

Sistemas de Informação
Disciplina: Fundamentos de Redes de Computadores - 3º Período
Professor: José Maurício S. Pinheiro

AULA 4: Princípios da Internet

A Internet ou *World Wide Web* é uma estrutura arquitetônica que permite o acesso a documentos vinculados, espalhados por milhões de máquinas pelo mundo. É uma rede pública de comunicação de dados, com controle descentralizado e que utiliza o conjunto de protocolos TCP/IP como base para a estrutura de comunicação e seus serviços de rede.

Conhecida ainda como *Web*, teve início em 1989, no CERN (*The European Organization for Nuclear Research*), centro europeu para pesquisa nuclear e nasceu da necessidade de fazer com que grupos de cientistas de diferentes nacionalidades pudessem colaborar uns com os outros através da troca de relatórios, plantas, desenhos, fotos e outros documentos. Em 1994, o CERN e o *Massachusetts Institute of Technology* (MIT) assinaram um acordo criando o *World Wide Web Consortium* (W3C), uma organização voltada para o desenvolvimento da *Web*, a padronização de protocolos e para o incentivo à interoperabilidade entre os sites.

A Internet tem as suas políticas controladas pelo IAB (*Internet Architecture Board*), um fórum patrocinado pela *Internet Society*, uma comunidade aberta formada por usuários, fabricantes, representantes governamentais e pesquisadores.

1. ARQUITETURA

Do ponto de vista dos usuários, a Internet é uma coleção de documentos, conhecidos como páginas *Web*. Cada página pode conter *links* (vínculos) para outras páginas. Os usuários podem seguir um *link* (por exemplo, dando um clique sobre ele), que os levará até a página indicada. A idéia de fazer uma página apontar para outra é chamada hipertexto ou *hiperlink*. As páginas são visualizadas com o auxílio de um programa denominado navegador ou *browser* que busca a página solicitada, interpreta seu texto e seus comandos de formatação e exibe a página, formatada de modo apropriado, na tela do computador. Também é possível usar a Internet para enviar e receber arquivos. Chama-se *download* a recepção de um arquivo através da *Web* e *upload* o envio de arquivos pela *Web*.

O modelo básico de funcionamento da *Web* é mostrado na Figura 1. Na figura, o navegador está exibindo uma página na máquina cliente. Quando o usuário clica em uma linha de texto vinculada a uma página no servidor *abcd.com*, o navegador segue o *hiperlink* enviando uma mensagem ao servidor *abcd.com* na qual solicita a página. Se essa página contiver um *hiperlink* para uma página no servidor *xyz.com* e ele for ativado por um clique, o navegador enviará uma solicitação da página a essa máquina, e assim por diante.

Além de terem texto comum (não sublinhado) e hipertexto (sublinhado), as páginas da *Web* também podem conter ícones, desenhos de linhas, mapas e

fotografias. Cada um deles pode (opcionalmente) estar vinculado a outra página.

