Web Caching Performance Evaluation : Quiz

Presenter: Sandeep K. S. Gupta

Reference:

- Chapter 2 Computer Networking, Kurose and Ross, Addison Wesley

Arizona State University

School of computing, informatics, and decision systems engineering

Web caching

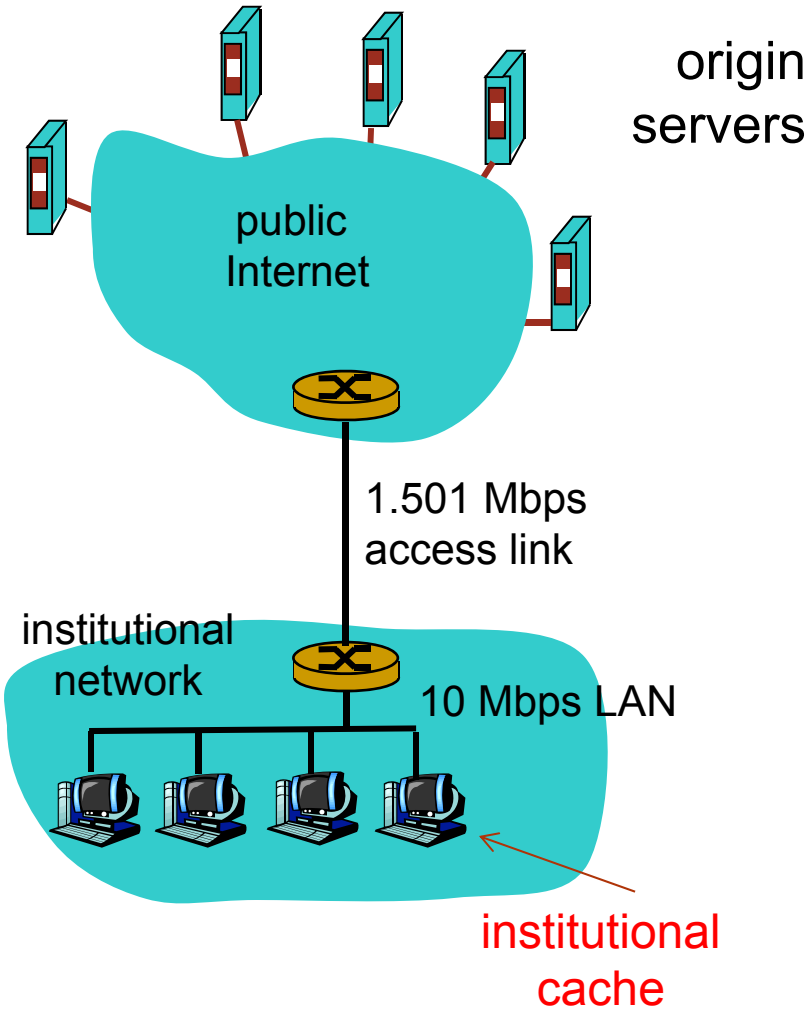
- Caches frequently accessed web pages "closer" to the user
- Cache acts as both client and server

Motivation

- ↓ response time for client request.
- ↓ traffic on an institution's access link.

Usage

- Cache is installed by Internet Service Provider (university, company, residential ISP)
- Internet dense with caches: enables "poor" content providers to effectively deliver content (but so does P2P file sharing)



