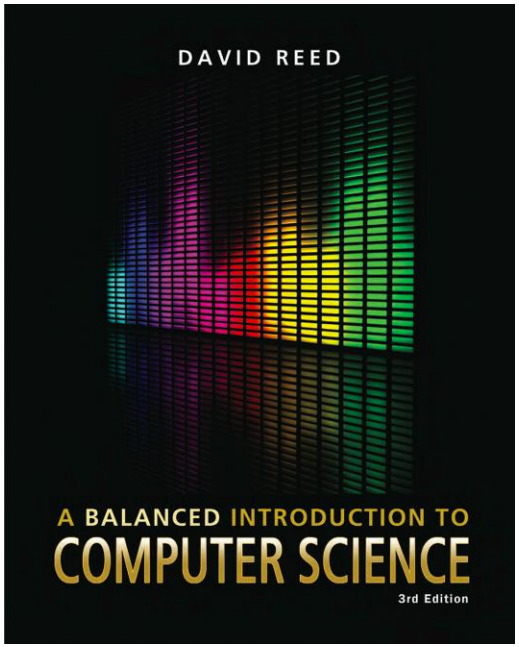
A Balanced Introduction to Computer Science, 3/E

David Reed, Creighton University

ISBN 978-0-13-216675-1



Chapter 3
The Internet and the Web

History of Internet

recall: the Internet is a vast, international network of computers

the Internet traces its roots back to the early 1960s

- MIT professor J.C.R. Licklider published a series of articles describing a "Galactic Network" of communicating computers
- in 1962, Licklider became head of computer research at the U.S. Department of Defense's *Advanced Research Project Agency (ARPA)*
- in 1967, Licklider hired Larry Roberts to design and implement his vision of a Galactic Network

the ARPANet (precursor to the Internet) became a reality in 1969

- it connected computers at four universities: UCLA, UCSB, SRI, and Utah
- it employed dedicated cables, buried underground

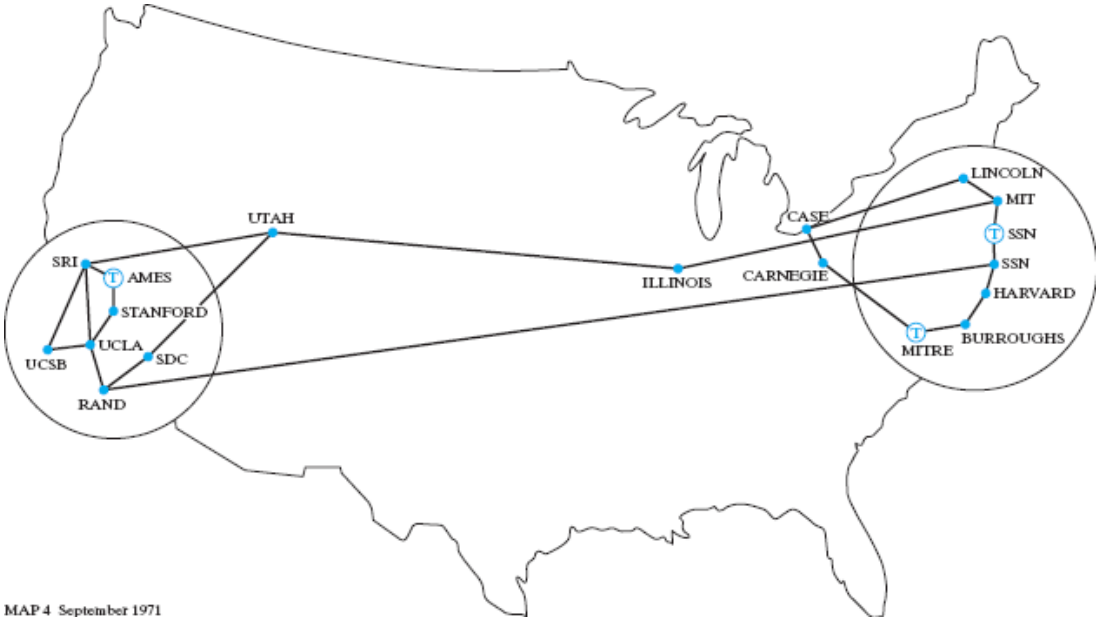
  the data transfer rate was 56K bits/sec, roughly the same as dial-up services today