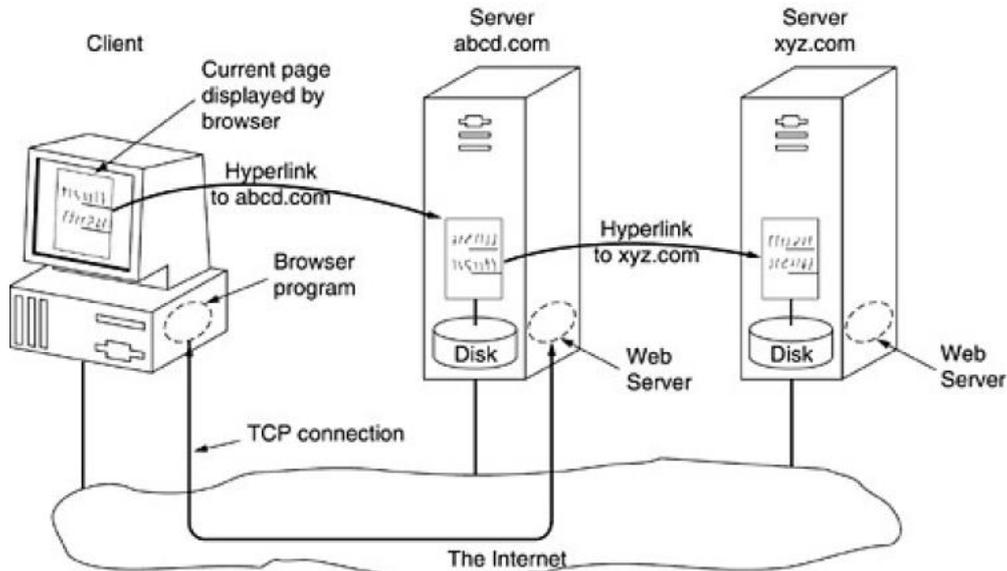


Figura 1 - Modelo Web básico

1.1. O LADO DO CLIENTE

Um navegador é um programa que pode exibir uma página da Internet. Quando um item é selecionado, o navegador segue o *hiperlink* e busca a página selecionada. Por essa razão, o *hiperlink* incorporado precisa de um meio para nomear qualquer outra página na Web. As páginas são nomeadas com o uso de URLs (*Uniform Resource Locators*).

O URL tem três partes: o protocolo (http, https), o nome DNS da máquina em que a página está localizada (www.abcd.com) e o nome do arquivo que contém a página. Quando um usuário clica em um *hiperlink*, o navegador executa uma série de etapas em ordem para buscar a página indicada.

As etapas que ocorrem quando um link é selecionado são as seguintes:

1. O navegador determina o URL
2. O navegador pergunta ao DNS qual é o endereço IP
3. O DNS responde com o IP correspondente (por exemplo, 157.108.192.32)
4. O navegador estabelece uma conexão TCP com a porta 80 em 157.108.192.32.
5. O navegador envia um comando solicitando o arquivo /home/index.html.
6. O servidor envia o arquivo /home/index.html.
7. A conexão TCP é encerrada.
8. O navegador exibe todo o texto de /home/index.html.
9. O navegador busca e exibe todas as imagens que o arquivo contém.

1.1.1. PLUG-INS E APLICAÇÕES AUXILIARES

Em vez de tornar os navegadores cada vez maiores, construindo interpretadores para uma coleção de tipos de arquivos que cresce com rapidez, a maioria dos navegadores opta por uma solução mais geral. Quando um servidor retorna uma página, ele também retorna algumas informações adicionais sobre a página. Há duas possibilidades: *plug-ins* e aplicações auxiliares.

Um *plug-in* é um módulo de código que o navegador busca em um diretório especial no disco e instala como uma extensão do próprio navegador.

A outra maneira de estender um navegador é usar uma aplicação auxiliar, um programa completo que é executado como um processo separado. A aplicação auxiliar não oferece nenhuma interface para o navegador e não faz uso de serviços do navegador. Em vez disso, em geral ele simplesmente aceita o nome de um arquivo de rascunho em que o arquivo de conteúdo é armazenado, abre o arquivo e exibe o conteúdo. De modo geral, os auxiliares são grandes programas que existem de forma independente do navegador, como aplicações para PDF e suítes de escritório.

1.2. O LADO DO SERVIDOR

Quando o usuário digita um URL ou clica em uma linha de hipertexto, o navegador analisa o URL e interpreta a parte entre `http://` e a barra seguinte como um nome DNS a ser pesquisado. Munido do endereço IP do servidor, o navegador estabelece uma conexão TCP para a porta 80 desse servidor. Em seguida, envia um comando contendo o restante do URL, que é o nome de um arquivo nesse servidor. O servidor então retorna o arquivo para ser exibido pelo navegador.

