Protocol".
Définit par RFC 791 en septembre 1981. Concu pour la transmission de "DataGrams" (Packets) entre machines au travers un réseau.

TCP: Transmission Control Protocol - A flow controlled, connection oriented, end-to-end reliable protocol designed to use IP.
TCP est defini par RFC 793, (1981).
TCP fourni une canaux de communication entre processus.
TCP ajout un "port" au addresses des machines.
Il peut fonctionner au travers les sortes de connections tres varié.

HTTP: HyperText Transfer Protocol
Le protocole HTTP est un protocole client/serveur permettant l'échange rapide de données pour les systèmes d'information intégrant des ressources distribuées de type multimédia.

FTP : File Transfer Protocol. Une protocol de communication de fichiers

MIME : Multipurpose Internet Mail Extension

Le format MIME est le format de transfert des informations de type multimédia.
A l'origine, Internet E-mail etait definit uniquement pour l'ascii, par RFC 822.
RFC 2045, 2046, 2047 et 2048 et 2049 ont definit une ensemble d'extensions connus comme MIME afin de communiquer les messages de format hybrids.
MIME etait adapté pour le WWW afin de permetre une contenus multi-média.

- Transfer d'informations de n'importe quel type (images, sons, textes formatés,...)
- Compatibilité avec les formats existants
- Ouvert aux formats à venir

Message MIME :

Message = entête + corps

Entete = Mime-version + Content Type + Content Transfer Encoding + Content ID + Content Desciption.

Numéro de version (courante : 1.0)

Type du contenu
- TEXT
- MULTIPART (combinaison de plusieurs parties)
- APPLICATION (binaire)
- MESSAGE (message encapsulé)
- IMAGE
- AUDIO
- VIDEO

MIME offre la possibilité de definir les nouveau types dynamiquement.

Les Types MIME prédefinit.

TEXT
- plain : texte non formaté
- richtext : texte enrichit d'un traitement de texte

APPLICATION
- octet-stream : données de type binaire
- postscript : programme postscript
- oda : informations encodées selon le standard ODA

IMAGE: jpeg / gif

AUDIO: basic (PCM, 8bits, 8kHz, monocanal)

VIDEO : mpeg

MESSAGE
- rfc822 : message rfc822 (de type mail)
- partial : gros message tronçonné
- external body : informations pour accéder aux données

MULTIPART
- mixed : parties indépendantes à lire séquentiellement
- alternative : une information / différentes représentations
- digest : messages compatibles avec la norme rfc822
- parallel : parties à lire simultanément

Codage de caractères accentués

```
Mime-Version: 1.0
Content-Type: text/plain ; charset=ISO-8859-1
Content-Transfer-Encoding: 8BIT
é è ç ô î
```

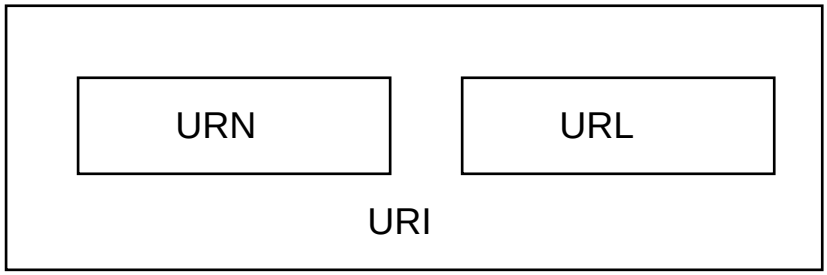
Codage pour une image

```
Mime-Version: 1.0
Content-Type: image/tiff ; name=monimage.tiff
Content-Transfer-Encoding: base64
```

Codage multi-parties

```
Mime-Version: 1.0
Content-Type: multipart/mixed ;
boundary="PART-BOUNDARY=.19801081431.ZM7315.raminis"
--PART-BOUNDARY=.19801081431.ZM7315.raminis
Content-Type: text/plain; charset=us-ascii
--PART-BOUNDARY=.19801081431.ZM7315.raminis
Content-Description: Text
Content-Type: text/plain ; name="monfichier.txt" ; charset=us-ascii
--PART-BOUNDARY=.19801081431.ZM7315.raminis
Content-Description: JPEG Image
Content-Type: image/jpeg ; name="monimage.jpg"
Content-Transfer-Encoding: base64
```

Universal Resource Identifier (URI)



URI - Universal Resource Identificateur

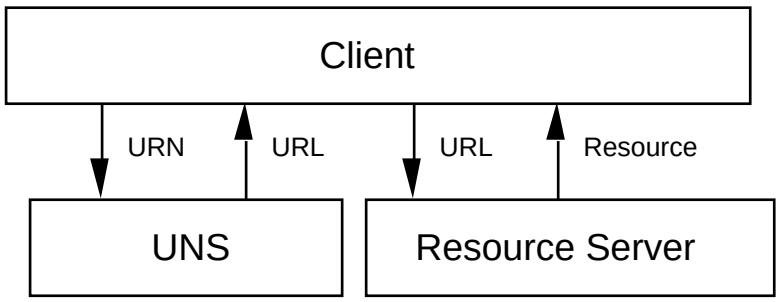
A universal set of names and addresses in a registered name space.