- the ARPANet demonstrated that researchers at different sites could communicate, share data, and run software remotely

ARPANet

the ARPANet was intended to connect only military installations and universities participating in government projects

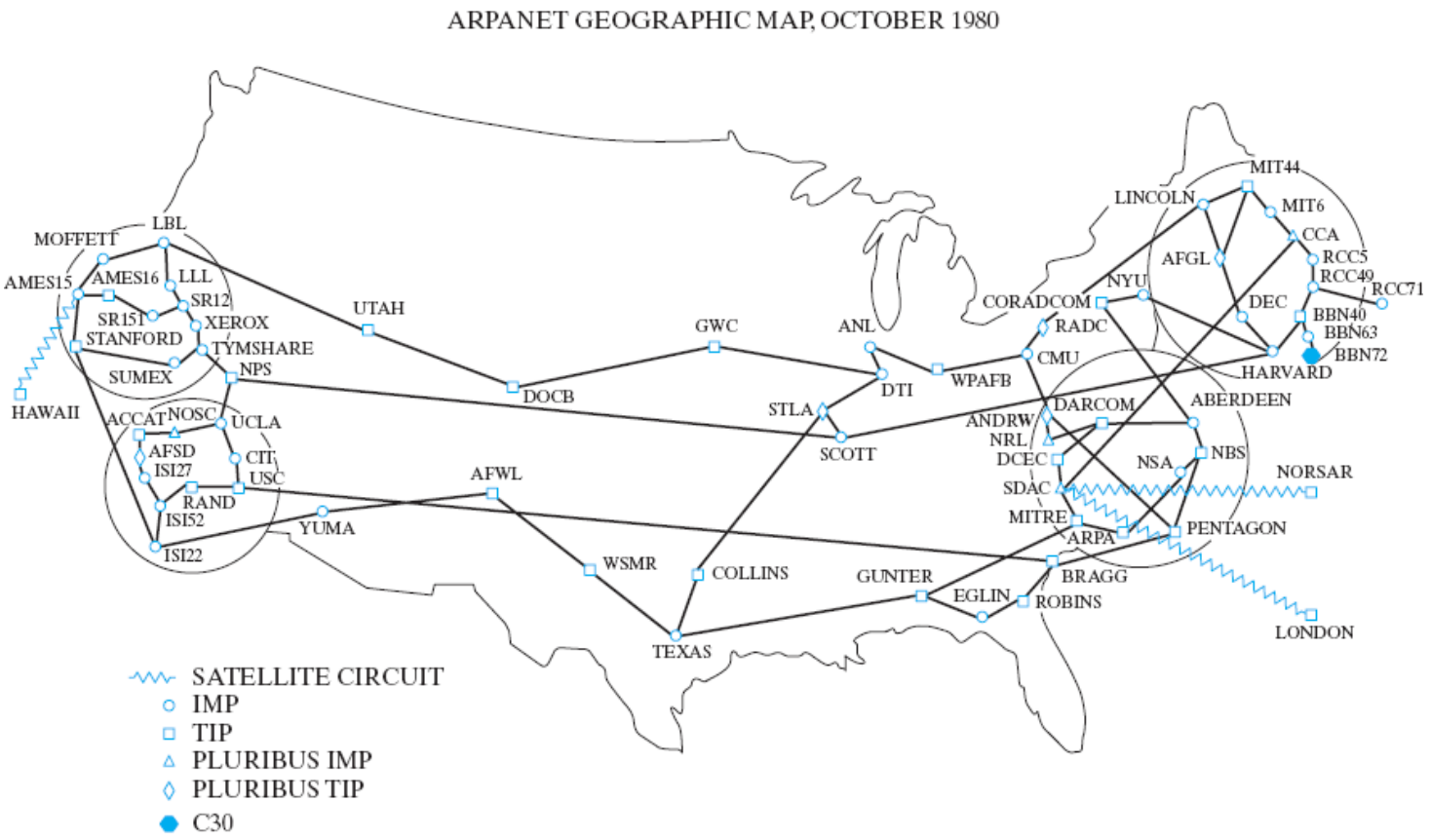
- by 1971, 18 sites were connected; most used Interface Message Processors (IMPs) which allowed up to 4 terminal connections at the site
- sites labeled with a T utilized Terminal Interface Processors (TIPs), which allowed up to 64 terminal connections at the site



ARPANet Growth

by 1980, close to 100 sites were connected to the ARPANet

- satellite connections provided links to select cities outside the continental U.S.



