Why Web caching?

- Cost
  - Original motivation for adopting caches (esp. internationally)
  - Caching saves bandwidth (bandwidth is expensive)
  - 50% byte hit rate cuts bandwidth costs in half

- Performance
  - User: Reduces latency
    - RTT to cache lower than to server
  - Server: Reduces load
    - Caches filter requests to server
  - Network: Reduces load
    - Requests that hit in the cache do not travel all the way to server

HTTP headers for cache control

- Expires header are supported by practically every cache.
  - Especially good for making static images cacheable.

- Cache-Control HTTP header
  - max-age=[seconds]: specifies the maximum amount of time that a cached copy is considered fresh.
  - no-cache: force caches to submit the request to the original server for validation before releasing a cached copy.
  - no-store: instruct caches not to keep a local copy under any conditions.
  - must-revalidate: tell caches that they must obey any freshness information you give them, e.g. cannot serve client with a stale page.

HTTP header example

```
HTTP/1.1 200 OK
Date: Fri, 30 Oct 1998 13:19:41 GMT
Server: Apache/1.3.3 (Unix)
Cache-Control: max-age=3600, must-revalidate
Expires: Fri, 30 Oct 1998 14:19:41 GMT
Last-Modified: Mon, 29 Jun 1998 02:28:12 GMT
ETag: "3e86-410-3596fbbc"
Content-Length: 1040
Content-Type: text/html
```

- **ETag**: unique identifiers generated by the server and changed every time when the request resource is changed. Used by the caches to validate the freshness of their local copies.

Quick exercise

- Can we cache a document sent through the HTTP response below?

```
HTTP/1.1 200 OK
Server: Apache
X-Rack-Cache: miss
ETag: "e6811cdbcedf972c5e8105a89f637d39-gzip"
Status: 200
Content-Type: text/html; charset=utf-8
Expires: Mon, 29 Apr 2013 21:44:55 GMT
Cache-Control: max-age=0, no-cache, no-store
Pragma: no-cache
Date: Mon, 29 Apr 2013 21:44:55 GMT
```

When not to cache?

- If the HTTP response header tell the cache not to keep it, it won't.
- If no validator (e.g. Last-Modified header is absent) in the HTTP response, it will be considered uncacheable.
- If the HTTP is encrypted, it won't be cached.

Question to ponder?

- Why we can only get approximately 50% hit rate at maximum upon using Web Caching?

Summary

- Application layer HTTP protocol messages need transportation over the network.
- Use a TCP connection client->server (cache).

- Question: How does the application gain access to the transport layer?