Os servidores da Internet fazem mais que apenas aceitar nomes e retornar arquivos. De fato, o processamento real de cada solicitação pode ficar bastante complicado. Por essa razão, em muitos servidores, cada módulo de processamento executa uma série de etapas usando algum subconjunto das etapas a seguir, dependendo de quais delas sejam necessárias para essa solicitação específica:

1. Resolver o nome da página da Web solicitada.
2. Autenticar o cliente.
3. Executar o controle de acesso no cliente.
4. Executar o controle de acesso na página da Web.
5. Verificar o *cache*.
6. Buscar a página solicitada no disco.
7. Determinar o tipo *plug-in* ou aplicação auxiliar na resposta.
8. Cuidar de diversas tarefas.
9. Retornar a resposta ao cliente.
10. Criar uma entrada no *log* do servidor.

2. URL'S — UNIFORM RESOURCE LOCATORS

Páginas da Internet podem conter *links* para outras páginas. Esse processo exige mecanismos de nomenclatura e localização de páginas, o que implica na resposta de três perguntas antes de uma página selecionada poder ser exibida:

1. Qual é o nome da página?
2. Onde a página está localizada?
3. Como a página pode ser acessada?

Se toda página recebesse de algum modo a atribuição de um nome exclusivo, não haveria nenhuma ambigüidade na identificação de páginas. Apesar disso, o problema não seria resolvido. A solução escolhida identifica as páginas de uma forma que resolve todos os três problemas ao mesmo tempo. A cada página é atribuído um *Uniform Resource Locator* (URL) que funciona como o nome universal da página. Os URL's têm três partes: o protocolo, o nome DNS da máquina em que a página está e um nome local que indica a página específica (normalmente, um nome de arquivo na máquina onde ele reside). Quando o texto é selecionado, o navegador procura o nome do host usando o DNS. Agora, de posse do endereço IP do host, o navegador estabelece uma conexão TCP com o host. Através dessa conexão, ele envia o nome do arquivo, usando o protocolo especificado.

Esse esquema de URL é aberto, pois é simples para os navegadores o uso de vários protocolos para acessar diferentes tipos de recursos. Na verdade, foram definidos URL's para vários outros protocolos comuns como exemplifica a Tabela 1.

Tabela 1 - Exemplos de protocolos comuns usados na Internet

Nome	Usado para	Exemplo
http	Hipertexto (HTML)	http://www.cs.vu.nl/~ast/
ftp	FTP	ftp://ftp.cs.vu.nl/pub/minix/README
telnet	Login remoto	telnet://www.w3.org:80

O protocolo HTTP (*Hypertext Transfer Protocol*) é a linguagem natural da Internet, que os servidores utilizam para se comunicar. O protocolo FTP (*File Transfer Protocol*) é o protocolo de transferência de arquivos da Internet. O protocolo TELNET é usado para estabelecer uma conexão on-line com uma máquina remota.

3. URN'S - UNIVERSAL RESOURCE NAMES

O crescente uso da Web gerou um enfraquecimento no esquema de URL. Um URL aponta para um host específico. No caso de páginas intensamente referenciadas, é interessante ter várias cópias afastadas entre si, a fim de reduzir o tráfego da rede.

O problema é que os URL's não oferecem um meio de fazer referência a uma página sem informar simultaneamente onde ela está. Para resolver esse problema e possibilitar a replicação de páginas, a *Internet Engineering Task*

Force (IETF) trabalha em um sistema de URN's (*Universal Resource Names*). Um URN pode ser entendido como um URL generalizado.

4. COOKIES

O nome deriva de uma antiga gíria dos programadores, na qual um programa chama um procedimento e recebe de volta algo que talvez precise apresentar mais tarde para conseguir a realização de algum trabalho.

Os *cookies* são arquivos ou *strings*, e não programas executáveis. Quando um cliente solicita uma página da Internet, o servidor pode fornecer informações adicionais junto com a página solicitada. Essas informações podem incluir um *cookie* com 4 KB no máximo. Os navegadores armazenam os cookies oferecidos em um diretório no disco rígido do cliente, a menos que o usuário tenha desativado os *cookies*.