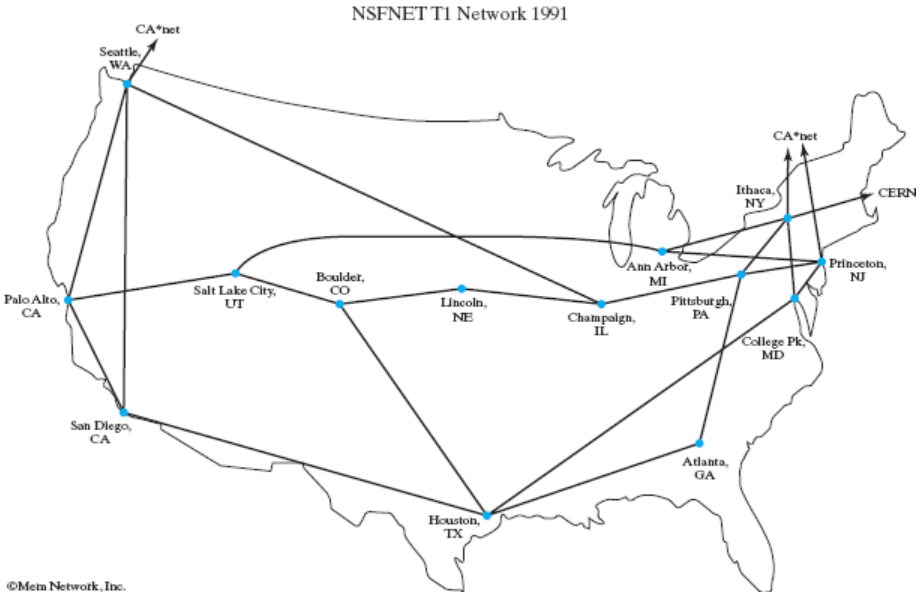
NSFNet

in the early 1980s, the ARPANet experienced an astounding growth spurt

- applications such as email, newsgroups, and remote logins were attractive to all colleges and universities
- by 1984, the ARPANet encompassed more than 1,000 sites

to accommodate further growth, the National Science Foundation (NSF) became involved with the ARPANet in 1984

- NSF funded the construction of high-speed transmission lines that would form the backbone of the expanding network



"Internet"

the term "Internet" was coined in recognition of the similarities between the NSFNet and the interstate highway system

- backbone connections provided fast communications between principal destinations, *analogous to interstate highways*
- connected to the backbone were slower transmission lines that linked secondary destinations, *analogous to state highways*
- local connections were required to reach individual computers, *analogous to city and neighborhood roads*

note: Al Gore did not INVENT the Internet, *nor did he ever claim to*

- *he sponsored legislation in the late 1980s to support growth and expand access*

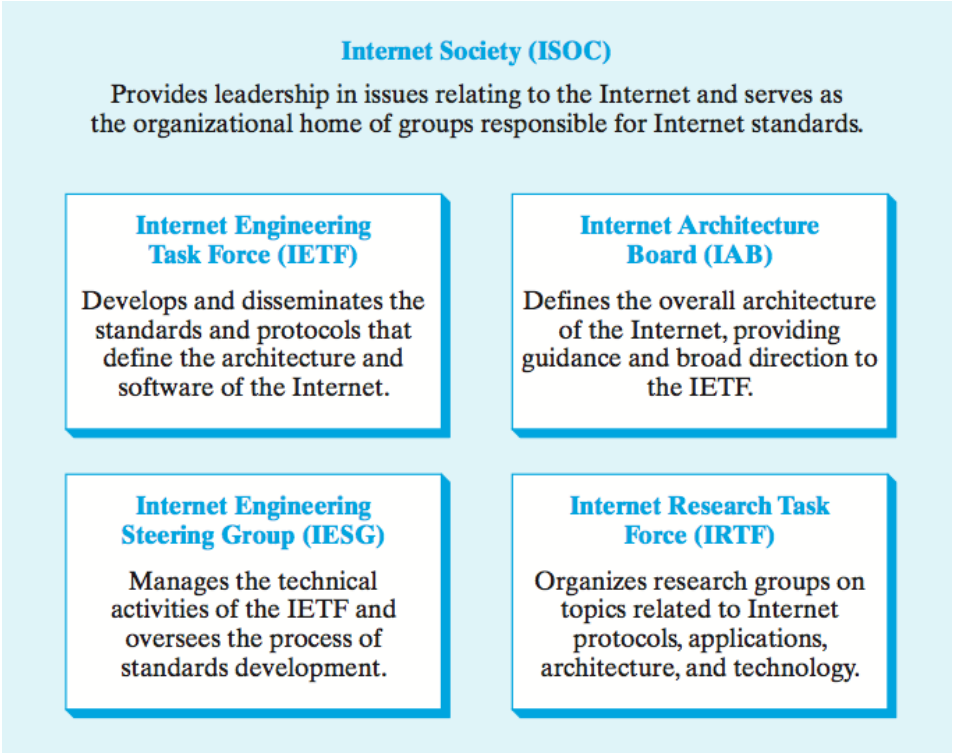
recognizing that continued growth would require significant funding and research, the government decided in the mid 90s to privatize the Internet