URL - Universal Resource Locator

The association of a phsycal address of an object and an acess scheme.
Exemple : pour les URL de HTTP, HTTP est la scheme d'adressage.

URN - Uniform Resource Name

A persistant, globally unique name assigned to an object.
Réalisé par une service de nommage qui associe une URN a une URL.



URI Syntaxe

"%" (Percent) Caractere d'Escape

"/" (Slash) - Indicateur de Hiérarchie.

"#" (Hash) - Identificateur de Fragment (ex les anchors).

"?" (Query) - Delimiteur de interrogation.

"+" utilisé pour les espace.

URL Syntaxe

Definit par RFC 1738 et 1808.

Syntaxe :

"//" [User [":" Password] "@"] host [":" port] "/" URL-PATH

Les exemples des "Schemes"

FTP - File Transfer Protocol (RFC 959).

Port par defaut - 21

User par defaut : Anonymous avec Password un address de Email.

HTTP : HyperText Transfer Protocol (HTTP).

Port par defaut - 80.

telnet : Une scheme de terminal a distance.

Port par defaut - 23

Exemples d'URL

```
file://
file:///Macintosh HD/Jim/WWW/jlc.html
ftp://ftp.imag.fr
news:imag.ragot
http://www-ufrima.imag.fr/
http://www-ufrima.imag.fr/FORMATION/DESS-GI/INFO-PLUS/
Plaquette/dessgi-plaquette.html#Association
telnet://babbage.imag.fr
http://cgest.grenet.fr/cgi-bin/a?name=Crowley
```

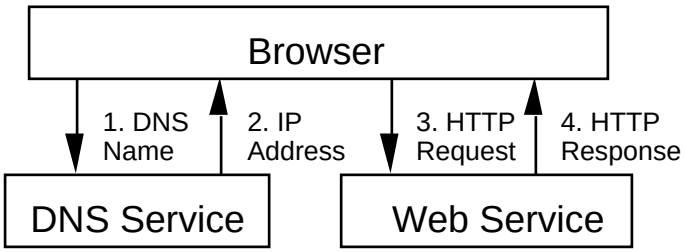
IP Address : une adresse de 32 bits qui est un identificateur unique pour chaque nœud du réseau Internet. Les formes, défini en RFC 791, sont de la forme
A.B.C.D

ex : le www.w3.org est 18.23.0.22

L'Amphi E de l'ENSIMAG est 195.221.228.31

Le DNS de l'ENSIMAG est 195.221.228.2

DNS : Domain Name Server - Crée en 1984. Actuellement définit par RFC 1034 et RFC 1035. Les formes symboliques de l'URL sont transformées en adresse exacte par un serveur de noms (le DNS). Le nom symbolique permet la migration des services entre machines physiques. (Par exemple www-prima.imag.fr vient de passer de pandora.imag.fr à sinope.imag.fr sans que les acess soit perturbé).



Les "Top Level Domaines" (TLD) sont la partie la plus abstrait des URLs. Il existe les TLD pour les pays (ccTLD) et les TLD génériques (gTLD). Les ccTLD sont gérés par les autorités national de chaque pays.

Les gTLD sont définis par RFC 1591 comme :

edu - "Education" - Réservé au Universités

com - "Commercial" - les organisations commerciaux (indépendant du pays).

net - "Network" - Les administrateurs des réseaux et les fournisseurs de services de réseau (Internet service providers).

gov - "Governmental" - Les agences de la gouvernement Américain.
Pour les autres pays, il faut le code du pays. Par exemple, en France en trouve *.gouv.fr

mil - "Military" - Le militaire des États-Unis

int - "International" - Les organismes internationaux crée par traité entre pays.
Exemple : La nations unies, ou la Commission Européenne.