Em princípio, um cookie poderia conter um vírus; no entanto, tendo em vista que os cookies são tratados como dados, oficialmente não existe nenhuma maneira do vírus funcionar e causar danos, embora sempre seja possível algum hacker explorar uma falha ou *bug* do navegador para provocar a ativação de um cookie. Um cookie pode conter até cinco campos conforme mostra a Tabela 2:

- **Domain** - informa de onde veio o cookie. Os navegadores devem confirmar que os servidores não estão mentindo a respeito de seu domínio. Cada domínio pode armazenar no máximo 20 cookies por cliente;
- **Path** - é um caminho na estrutura de diretórios do servidor que identifica as partes da árvore de arquivos do servidor que podem usar o cookie. Frequentemente, ele contém o símbolo / (barra), que representa a árvore inteira.
- **Content** - tem a forma nome = valor. Tanto nome quanto valor podem ser o que o servidor quiser. Esse campo é onde fica armazenado o conteúdo do cookie.
- **Expires** - especifica quando o cookie irá expirar. Se esse campo estiver ausente, o navegador descartará o cookie ao terminar (chamado cookie não persistente). Se forem fornecidas hora e data, o cookie é chamado persistente e será mantido até expirar. A hora de expiração (vencimento) é dada pelo horário de Greenwich (GMT). Para remover um cookie do disco rígido do cliente, basta o servidor enviá-lo novamente com uma data de expiração vencida.
- **Secure** - pode ser ajustado para indicar que o navegador só deve retornar o cookie a um servidor seguro. Esse recurso é usado para comércio eletrônico, transações bancárias e outras aplicações seguras.

Tabela 2 - Campos encontrados nos cookies

Domain	Path	Content	Expires	Secure
toms-casino.com	/	CustomerID=497793521	15-10-02 17:00	Yes
joes-store.com	/	Cart=1-00501;1-07031;2-13721	11-10-02 14:22	No

O primeiro cookie definido por *toms-casino.com* e é usado para identificar o visitante. Quando o visitante se conectar da próxima vez, o navegador enviará o cookie para que o servidor saiba quem ele é. Munido dessa identificação, o servidor poderá pesquisar o registro em um banco de dados e usar essa informação para construir uma página apropriada para exibição de acordo com os hábitos do visitante.

O segundo cookie definido por *joes-store.com* é de um site de vendas pela Internet. Ao encontrar um produto do seu interesse, o visitante clica sobre o artigo e o servidor monta um cookie contendo o número de itens e o código dos produtos para enviar de volta ao visitante. À medida que o visitante continua a percorrer a loja, o cookie é transmitido a cada nova página solicitada. Conforme as compras se acumulam, o servidor adiciona as mercadorias ao cookie. Finalmente, quando o visitante finaliza a compra, o cookie, que agora contém a lista completa de compras, é enviado junto com a solicitação. Desse modo, o servidor sabe exatamente o que foi adquirido.

Os cookies também podem ser usados em benefício do próprio servidor. Por exemplo, deseja-se saber quantos visitantes distintos um servidor recebeu e quantas páginas cada um percorreu antes de deixar o site. Quando a primeira solicitação entrar, ela não será acompanhada por nenhum cookie, e assim o servidor enviará de volta um cookie contendo Counter = 1. Cliques subsequentes nesse site farão o cookie ser enviado de volta ao servidor. Toda vez que isso ocorre, o contador é incrementado e devolvido ao cliente. Controlando os contadores, o servidor pode verificar quantas pessoas desistiram depois de verem a primeira página, quantas observaram duas páginas e assim por diante.