- control of the network's hardware was turned over to telecommunications companies and research organizations (e.g., AT&T, Verizon, Qwest, Sprint)
- research and design are administered by the *Internet Society*

Internet Society

Internet Society is an international nonprofit organization (founded in 1992)

- it maintains and enforces standards, ensuring that all computers on the Internet are able to communicate with each other
- it also organizes committees that propose and approve new Internet-related technologies and software



Internet Growth

until recently, the Internet more than doubled in size every 1 or 2 years

- why has this trend slowed? will it continue?

Year	Computers on the Internet[1]
2010	758,081,484
2008	570,937,778
2006	439,286,364
2004	285,139,107
2002	162,128,493
2000	93,047,785
1998	36,739,000
1996	12,881,000
1994	3,212,000
1992	992,000
1990	313,000
1988	56,000
1986	5,089
1984	1,024
1982	235

(Internet Software Consortium, April 2010.)

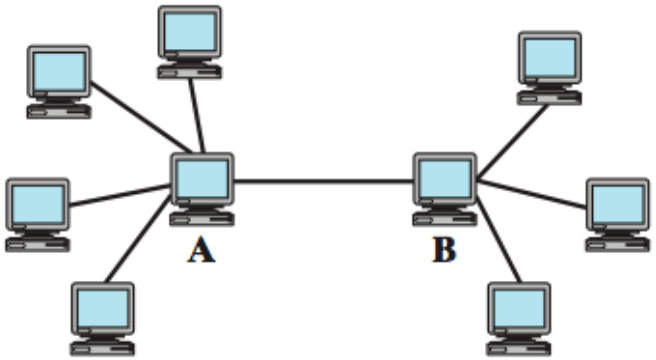
Distributed Networks

the design of the ARPANet was influenced by the ideas of Paul Baran, a researcher at the RAND Institute

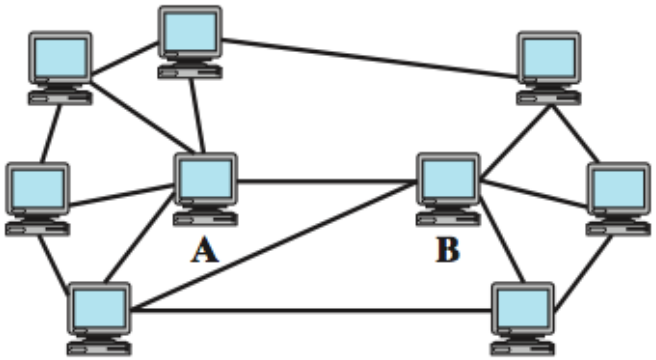
- Baran proposed 2 key ideas: *distributed network* and *packet-switching*

recall: the ARPANet was funded by the Dept of Defense for communications

- as such, it needed to be resistant to attack or mechanical failure



In a centralized network, the failure of a single machine or connection (for example, between A and B), can isolate large portions of the network.