org - Organisation qui sont ni commerciaux, ni gouvernementaux

HTTP : Hyper Text Transfert Protocole

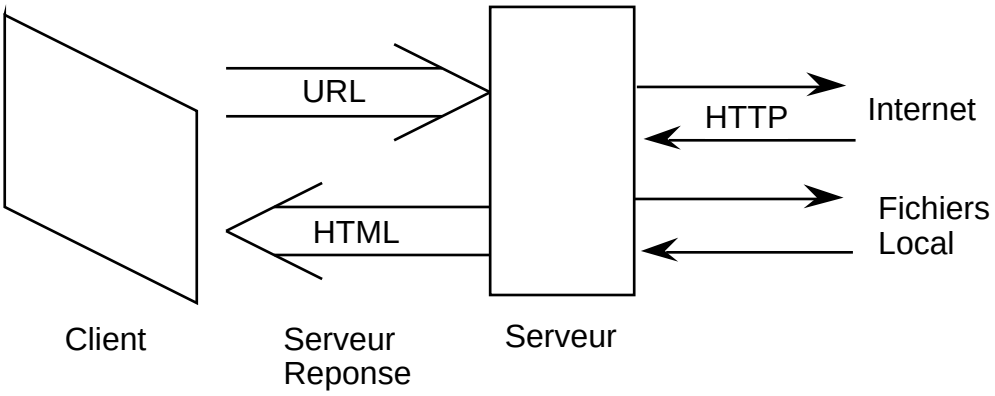
Modèle client-serveur pour le transfert des documents hypertextes.

Protocole utilisé par les serveurs WWW depuis 1990.

Échange de messages codés dans un format similaire au type MIME.

Pourquoi un nouveau protocole?

1. Transfert de fichiers,
2. Recherche par requête,
3. Négociation automatique de format du entre client et serveur,
4. Capacité de reporter le client sur un autre serveur.



URL = Uniform Ressource Locator

methode://machine[:port]/fichier[#ancre|?params]

file	accès local ou protocole FTP
ftp	protocole FTP
http	protocole HTTP
telnet	session interactive TELNET
gopher	protocole GOPHER
wais	version WAIS du protocole Z39.50
news	protocole NTTP
mailto	adresse de courrier électronique

Exemple :

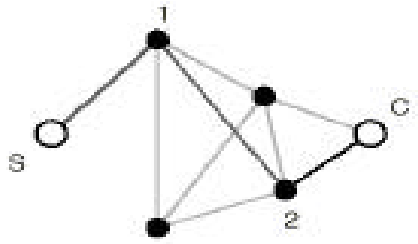
On peut exécuter des commandes Unix dans un page html.
Le page html suivant execute le commande shell "mailto".

```
<HTML>
<TITLE>Send me mail</TITLE>
<BODY>
Click here to <A HREF="mailto:jlc@imag.fr">send me email</A>
</B>
</BODY>
</HTML>
```

Exemples d'URL

file://
file:///Macintosh HD/Jim/WWW/jlc.html
ftp://ftp.imag.fr
news:imag.ragot
http://www-ufrima.imag.fr/
http://www-ufrima.imag.fr/FORMATION/DESS-GI/INFO-PLUS/Plaquette/dessgi-plaquette.html#Association
telnet://babbage.imag.fr
http://cgest.grenet.fr/cgi-bin/a?name=Crowley&prenom=James

cgest.grenet.fr est traduit en numéro IP par une "Name_Serveur".
exemple d'un numéro IP : 195.221.224.119 (ima-118.imag.fr)
Les requêtes URL Transit l'Internet

