CSE 454
Advanced Internet & Web Services

- **Prof: Dan Weld**
  - Most lectures, concepts, perspective.
- **TA: Alan Liu**
  - Machine/environment/software, project details
- **Expectations:**
  - Project (multiple parts, on time!)
  - Reading (papers, web - no formal text)
  - Class participation / development
- **Caveat: Life on the cutting edge**

My Background

- **Research on Intelligent Internet Systems [1991-**
  - Internet Softbot (Discover award finalist '95)
  - Webcrawler by Brian Pinkerton
  - Metacrawler by Eric Selberg & Oren Etzioni
  - Mulder (first automated WWW question answerer)
  - KnowItAll – massive, autonomous information extraction
- **Co-founded**
  - Netbot
  - AdRelevance
  - Nimble Technology
  - Asta Networks
- **Leaves of absence**
  - VP Engineering at Netbot
  - Venture Partner w/ Madrona Venture Group.
- **Incredible shortage of software engineers!**
- **Dearth of training**

Your Background?

- **Classes?**
  - 444, 451, 461, 473
- **Concepts?**
  - Threads, race condition, deadlock
  - Naïve Bayes classifier
  - Hybrid hash join algorithm
  - Precision, recall
  - Fingerprint algorithm
  - LRU cache replacement policy
- **Programming Background?**
  - Java, .NET, J2EE, XML, admin own webserver

Course Outcomes

- **After this course, you should know:**
  - How search engines work
  - How to build scalable web sites
  - How Amazon generates personalized recommendations
  - How digital cash works
  - Issues in e-commerce
  - How to build peer2peer systems (overlay networks)
- **Focus: search! (why?)**

Why Search?

- **A billion or so searches per day…**
- **Boost to productivity**
  - Intellectual & economic
- **Search is 'hot'**
  - Google IPO
  - Amazon's book search feature
- **Fascinating research problem.**
- **You can learn to be a something of a search expert in one quarter!**

Syllabus

- **Introduction**
  - History, networking overview, web server architecture
- **Information Retrieval on the Web**
  - Crawling, indexing, scaleup issues
  - Vector space model,
  - Hyperlink analysis
- **Data Mining**
  - collaborative filtering, clustering, classification
- **Web Services**
  - Protocols, brokers, meta-search, data integration
- **Information Extraction**
  - Question answering
  - The future of search
- **Special Topics**
  - Semantic web, e-commerce, security, peer-to-peer, Time permitting

What This Course Is Not

> … there is a difference between training and education. If computer science is a fundamental discipline, then university education in this field should emphasize enduring fundamental principles rather than transient current technology.
> -Peter Wegner, Three Computing Cultures. 1970.

- **We won't:**
  - Teach you how to be a web master
  - Teach all the latest x-buzzwords in technology
    - XML/SOAP/WSDL
      - (okay, may be a little).
  - Teach web/javascript/java/jdbc… programming

Grading

- **Group Project**
  - 50% The artifact itself
  - 25% Written report
  - 25% Oral presentation and class participation
- **Note: 454 is a capstone design class**

Project: Webcam Search

Why?

- **Finding webcams**
- **Classifying them**
- **Search interface**

Good news: we'll rely on **Nutch** rather than building an engine from scratch.
Team Project (groups of 3)

Warning

- **No textbook**
- **Large project component**
- **Poorly documented, unstable systems**
- **Field changes quickly**
  - Each year is essentially a new course
- **Need students to help debug class!**

History

Pre-history: Census, Dewey Decimal system
and other bizarre medieval rituals performed by hand.
1950s: "Information Retrieval" (IR) term coined
1960 Ted Nelson proposes Xanadu
Hypertext vision of WWW
1961 Kleinrock paper on packet switching
Contrast with phone lines, which are circuit switched.
1965 Gordon Moore proposes law
1966 Design of ARPAnet

History

1968 Doug Engelbart: the first WIMP
Gerald Salton SMART system (Cornell)
vector space model, "father of IR"
1969 First ARPAnet message UCLA -> SRI
1970 ARPAnet spans country, has 5 nodes
1971 ARPAnet has 15 nodes
1972 First email programs, FTP spec
1973 Ethernet operation at Xerox PARC

History

1974 Intel launches 8080;
TCP design
1975 Gates/Allen write Basic for Altair 8800
1976 Apple Computer formed by Jobs/Wozniak
1977 111 hosts on ARPAnet
1979 Visicalc
1980s: Proprietary document DBs
Lexis-Nexis, Medline
1981 Microsoft has 40 employees; IBM PC

History

1983 ARPAnet uses TCP/IP
Birth of internet
1983 Design of DNS
1984 Launch of Macintosh;
1000 hosts on ARPAnet
1985 Symbolic.com first registered domain name
1989 100,000 hosts on Internet
1990 Cisco Systems goes public $288 M
Tim Berners-Lee creates WWW at CERN

History

1993 Mosaic developed at UIUC
Web grows by 341,000% in a year
1994 Webcrawler built (UW class project!)
Yahoo launched, Netscape & Amazon formed
1995 Netscape IPO, Windows 95, MetaCrawler
1997 Amazon IPO
2000 Internet "bubble" bursts.
2001

History

1990: Archie (index file names, anon. ftp servers)
1991: Gopher (menus, links, to servers)
1992: Veronica (index of menu items on gophers)
1993: Jughead (keyword + boolean search)
Mosaic developed at UIUC
Web grows by 341,000% in a year
1993: WWW Wonderer (first crawler)
1994: WebCrawler (UW class project!), Lycos (first popular SEs)
1994: Yahoo directory
1995: MetaCrawler (first major meta-search engine)
Netscape IPO

Approaching the Present

1997: goto.com ("sponsored links" pay-per-click)
AskJeeves (question answering)
Netbot (comparison-shopping search)
Amazon IPO
1998: Open directory launched
1998: Google, pagerank algorithm
1999: SE becomes portal (Yahoo, Excite)
"Search is a commodity"
2000: Flipdog (information extraction)
2001-?: Ascendance of Google
"search is nirvana"
Dominance of advertising model

The Future?