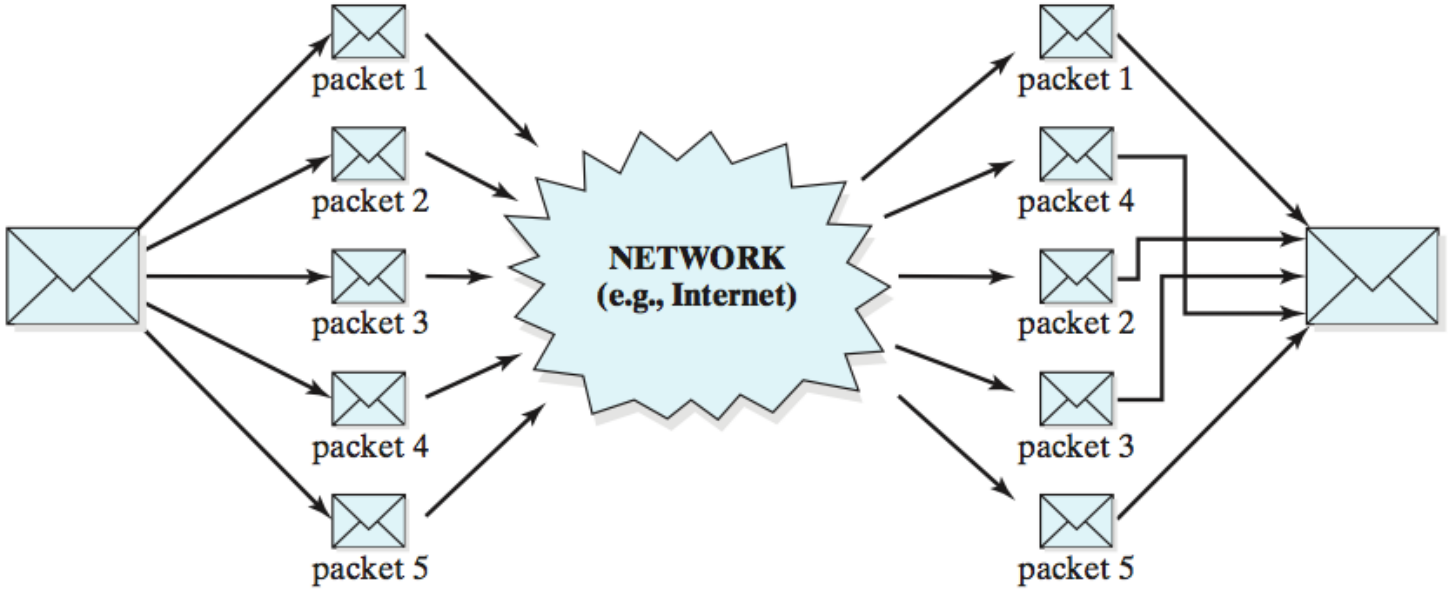


In a distributed network, redundant connections provide alternate routes for messages, allowing them to circumvent failed machines or connections.

Packet Switching

in a packet-switching network, messages to be sent over the network are first broken into small pieces known as *packets*

- these packets are sent independently to their final destination



1. The message is broken into small packets, each labeled for delivery.

2. The packets travel independently across the network, perhaps arriving at their destination out of order.

3. The packets are reassembled in the correct order to obtain the original message.

Advantages of Packets

1. sending information in smaller units increases the efficient use of connections
   - large messages can't monopolize the connection
   - *analogy: limiting call lengths at a pay phone to limit waiting*

2. transmitting packets independently allows the network to react to failures or network congestion
   - routers (special-purpose computers that direct the flow of messages) can recognize failures or congestion and reroute the packet around trouble areas

3. breaking the message into packets can improve reliability
   - since the packets are transmitted independently, it is likely that at least part of the message will arrive (even if some failures occur within the network)
   - software at the destination can recognize which packets are missing and request retransmission

Protocols and Addresses

the Internet allows different types of computers from around the world to communicate

- this is possible because the computing community agreed upon common *protocols* (sets of rules that describe how communication takes place)
- the two central protocols that control Internet communication are:
  1. *Transmission Control Protocol* (*TCP*)
  2. *Internet Protocol* (*IP*)

these protocols rely on each computer having a unique identifier (known as an *IP address*)

- *analogy: street address + zip code provide unique address for your house/dorm*
  using this address, anyone in the world can send you a letter

- an IP address is a number, written as a dotted sequence such as 147.134.2.84
- each computer is assigned an IP address by its Internet Service Provider (ISP)
- some ISPs (e.g., AOL, most colleges) maintain a pool of IP addresses and assign them dynamically to computers each time they connect

