Improving Web Performance

These slides are based on the slides made available by the authors of
Computer Networking: A Top Down Approach
Featuring the Internet, 2nd edition.
Jim Kurose, Keith Ross
Addison-Wesley, July 2002.

- HTTP
  - Persistent connections (discussed earlier)
- Web Caching
- Content Delivery Networks
Web caches (proxy server)

**Goal:** satisfy client request without involving origin server

- user sets browser: Web accesses via cache
- browser sends all HTTP requests to cache
  - object in cache: cache returns object
  - else cache requests object from origin server, then returns object to client

More about Web caching

- Cache acts as both client and server
- Cache can do up-to-date check using If-modified-since HTTP header
  - Issue: should cache take risk and deliver cached object without checking?
  - Heuristics are used.
- Typically cache is installed by ISP (university, company, residential ISP)

**Why Web caching?**

- Reduce response time for client request.
- Reduce traffic on an institution's access link.
- Internet dense with caches enables "poor" content providers to effectively deliver content
Caching example (1)

**Assumptions**
- average object size = 100,000 bits
- avg. request rate from institution's browser to origin serves = 15/sec
- delay from institutional router to any origin server and back to router = 2 sec

**Consequences**
- utilization on LAN = 15%
- utilization on access link = 100%
- total delay = Internet delay + access delay + LAN delay
  = 2 sec + minutes + milliseconds

Caching example (2)

**Possible solution**
- increase bandwidth of access link to, say, 10 Mbps

**Consequences**
- utilization on LAN = 15%
- utilization on access link = 15%
- Total delay = Internet delay + access delay + LAN delay
  = 2 sec + msecs + msecs
- often a costly upgrade
Caching example (3)

Install cache
- suppose hit rate is .4

Consequence
- 40% requests will be satisfied almost immediately
- 60% requests satisfied by origin server
- utilization of access link reduced to 60%, resulting in negligible delays (say 10 msec)
- total delay = Internet delay + access delay + LAN delay
  = .6*2 sec + .6*.01 secs + milliseconds < 1.3 secs

Content distribution networks (CDNs)

- The content providers are the CDN customers.

Content replication
- CDN company installs hundreds of CDN servers throughout Internet
  - in lower-tier ISPs, close to users
- CDN replicates its customers’ content in CDN servers. When provider updates content, CDN updates servers
**CDN example**

1. Origin server
2. DNS query for www.cdn.com
   - Origin server
   - CDNs authoritative DNS server
   - Nearby CDN server

**origin server**
- www.foo.com
- distributes HTML
- Replaces:
  - http://www.foo.com/sports.ruth.gif
  - with

**CDN company**
- cdn.com
- distributes gif files
- uses its authoritative DNS server to route redirect requests

**More about CDNs**

**routing requests**
- CDN creates a "map", indicating distances from leaf ISPs and CDN nodes
- when query arrives at authoritative DNS server:
  - server determines ISP from which query originates
  - uses "map" to determine best CDN server

**not just Web pages**
- streaming stored audio/video
- streaming real-time audio/video
- CDN nodes create application-layer overlay network