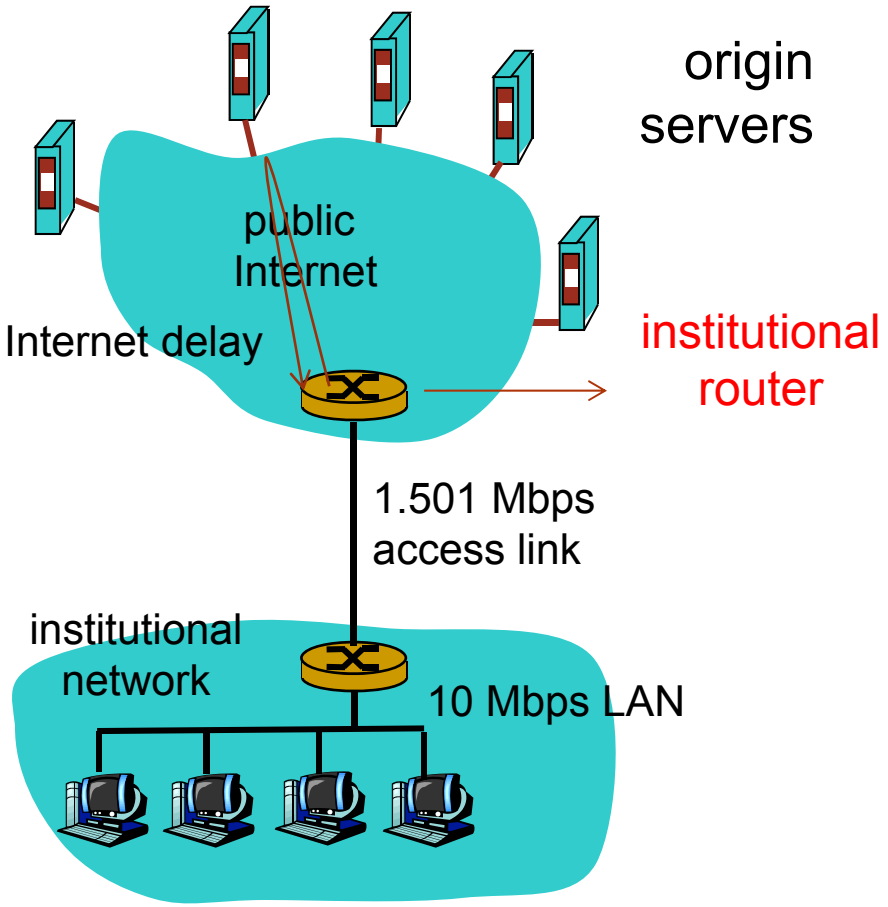
Initial Scenario

Assumptions

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

Questions:

- What is the utilization of LAN?
- What is the utilization of access link?
- What is the total delay (latency) per request?
  Total delay = Internet delay + access delay + LAN delay
  - only account for data transmission delay and ignore transmission delay for the request.



Improving Performance

- Performance Goal : avg. access delay < 3sec.
- Option 1: Fatter access link – 10Mbps
- Option 2: Adding Cache on the LAN
  - The estimated hit rate of 80%

Question: Which option is better in terms of average access delay?

Option 1: Fatter access link

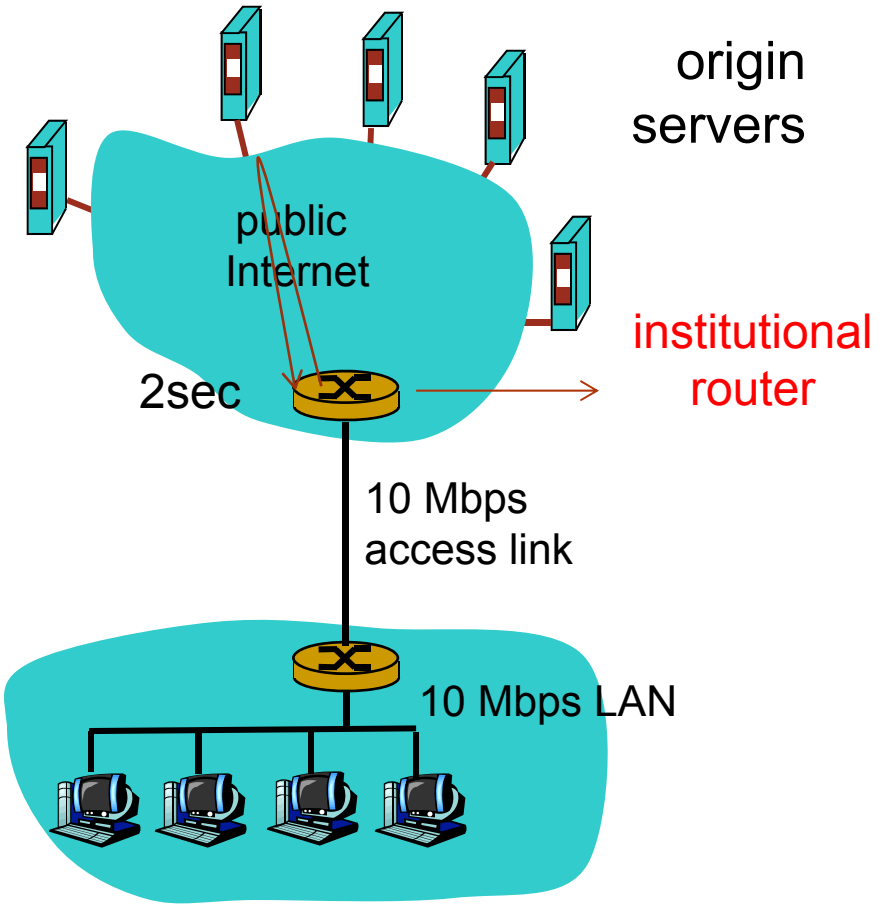
Possible solution

- increase bandwidth of access link to, say, 10 Mbps
  - often a costly upgrade

Performance Evaluation (Hint)

- utilization on LAN = ?
- utilization on access link = ?
- Total delay = Internet delay + access delay + LAN delay

= 2 sec + ? + ?