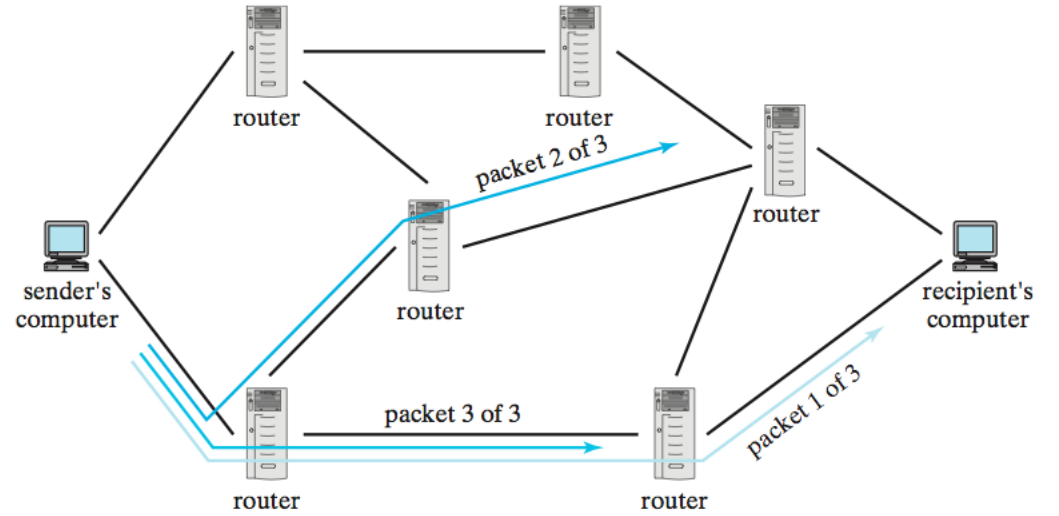
TCP/IP

Transmission Control Protocol (TCP)

- controls the method by which messages are broken down into packets and then reassembled when they reach their final destination

Internet Protocol (IP)

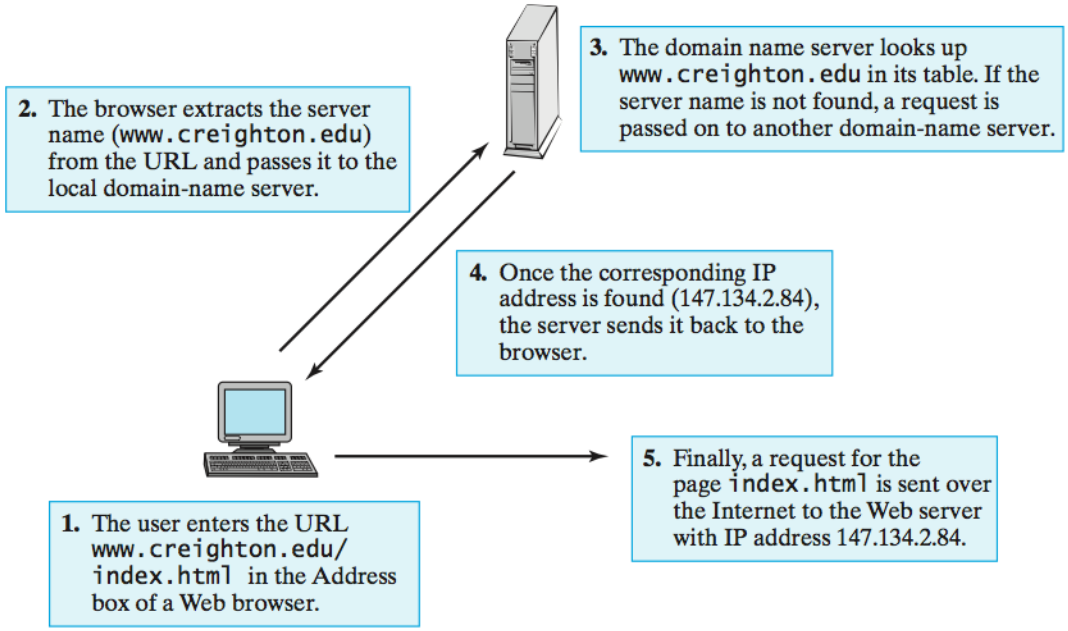
- concerned with labeling the packets for delivery and controlling the packets' paths from sender to recipient



Routers and DNS

the Internet relies on special purpose computers in the network

- *routers* are computers that receive packets, access the routing information, and pass the packets on toward their destination
- *domain name servers* are computers that store mappings between domain names and IP addresses
  - *domain names* are hierarchical names for computers (e.g., bluejay.creighton.edu)
    they are much easier to remember and type than IP addresses
  - domain name servers translate the names into their corresponding IP addresses



