Means of Content Distribution

- By content providers, e.g. www.yahoo.com
  - replicated, distributed Internet sites
- By content consumers (or clients), e.g., UR
  - reduce bandwidth consumption of an organization through caching - Web caching
- By a third-party
  - content distribution network

A Quantitative Study: the Base Case

Assumptions
- 1.5Mbps wide-area access link, 10Mbps local-area network
- wide-area delay 2 seconds, local-area delay 10 milliseconds

Performance
- effective bandwidth to wide-area content - 1.5Mbps
- average access delay = Internet delay + LAN delay = 2.01 seconds

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%
  = 2.02 * 50% + 0.01 * 50%
  = 1.015 seconds

Dynamic Content Distribution

Christopher Stewart

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

- What is dynamic content?
- How do we create these sites?
- What are the tradeoffs to distributing dynamic content?
- Our first approach to distributing dynamic content.
- Future work

Dynamic Web Content - Definition

Dynamic Web Content – Client-side Implementation

- Developers write code that uses data stored on the user’s hard drive.
  - Cookies, Temporary files
- Users Browser interprets the code and presents dynamic data
- Examples: JavaScript and VB script

Client-side Implementation Problems

- Minor
  - Server has no control over data
  - Browsers interpret scripts differently
- Major
  - File system specific – can’t switch machines
  - User specific – can’t combine data from different sources
Dynamic Web Content – Server-side Implementation

Sees dynamic content as the server sends her updated pages.
- User sends request information to server.
- Web server accepts request. Forwards information as parameter to a processing program.
- The processing program examines data. Communicates with database. Sends response to web server.
- Web server formats the program’s data, and returns HTML to the user.

Web server writes web pages based on content in a local database.

Server-side Implementation Problems

Minor
- Increased resource consumption on the server.
- Centralized data single point of failure.

Major
- Increase in Response time.

Standard Dynamic Content Server Setup

Request Parsers

I need Index.html
Here is index.html

Bad Form
Resubmit

Log In Form
You are a member

Database

Client-Side vs. Server-Side

Which is better?
- If the site requires no or user-specific data?
- If the site requires a lot of data merged from different places?
- A Search Engine?
- Yahoo!? 
Pertinent Issues – review


Throughput
Response Time
Bandwidth Consumption
IO & CPU Resource Consumption

Dynamic Content Distribution - Naive

- Major reduction in CPU consumption and IO at the primary server site.
- Requires full replication of data and code.
- Why won’t it work this time?

Difficulties of Distributing Dynamic Content

- Duplicating data is difficult & against business logic.
  - Equifax cannot duplicate its data onto untrustworthy nodes.
  - Yahoo cannot easily duplicate its mail server because data changes so frequently.

Realistic Tradeoffs

- Increase in Response time. (Bad)
- Increase in Throughput. (Good)
- Decrease in server site resource consumption. (Good)
- Decrease in server site bandwidth? Hmmm.
  Sometimes Content Distribution is good.
  Sometimes Content Distribution is bad.
Outline

- What is dynamic content?
- How do we create these sites?
- What are the tradeoffs to distributing dynamic content?
- Our first approach to distributing dynamic content.
- Future work

General Rules Governing When to Distribute Dynamic Content

- Response Time should not exceed some constant $k$ for any user.
- Must reduce bandwidth to server site.

Understanding When to Distribute using Component Profiles

- A component profile for Component A is two vectors
  - The first vector quantifies bandwidth consumed between A and all other components.
  - The second vector counts roundtrips between A and all other components.
    - A roundtrip = 1 full blocking network read write between components.

What Effects Response Time

- Workload (i.e. Resource consumption)
  - Assume low level
- Network latency
  - We can estimate network latency and therefore response time at low concurrency levels.
  
  \[
  \text{Network Latency} = \sum_{\text{each wide area link } l} \left( RT_l \times LT_l(a) + \left\lfloor \frac{DV_l}{BW_l(a)} \right\rfloor \right)
  \]
Graph Version of Application Profiles: Bandwidth

What does the Min-Cut across this graph mean?

Graph Version of Component Profiles: Roundtrip

What does the Min-cut across this graph mean?

Four Distribution Strategies

- All Components at host server
- Min BW
- Min RT
- Min BW with Response Time Bound

Results of Different Strategies

Response Time Bound: 500ms

We see Min-BW cut saves bandwidth significantly but exceeds response time threshold.

Min-BW with response time bound performs well.
Future Work

- Consider Workload
- Globally consider all clients
- Dynamically adjust
  - Changing request mixes
  - Changing workloads