Um uso controverso dos cookies tem a finalidade de reunir secretamente informações sobre os hábitos de navegação dos usuários. Na teoria, os cookies só devem voltar ao site que os originou, mas os hackers têm explorado numerosos bugs nos navegadores para capturar cookies não destinados a eles. Tendo em vista que alguns sites de e-commerce inserem números de cartões de crédito em cookies, fica claro o potencial para abusos.

Para manter alguma privacidade, alguns usuários configuram seus navegadores para rejeitar todos os cookies. Porém, isso pode trazer problemas com sites legítimos que utilizam cookies. Para resolver essa questão, às vezes os usuários instalam software para eliminar cookies. Esses produtos de software são programas especiais que inspecionam cada cookie recebido assim que ele chega e aceitam ou descartam o cookie, dependendo das opções fornecidas pelo usuário (por exemplo, sobre os Web sites confiáveis). Isso dá ao usuário um controle preciso sobre os cookies que devem ser aceitos e rejeitados. Os navegadores modernos têm controles internos elaborados sobre cookies, definidos pelo usuário.

5. PROTOCOLOS WEB

A Internet é dita ser um sistema aberto, uma vez que todos os seus serviços básicos assim como as aplicações são definidas publicamente, podendo ser implementadas e utilizadas sem pagamento de royalties ou licenças para outras instituições. Neste contexto, um protocolo permite a comunicação entre processos (que se executam eventualmente em diferentes máquinas), isto é,

um conjunto de regras e procedimentos para transmitir e receber dados numa rede. O Modelo de Referência OSI é útil como ferramenta conceitual para a compreensão das camadas dos protocolos. Os protocolos são implementações reais, no código do programa e no hardware, das normas conceituais definidas no Modelo de Referência OSI. As tarefas descritas no Modelo OSI podem ser executadas de várias maneiras, dependendo dos objetivos dos projetistas da rede. Além disso, vários conjuntos de protocolos são compartilhados (Figura 2). Alguns protocolos e conjuntos de protocolos já existiam antes do Modelo de Referência OSI ter sido divulgado e podem ser vagamente comparados ao modelo de sete camadas. Entre os serviços mais conhecidos da Internet estão o correio eletrônico (protocolos SMTP, POP3), a transferência de arquivos (FTP), a emulação remota de terminal (Telnet) e o acesso à informação hipermídia (HTTP).

5.1. CONJUNTO TCP/IP

O conjunto de protocolos da Internet, também chamado de conjunto de protocolos TCP/IP, foi projetado especialmente para ser utilizado na Internet. Sua característica principal é o suporte direto à comunicação entre redes de diversos tipos. Neste caso, a arquitetura TCP/IP é independente da infraestrutura de rede física ou lógica empregada.

O que torna a comunicação da Internet possível são estes dois protocolos, e eles estão ligados para permitir uma comunicação eficiente. O TCP separa os dados em “pacotes”, e o IP endereça-os e é o responsável por fazer esses dados chegarem ao seu destino, ou seja, estes dois protocolos dividem os dados em seções (pacotes) e entregam-nos no seu devido destino. Depois de serem entregues, reconstroem os pacotes na forma original para eles poderem ser visualizados e utilizados.

O TCP/IP foi desenvolvido originalmente pelo Departamento de Defesa norte-americano (DoD), visando oferecer um serviço confiável nas inter-redes amplas que incorporam vários tipos de computadores.

5.2. INTERNET PROTOCOL (IP)

O *Internet Protocol* (IP) é o protocolo da Internet que corresponde à camada 3 do Modelo OSI. É este protocolo que identifica, localiza e estabelece conexão entre computadores ligados à Internet. Trata-se de um protocolo sem conexão que fornece serviço de datagrama e cujos pacotes são mais comumente conhecidos como datagramas IP. O IP é um protocolo de comutação de pacotes que realiza endereçamento e seleção de rota. Um cabeçalho do IP é incluído nos pacotes, que são transmitidos como frames pelos protocolos de nível inferior.