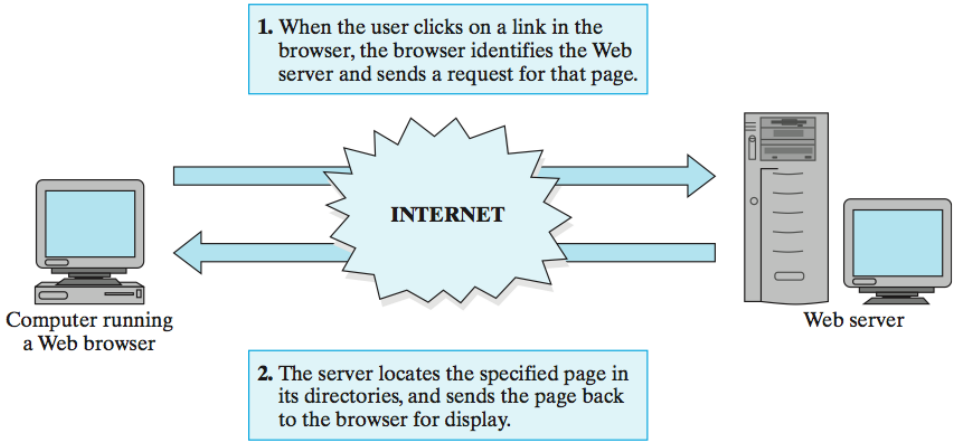
History of the Web

the World Wide Web is a multimedia environment in which documents can be seamlessly linked over the Internet

- proposed by Tim Berners-Lee at the European Laboratory for Particle Physics (CERN) in 1989
- designed to facilitate sharing information among researchers located all over Europe and using different types of computers and software

Berners-Lee's design of the Web integrated two key ideas

1. hypertext (documents with interlinked text and media)
   - Web pages can contain images and links to other pages
2. the distributed nature of the Internet
   - pages can be stored on machines all across the Internet, known as *Web servers*
   - logical connections between pages are independent of physical locations



Web Timeline

1990: Berners-Lee produced working prototypes of a Web server and browser

1991: Berners-Lee made his software available for free over the Internet

1993: Marc Andreesen and Eric Bina of the University of Illinois' National Center for Supercomputing Association (NCSA), wrote the first graphical browser: Mosaic
- Mosaic integrated text, image & links, made browsing more intuitive

1994: Andreesen founded Netscape, which marketed the Netscape Navigator

1995: Microsoft released Internet Explorer → the browser wars begin!

1999: Internet Explorer becomes the most popular browser (~90% of market in 2002)

2009: Mozilla Firefox (Netscape descendent) grows in popularity, IE share drops to 62%

Year	Computers on the Internet[5]	Web Servers on on the Internet[6]
2010	758,081,484	205,368,103
2008	570,937,778	175,480,931
2006	439,286,364	88,166,395
2004	285,139,107	52,131,889
2002	162,128,493	33,082,657
2000	93,047,785	18,169,498
1998	36,739,000	4,279,000
1996	12,881,000	300,000
1994	3,212,000	3,000
1992	992,000	50

in 2005, Google indexed more than 8 billion Web pages

by 2009, various sources have estimated >50 billion Web pages

(Internet Software Consortium & Netcraft, April 2010.)

How the Web Works

like Internet communications, the Web relies on protocols to ensure that pages are accessible to any computer

- HyperText Markup Language (HTML) defines the form of Web page content
  - HTML5 is the current draft standard, supported by all modern browsers
- HyperText Transfer Protocol (HTTP) defines how messages exchanged between browsers and servers are formatted
  - the prefix `http://` in a URL specifies that the HTTP protocol is to be used in communicating with the server
  - the prefix is NOT used for local file access since no server communication is necessary

for efficiency reasons, browsers will sometimes *cache* pages/images

- to avoid redundant downloads, the browser will store a copy of a page/image on the hard drive (along with a time stamp)
- the next time the page/image is requested, it will first check the cache
  - if a copy is found, it sends a *conditional* request to the server
    - essentially: "send this page/image only if it has been changed since the timestamp"
    - if the server copy has not changed, the server sends back a brief message and the browser simply uses the cached copy