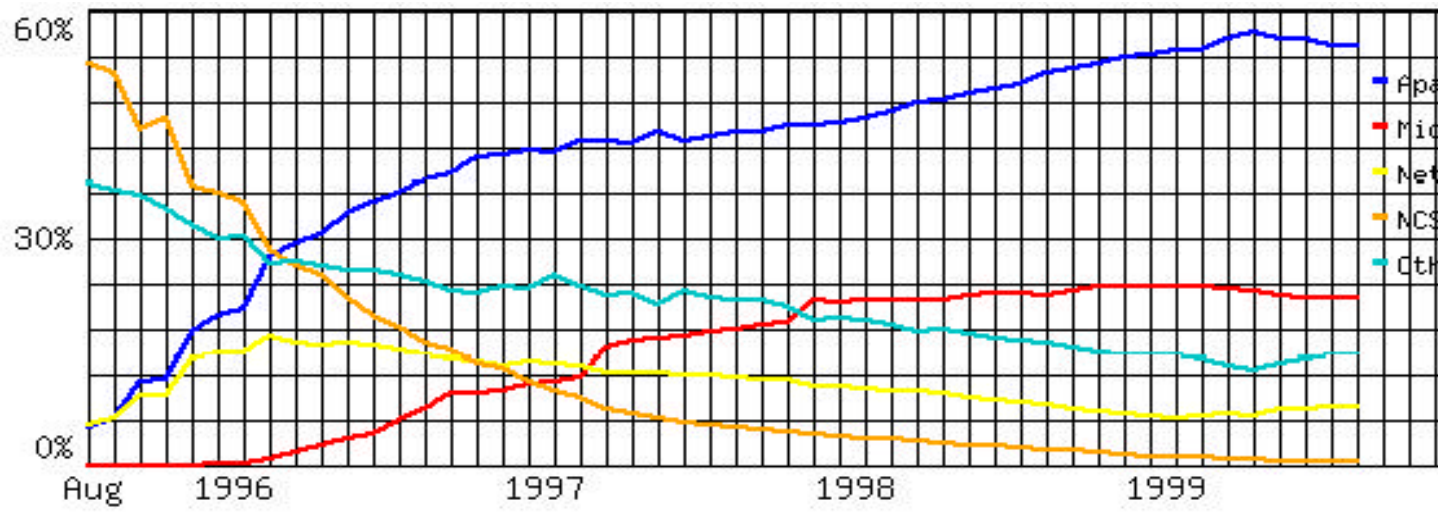


S : Serveur HTTP C : Client 1, 2 : Machines intermédiaires.
Les machines intermédiares disposent d'un cache.

Quelques serveurs HTTP

(etat en 2000 - maintant APACHE domine completement le marché).

Serveur	Plate-Forme(OS)	Caractéristiques.
Apache	Unix	Directives d'insertion, accès restreint par mot de passe, nom de domaine, adresse IP, support de SSL, gratuit
Microsoft	WinNT	Support de SSL, interface graphique, support produit Microsoft
Netscape	WinNT/Unix	Archivage des accès, directives d'insertion, accès restreint par mot de passe, nom de domaine, adresse IP, support de SSL, interface graphique
NCSA	Unix	Directives d'insertion, accès restreint par mot de passe, nom de domaine, adresse IP, gratuit chez O'Reilly



Source : "http://www.netcradt.com/survey/"

Installation d'un serveur HTTP

1. Installation du démon,
2. Configuration du serveur (fichier http.conf),
3. Création des pages (au moins la première),
4. Exécution du démon.

Exemple : Configuration du démon Apache.

Configuration du serveur Apache (1)

- ServerName pandora.inrialpes.fr
  ServerAdmin Jerome.Martin@inrialpes.fr
  Nom du serveur et courrier de son administrateur.

- ServerType standalone
  Type de serveur : standalone ou inetd

- Port 80
  Numéro du port du serveur.

- User nobody
  Group nogroup
  Nom et groupe de l'utilisateur du serveur

- ServerRoot /www/http/apache_1.3b2/
  DocumentRoot /www/html/
  Racine d'installation du serveur et des pages html

- TransferLog logs/access_log
  ErrorLog logs/error_log
  Emplacement des fichiers de trace des accès et erreurs

- PidFile logs/httpd.pid
  Fichier contenant le numéro de processus du démon.

- Timeout 300
  KeepAliveTimeout 15
  Temps d'attente avant d'envoyer un timeout ou entre deux requêtes.

- MaxClients 150
  Limitation du nombre de clients pouvant se connecter au serveur.

Connexion HTTP directe

Transaction HTTP:

1. Le client envoie une requête (port 80 du serveur)
2. Le serveur fournit ou non la ressource demandée
3. Fermeture de la connexion

En-tête de la requête HTTP

From	Contient l'adresse électronique de l'utilisateur-client
If-Modified-Since	Permet à la méthode GET d'accéder à la ressource si celle-ci a été modifiée depuis la date donnée
Referer	URL d'origine de la requête
User-Agent	Informations sur le client (pour maintenir des statistiques ou adapter la réponse selon le client)

Descriptif de la requête HTTP

Accept	Liste des types MIME supportés par le client (exemples : image/gif, image/jpeg)
Content-Encoding	Description du codage appliqué au corps de la requête
Content-Length	Taille en octets du corps de la requête
Content-Type	Le format MIME du corps de la requête

Statut de la réponse HTTP

Informe le client sur le déroulement du traitement de la requête par le serveur.
4 classes de réponses :

classe 2 :	Succès,
classe 3 :	Redirection / traitement incomplet,
classe 4 :	Erreur client,
classe 5 :	Erreur serveur.

Statuts de classe 2

Succès de la requête

200 OK	Requête a été traitée
201 Created	.Requête a été traitée et a abouti à la création d'une nouvelle ressource.
202 Accepted	La requête a été reçue et est en cours de traitement. La connexion peut être interrompue
204 No content	La requête a été traitée mais ne contient pas de document.