O IP roteia pacotes pelas interredes, utilizando as tabelas de roteamento dinâmicas que são consultadas a cada salto. O roteamento pode ser determinado por meio da consulta de informações dos dispositivos lógico e físico da rede, como fornecido pelo *Address Resolution Protocol* (ARP). O IP desmonta e remonta o pacote quando requerido pelas limitações de tamanho de pacote definidas para as camadas de link de dados e física que estão sendo implementadas. O IP também realiza a verificação de erros nos dados do

cabeçalho usando um checksum, embora os dados das camadas superiores não suportem essa verificação.

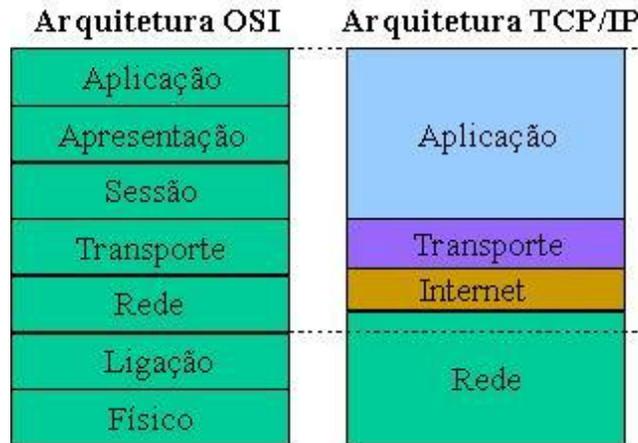


Figura 2 - Comparação entre o Modelo OSI e TCP/IP

5.3. INTRANET

Uma Intranet é a aplicação da tecnologia criada na Internet e do conjunto de protocolos de transporte e de aplicação TCP/IP numa rede privada, interna ou numa empresa. Na intranet, não só a infra-estrutura de comunicação é baseada em TCP/IP, mas também grande quantidade de informações e aplicações são disponibilizadas por meio dos sistemas Web (protocolo HTTP) e correio eletrônico.

5.4. EXTRANET

A Extranet ou *extended intranet* é a extensão dos serviços da intranet de uma empresa para interligar e fornecer aplicações para outras empresas, como clientes, fornecedores, parceiros etc. É a utilização de tecnologias Web e de correio eletrônico para simplificar a comunicação e a troca de informações entre empresas.

6. MÁQUINAS DE BUSCA

Há uma enorme quantidade de informações na Internet e estas informações são, na sua maior parte, renovadas diariamente. Uma máquina (ou motor) de busca ou ainda *Searchbot*, é uma máquina específica encarregada de indexar páginas web para permitir uma busca com a ajuda de palavras-chave, num formulário de busca.

Máquinas de busca são dotadas de *robots* (software), chamados *spiders* (aranhas) que percorrem recursivamente as ligações das páginas Web e indexam seu conteúdo em bases de dados para permitir uma consulta posterior. Nenhuma máquina de busca pode percorrer a totalidade das páginas num só dia (este processo demora geralmente várias semanas), assim cada

máquina adota uma estratégia diferente para a frequência de atualização dos sites.

Quando um usuário preenche o formulário da máquina de busca, especifica as palavras que procura (e eventualmente aquelas que não deseja) graças aos operadores binários “e”, “ou”, “não”... (simbolizados por +, -,...); a solicitação é enviada ao motor de busca que consulta as suas bases de dados para cada uma das palavras e seguidamente refina a busca retirando as páginas que não correspondem à procura. As respostas são classificadas numa ordem de pertinência, correspondendo a critérios próprios da máquina de busca, como a percentagem de palavras que correspondem à busca, o seu índice de densidade (o número de vezes que cada uma das palavras-chave aparece na página) etc.

Convém mencionar ainda a existência de Metamotores de busca. Chama-se “metamotor” a máquina de busca cujo funcionamento se baseia na exploração dos resultados obtidos por outros motores de busca.