Multi-media IR
images.google.com
Comparison shopping
mysimon.com, froogle.google.com)
Open-source search
Nutch.
Desktop Search
Relevance spamming

Networking Overview

- **Network – collection of nodes and links that cooperate for communication**
- **Nodes – computer systems**
  - Internal (routers, bridges, switches)
  - Terminal (workstations)

- **Links – connections for transmitting data**
- **Protocol – standards for formatting and interpreting data and control information**

Adapted from slides by Neil Spring

Getting Data Across (imperfect wires)

- **Split big files into small pieces (packets)**
- **Packets (~1500 bytes) are sent separately**
  - Can be corrupted (noise, bugs)
  - Can be dropped (if corrupted, overloaded)
  - Can be reordered (if retransmitted, different paths)
- **Allows packets from different flows to be multiplexed along the same link**

Layers

- **Each layer abstracts the services of various lower layers.**

OSI Reference	Reality	Packet Format
Application	HTTP	
Presentation		
Session		
Transport	TCP	App data
Network	IP	IP Payload
Data-Link	Ethernet	
Physical	Twisted Pair	Ethernet Payload

The Internet Protocol (IP)

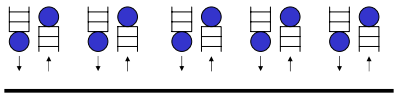
- **Connects disparate networks**
  - Single (hierarchical) address space
  - Single network header
- **Assumes data link is unreliable,**
- **Provides unreliable service**
  - Loss: A B D E
  - Duplication: A B B C D E
  - Corruption: A Q C D E
  - Reordering: A C D B E

IP Addresses

- **32 bits long, split into 4 octets:**
  - For example, 128.95.2.24
- **Hierarchical:**
  - First bits describe which network
  - Last bits describe which host on the network
- **UW subnets include:**
  - 128.95/16, 140.142/16 …
- **UW CSE subnets include:**
  - 128.95.2/24, 128.95.4/24, 128.95.219/24…

Packet Forwarding

- **Buffer incoming packets**
- **Decide which output link**
- **Buffer outgoing packets**
- **Send packet**



Routing

- **How do nodes determine which output link to use to reach a destination?**
- **Distributed algorithm for converging on shortest path tree**
- **Nodes exchange reachability information:**
  - "I can get to 128.95.2/24 in 3 hops"

TCP Service Model

- **Provide Reliability & Ordering**
  - Built on top of the unreliable, unordered IP
- **Bytestream Oriented**
  - When using TCP
  - You can think about bytes, not about packets.

TCP Ports

- **Connections are identified by the tuple:**
  - IP source address
  - IP destination address
  - IP source port
  - IP destination port
- **Lets two machines talk with**
  - Multiple connections at same time
  - Multiple application protocols
- **Well known ports for some applications**
  - Web: 80
  - Telnet: 23
  - Mail:25
  - DNS: 53

Domain Name System

- **We like to use names to refer to computers:**
  - www.cs.washington.edu…
- **But the network uses 4-tuple addresses!**
- **Simple solution: /etc/hosts**
  - Text file lists names and addresses
- **Scalable solution: DNS**
  - Distributed database of name to address mappings

DNS Name hierarchy

No accident DNS names are hierarchical

Allows distributed administration

CS dept administers cs.washington.edu zone

(Just like it administers 128.95.2/24)

Root servers know about

Servers for .edu, .com, .au, .uk, …

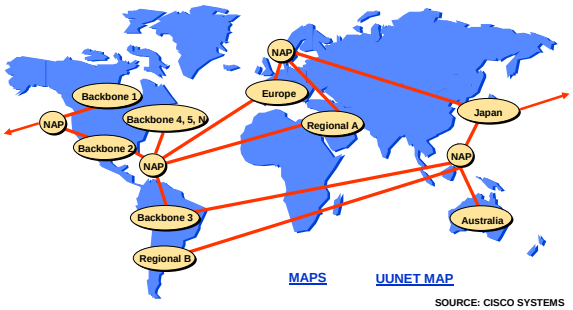
.edu servers know about

ucsd.edu, mit.edu, washington.edu…

Internet Backbone Structure

- **Level 1 (interconnect level, NAPs)**
  - billions of pages per day
- **Level 2 (national backbone, MAE, FIX)**
  - Federal Internet eXchange Points
  - Peering agreements: connect, share routing info)
- **Level 3 (regional providers, state level)**
- **Level 4 (local ISP)**
- **Level 5 (companies, individuals)**
- **Level 6 (routers)**

Structure of the Internet



MAPS UUNET MAP

SOURCE: CISCO SYSTEMS