Statuts de classe 3

Redirection, traitement incomplet de la requête

301 Moved Permanently	La ressource a été assignée à une nouvelle adresse. L'URL est donnée par le champ Location .
301 Moved Temporarily	La ressource a été assignée temporairement à une nouvelle adresse. L'URL est donnée par le champ Location .
304 Not Modified	La ressource n'a pas été modifiée depuis la date précisée par le champ If-Modified-Since de la requête.

Statuts de classe 4

Erreur client

400 Bad Request	Erreur de syntaxe.
401 Unauthorized	La requête nécessite une identification préalable de l'utilisateur.
403 Forbidden	Le serveur refuse de traiter la requête.
404 Not Found	Le serveur n'a pas trouvé la ressource demandée.

Statuts de classe 5

Erreur serveur

500 Internal Server Error	Erreur propre au serveur.
501 Not Implemented	Le serveur ne possède pas la fonctionnalité pour traiter la requête.
502 Bad Gateway	Le serveur agissant en tant que gateway ou proxy n'a pas pu traiter la requête.
503 Service Unavailable	Le serveur n'est pas en mesure de traiter la requête pour des raisons de surcharge ou de maintenance.

En-tête de la réponse HTTP

Location	Identifie l'URL exacte de la ressource demandée.
Server	Informations sur le serveur sollicité.
Exemple :	Server: CERN/3.0 libwww/2.17

Descriptif de la réponse HTTP

Content-Encoding	Description du codage appliqué au corps de la requête.
Content-Length	Taille en octets du corps de la requête.
Content-Type	Le format MIME du corps de la requête.
Date	Date et heure de la génération de la réponse.
Expires	Date et heure d'expiration du document.
Last-Modified	Date et heure de la dernière modification du document.

Exemple de réponse correcte

Requête :

```
GET /fichier1.html HTTP/1.0
Accept: text/html
```

Réponse :

```
HTTP/1.1 200 OK
Date: Fri, 09 Jan 1998 09:49:11 GMT
Server: Apache/1.3b2
Last-Modified: Tue, 19 Aug 1997 11:57:17 GMT
Content-Length: 118
Accept-Ranges: bytes
Connection: close
Content-Type: text/html
<HTML>
<HEAD>
<TITLE>Ceci est le titre de ma page</TITLE>
</HEAD>
<BODY>
Ceci est le corps du document
</BODY>
</HTML>
```

Exemple de réponse d'erreur

Requête

```
GET /toto.html HTTP/1.0
Accept: text/html
```

Réponse

```
HTTP/1.1 404 Not Found
Date: Fri, 09 Jan 1998 09:51:35 GMT
Server: Apache/1.3b2
Connection: close
Content-Type: text/html
<HTML>
<HEAD>
<TITLE>404 Not Found</TITLE>
</HEAD>
<BODY>
<H1>Not Found</H1>
The requested URL /toto was not found on this server.<P>
</BODY>
</HTML>
```

Cookies HTTP

Mécanisme de stockage d'informations chez le client pris en compte par le serveur à chaque accès.

Exemples d'utilisation :

- Sauvegarde d'option,
- Validité d'accès à un serveur payant,

Initialisation d'un cookie par le serveur HTTP

Set-Cookie: expires = <DATE>; path = <CHEMIN>;
domaine = <NomDomaine>; secure

expires=Date	Date d'échéance du cookie.
domaine=NomDomain	Identification du cookie correspondant au serveur accédait. Par défaut, il correspond au serveur HTTP.
path=Chemin	Association du cookie à un ensemble de ressources.
secure	Utilisation d'une connexion client-serveur serveur sécurisé (protocole HTTPS)

Insertion d'un cookie par le client

Lorsque le client établit une requête pour accéder à une URL, il recherche parmi les cookies mémorisés ceux s'appliquant au serveur/URL.

La requête contient une ligne avec les paires nom/valeur correspondantes :
Cookie: Nom1=Valeur1; Nom2=Valeur2; ...

A noter sur les cookies

- Plusieurs directives Set-Cookie peuvent êtres insérés par le serveur,
- Un client peut mémoriser 300 cookies de taille maximale de 4000 octets et 20 cookies par serveur
- Un script CGI peut effacer ou remettre à jour des cookies en spécifiant l'attribut expires une date expirée.

Exemple de transaction avec un cookie

Le client établit une requête,
La réponse du serveur est :
Set-Cookie: CUSTOMER=WILE_E_COYOTE; path=/; expires=Wednesday,09 Nov-99 23:12:40GMT
Le client accède aux URLs du serveur en insérant :
Cookie: CUSTOMER=WILE_E_COYOTE