CS 456 Hao Pan
Salahuddin, Mohammad
Spring 2018

Chapter 2.6 Video Streaming and Content Distribution Networks (CDNs)

07/06/2018 [Th]

2.6.1 Context

- Video traffic is a major consumer of Internet bandwidth, with Netflix and YouTube taking up 37% and 16% of residential downstream traffic respectively.
- With over 1 billion YouTube users and 75 million Netflix users, there is a challenge in how to reach so many users.
- A single mega-video server would *not* work because different users have *different capabilities* (wired vs. mobile, bandwidth rich vs. bandwidth poor...). There are also other reasons which will be explained later on.
- The solution is to use *distributed application-level infrastructure*.

2.6.2 Multimedia Video

- A video consists of a sequence of images being displayed at a constant rate.
- A digital image is made up of an array of pixels, where each pixel is represented by bits.
- In practice, we use **coding**: redundancy *within* and *between* images to decrease the number of bits used to encode an image.
- There are *two* types of coding:
  - **Spacial** (within image)
    - **Ex.** Instead of sending N values of the same color, send only two values: color value and number of repeated values.
  - **Temporal** (from one image to the next)
    - **Ex.** Instead of sending a complete frame, only send *differences* from previous frame.
- **CBR** (*constant bit rate*) is when the video encoding rate is fixed.
- **VBR** (*variable bit rate*) is when the video encoding rate changes as the amount of spatial and temporal coding changes.
- Ex. *MPEG1 (CD-ROM) [1.5Mbps], MPEG2 (DVD) [3-6 Mbps], MPEG4 (often used in Internet) [¡1 Mbps]*

2.6.3 Streaming Multimedia: DASH

- Stands for **Dynamic Adaptive Streaming over HTTP**
- The *server* divides a video file into *multiple chunks*, which are stored and encoded at different rates.
- A **manifest file** provides URLs for different chunks.
- The *client* periodically measures server-to-client bandwidth.
- Consulting manifest, the client requests one chunk at a time and chooses the *maximum coding rate sustainable* given its *current bandwidth*. Note that different coding rates can be chosen at different points in time.

- *"Intelligence" at client* exists, where the client determines:
  - *When* to request chunks so that buffer starvation and overflow does not occur.
  - *What encoding rate* to request, where higher quality rates are chosen when more bandwidth is available.
  - *Where* to request chunks, so that it can request chunks from a URL server that's closer or has higher available bandwidth.

2.6.4 Challenge

How to stream content, selected from millions of videos, to millions of simultaneous users?

2.6.4.1 Option 1

- Use a single large "mega-server".
- This method is not very reliable because:
  - There is a single point of failure.
  - There is a single point of network congestion.
  - Clients far away have long paths.
  - Multiple copies of a video may be sent over an outgoing link.
  - Simply put, *it does not scale*.

2.6.4.2 Option 2

- Store and serve multiple copies of videos at multiple geographically distributed sites (CDN)
- **Enter deep**: Pushing CDN servers into various access networks so that they're closer to users.
- **Bring home**: Smaller number (10s) of larger clusters in POPs near, but not within, access networks.

2.6.5 A Closer Look

- CDN stores copies of content at CDN nodes (**ex.** Netflix stores copies of MadMen).
- Subscribers request content from the CDN. The request is either directed to a nearby copy, which retrieves the content, or it is directed to a different copy if the network path is congested.
- See example of steps of a video request on slide 2-96.
- See example of how Netflix works on slide 2-97.