

## Web Performance Acceleration

Kai Shen

11/10/2014

CSC 257/457 - Fall 2014

1

## Web Performance

- Web performance problems with popular content at a central site:
  - Limited bandwidth at the central site;
  - Many users are far away from the central site.

11/10/2014

CSC 257/457 - Fall 2014

2

## Means of Web Acceleration

- By content providers, e.g., [www.yahoo.com](http://www.yahoo.com)
  - Replicated, distributed Internet sites
- By content consumers (or clients)
  - Web caching
  - Web prefetching
- By a third-party
  - Content distribution network

11/10/2014

CSC 257/457 - Fall 2014

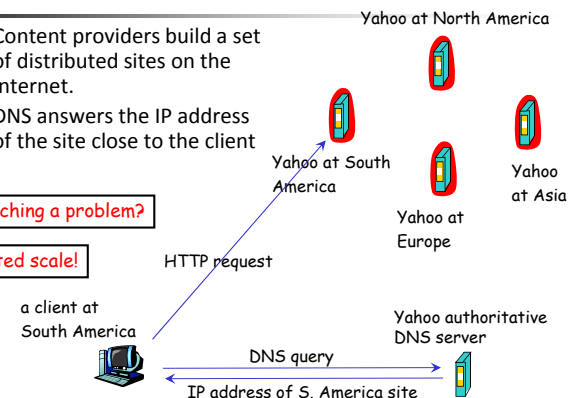
3

## Replicated Internet Sites

- Content providers build a set of distributed sites on the Internet.
- DNS answers the IP address of the site close to the client

Is DNS caching a problem?

limited scale!



11/10/2014

CSC 257/457 - Fall 2014

4

## Determining Nearby Content Server

- We create a “map”, indicating distances between Internet machines (or networks) and content servers
- When a DNS query arrives at authoritative DNS server:
  - server determines network location from which query originates
  - uses “map” to determine a nearby content server
- How to build the map?
  - offline pings from content servers to all network locations
  - just-in-time pings from content servers to the source
    - What if the source does not respond to pings?

11/10/2014

CSC 257/457 - Fall 2014

5

## A Related Question

- Scalable estimation of node-to-node relative proximity
  - A specific problem: given any node X, find a few nearby nodes in a group of 10,000 nodes ( $N_1, \dots, N_{10000}$ ).
  - Too costly to measure the distance from X to 10,000 nodes.
- Use a landmark L:
  - Measure the distances from  $N_1, \dots, N_{10000}$  to L offline.
  - Given node X, we measure the distance from X to L.
    - For node N, if distance(X,L) differs significantly from distance(N,L), X and N are likely far apart.
    - If distance(X,L) is similar to distance(N,L), does it mean X and N are nearby?
- Use multiple landmarks

11/10/2014

CSC 257/457 - Fall 2014

6

## Another Related Question

- How to build a map indicating the geographical location of machines or local networks?
  - User registrations
  - GPS information

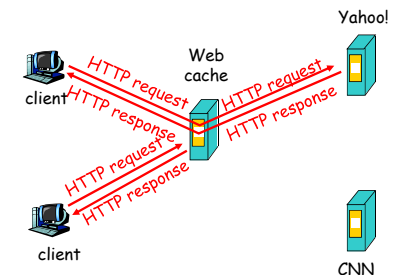
11/10/2014

CSC 257/457 - Fall 2014

7

## Web caches (proxy server)

- Cache is installed and shared by users (university, company, residential ISP)
- **Goal:** satisfy client requests without involving the original server.
- User sets browser: accesses via Web cache
- Browser sends all HTTP requests to Web cache
  - object in cache: cache returns object
  - otherwise cache requests object from the original server, then returns to client



11/10/2014

CSC 257/457 - Fall 2014

8

## A Quantitative Study: the Base Case

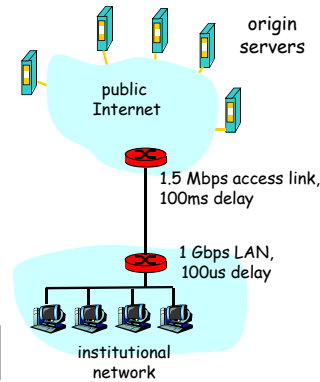
### Assumptions

- 1.5Mbps wide-area access link, 1Gbps local-area network
- wide-area delay 100 milliseconds, local-area delay 100 microseconds

### Performance

- bandwidth to wide-area content – 1.5Mbps
- average access delay  
= Internet delay + LAN delay  
= 100.1 milliseconds

Upgrade the access link BW is costly, doesn't help the access delay.