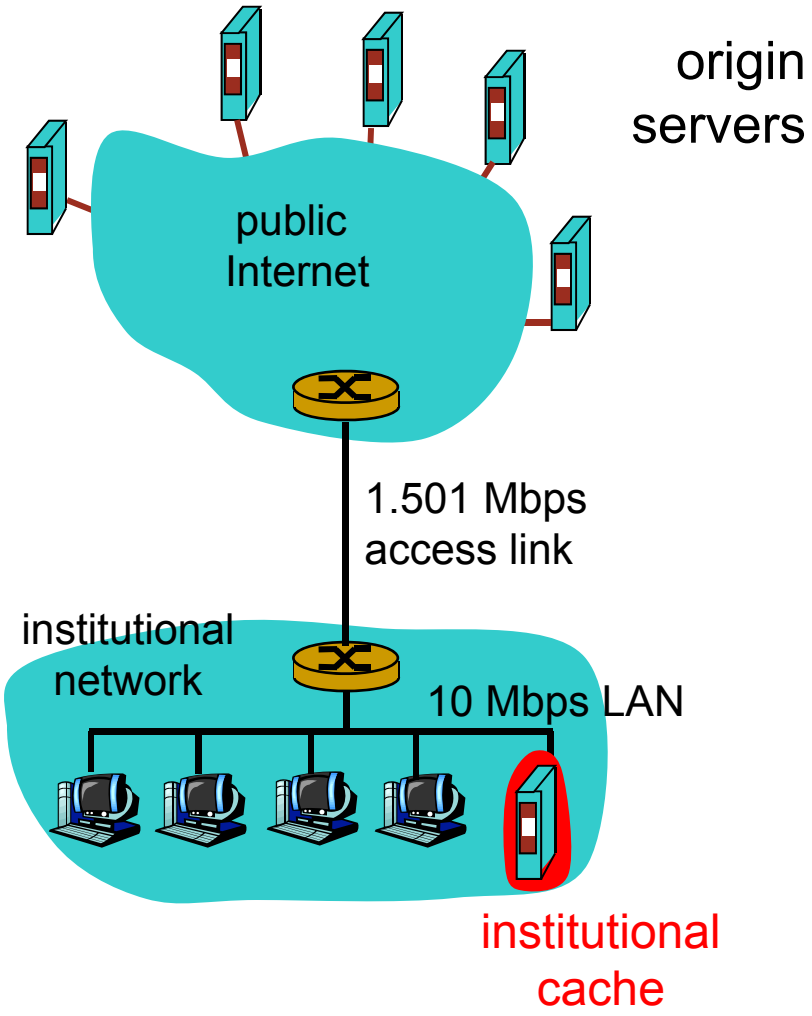
Option 2: Adding Cache on LAN

Install cache

- hit rate is .8

Performance Evaluation (hint)

- How many requests are satisfies locally on LAN?
- What is the effective request rate on the access link?
- What is the delay on the access link?
- total avg. delay ?



Evaluating Initial Performance

Assumptions

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

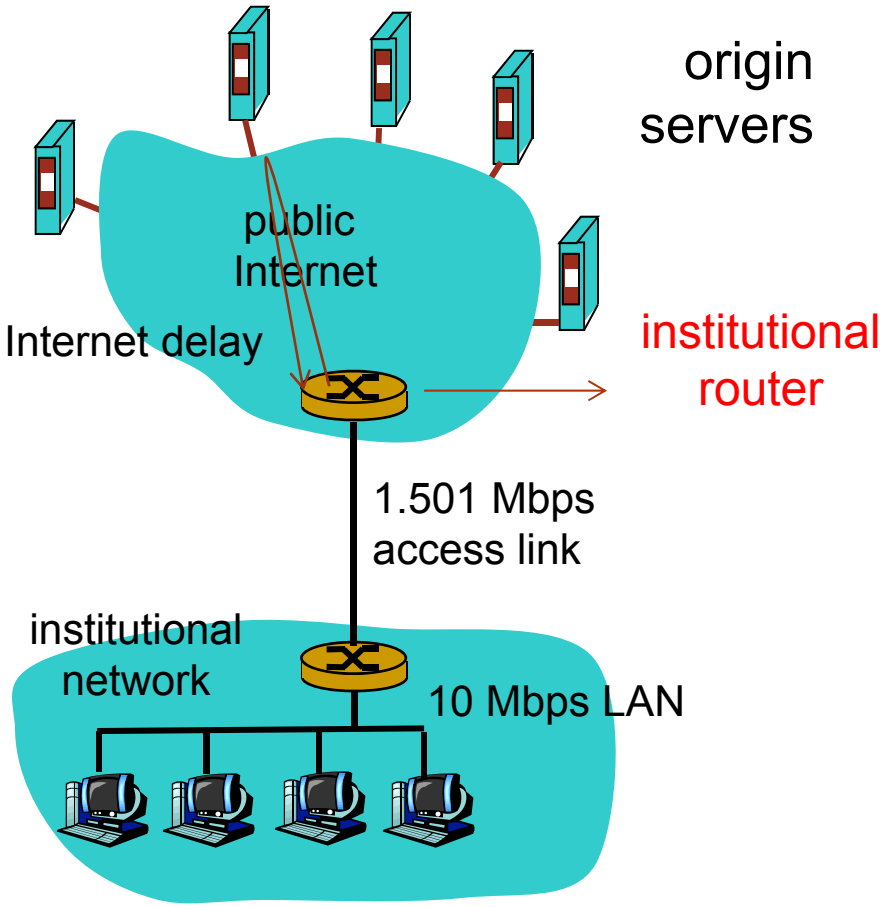
Performance Evaluation

- utilization on LAN = $15*10^5/10^7 = 15\%$
- utilization on access link (modeled as M/M/1 queuing system)
  - $\mu = 15.01$/sec
  - $\lambda = 15$/sec
  - $P = \lambda / \mu = .99$
- total delay = Internet delay + access delay + LAN delay

$= 2 \text{ sec} + 1/(\mu - \lambda) + 10^5/10^7$

$= 2 \text{ sec} + 100 \text{ sec} + 10 \text{ msec}$

$> 3\text{sec}$



Option 1: Fatter access link

Possible solution

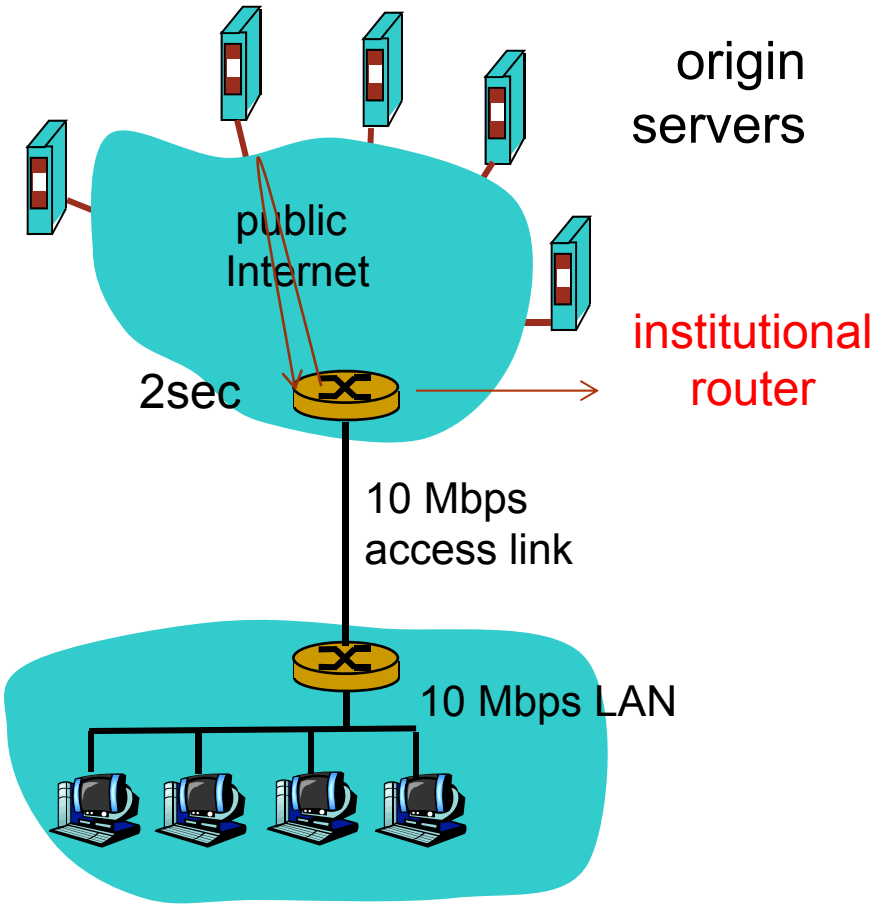
- increase bandwidth of access link to, say, 10 Mbps
  - often a costly upgrade

Performance Evaluation

- utilization on LAN = 15%
- utilization on access link = 15%
- Total delay = Internet delay + access delay + LAN delay

= 2 sec + 10 msec+ 10 msec

= 2.2 sec



Option 2: Adding Cache on LAN

Install cache

- hit rate is .8

Performance Evaluation

- 80% requests will be satisfied almost immediately
- 20% requests satisfied by origin server
- utilization of access link reduced to 20%,
  - resulting in negligible delays
  - 1/(15- (1/5)*15) = 1/12 = 80ms
- total avg delay = Internet delay + access delay + LAN delay

  = .2*(2.008) secs + .8*10 ms

  < 1 sec.