11/10/2014

CSC 257/457 - Fall 2014

9

## A Quantitative Study: Web Caching

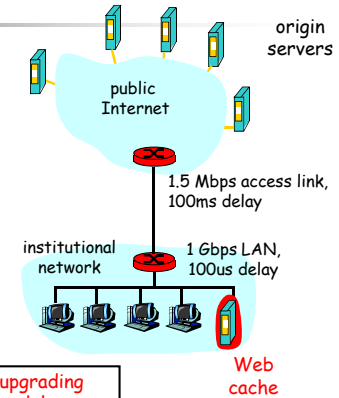
### Install a Web cache

- suppose hit rate is 50%

### Performance

- effective bandwidth to wide-area content  
= 3.0Mbps
- average access delay  
= cache miss delay \* 50% + cache hit delay \* 50%  
= 100.2 \* 50% + 0.1 \* 50%  
= 50.15 milliseconds

Not only higher bandwidth without upgrading access link, but also shorter access delay.



11/10/2014

CSC 257/457 - Fall 2014

10

## Cooperating Web Caching

- Caching cooperation
  - Several caches (often nearby) cooperate with each other
  - Benefits: more cache hits; less load on original servers
- Probing all caches at each access is too costly
- Challenge: knowing (roughly) each other's cache content
  - Each cache periodically broadcast its content to other caches

11/10/2014

CSC 257/457 - Fall 2014

11

## Cache Content Staleness

- Content providers lose direct control of cache content
  - retain some control through cached content staleness
- Is the cached page up-to-date?
  - using If-modified-since HTTP header.
  - removing pages that are too old.

11/10/2014

CSC 257/457 - Fall 2014

12

## Mining of Cache Logs

- Web cache logs contain a wealth of information
  - List of who accessed what at when
- Mine to learn useful statistics
  - Most popular web objects, distribution of web object popularity
  - General user access models (think time between accesses, pattern of page browsing sequence, ...)
  - ...
- User privacy

11/10/2014

CSC 257/457 - Fall 2014

13

## Web Prefetching

- Prefetch a web page before client makes access
  - Reduce latency, but not bandwidth
- Prefetching heuristics?
  - Hyperlinks in the current page (assume client may click some of them)
  - Predict future accesses based on past browsing history
- Benefit vs. cost

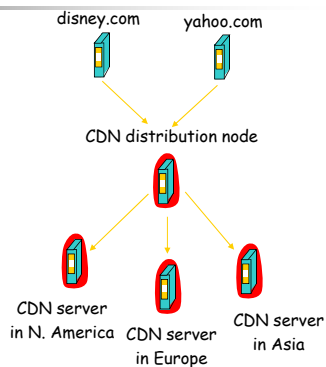
11/10/2014

CSC 257/457 - Fall 2014

14

## Content Distribution Network

- Motivation:
  - limited scale for replication by content providers
  - lose control of content by caching
- Content distribution network:
  - hundreds of CDN servers throughout Internet
  - content providers' content replicated on CDN servers
- Goals:
  - transparent to clients
  - content providers retain primary control

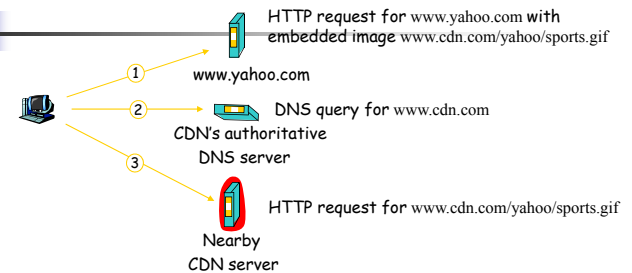


11/10/2014

CSC 257/457 - Fall 2014

15

## CDN in Action



### Content provider

- www.yahoo.com
- Replaces www.yahoo.com/sports.gif with www.cdn.com/yahoo/sports.gif

### CDN company

- replicates content at CDN servers
- uses its authoritative DNS server to redirect requests to a nearby CDN server

11/10/2014

CSC 257/457 - Fall 2014

16



## Summary

---

- Several means:
  - By content providers - replicated, distributed Internet sites
  - By content consumers (or clients) - Web caching, prefetching
  - By a third-party - content distribution network
- Differentiate them on the following:
  - Scalability
  - Content staleness
  - Transparency to clients
  - Reliance on the Domain Name System