A Joint Online Transcoding and Delivery Approach for Dynamic Adaptive Streaming

ALAN ZHUANG a@pear.hk

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

BACKGROUND

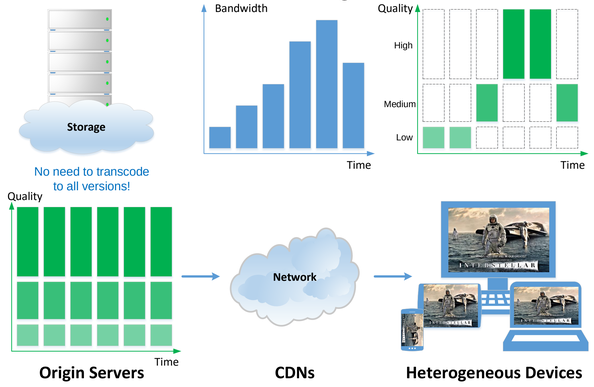
Popular online video sites

YouTube NETFLIX 优酷 HD

transcode their vidoes to multiple versions

$$\rightarrow \{\, mp4, [H.264, AAC] \,\} \begin{cases} version_1 & (x_1\ kbps) \\ version_2 & (x_2\ kbps) \\ \vdots & \vdots \\ version_n & (x_n\ kbps) \end{cases}$$

and then stream them through the network,



so that users can enjoy a better watching experience according to their current bandwidths and device decoding abilities.

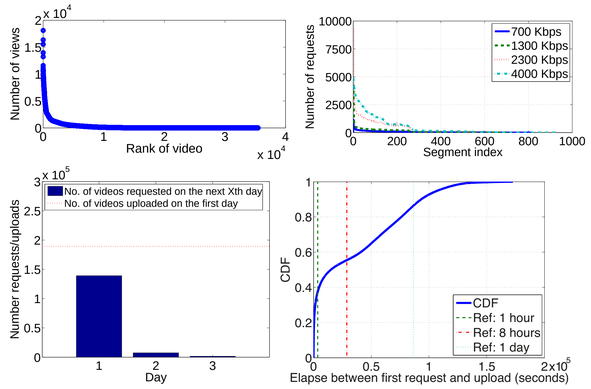
PROBLEM

They all transcode origin videos to all versions. In such scheme,

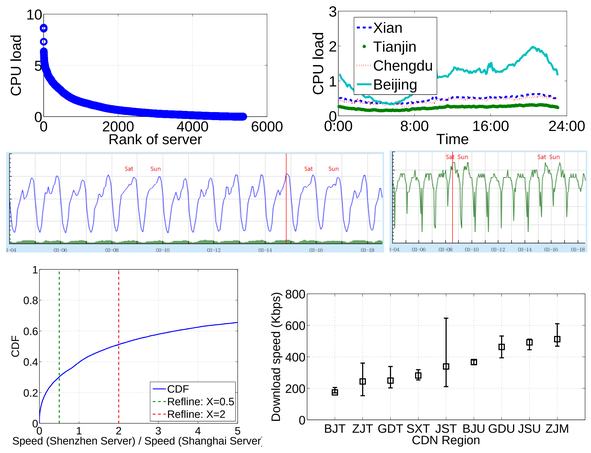
- only a small set of candidate bitrates to manually choose from; cannot effectively adapt to the changing network conditions
- huge computing resource consumption
  - H.264: 1/3 to 2/3 of playback time
  - H.265: 30+ times of playback time
  - 1 CPU: 1-2 concurrent coding tasks
- oblivious of users' *preferences* of different *peering servers*

OBSERVATIONS & INSIGHTS

Video Viewing Patterns (Pro. & UGC)



CDN Patterns (CPU load & bandwidth)



Insights:

⋆§ No necessary to pre-transcode/store all

⋆§ Most backend servers in CDNs are predictably(stably) idle

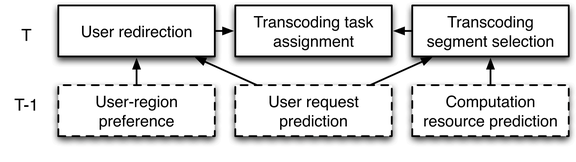
Can be scheduled for transcoding?

⋆§ users have preferences in CDN regions

⋆§ regions have preferences in media versions

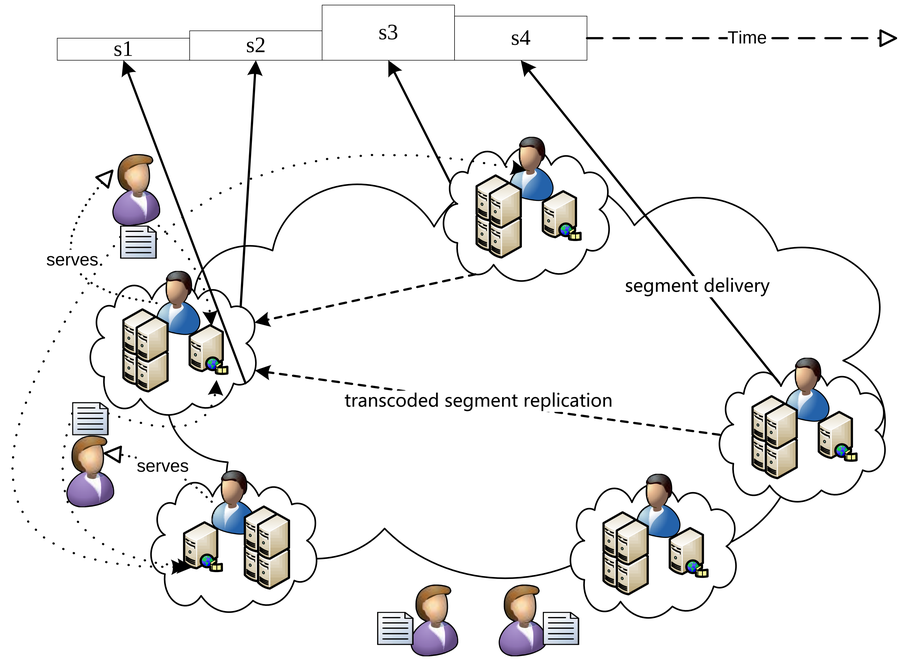
FRAMEWORK

3 phases joint optimization



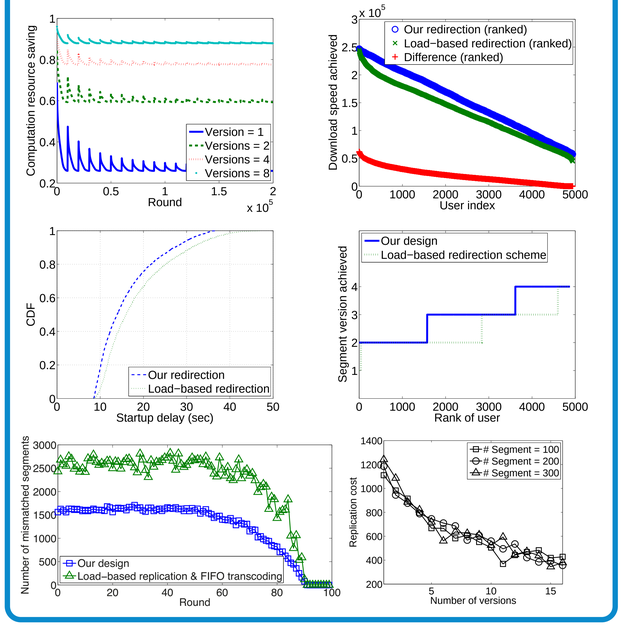
METHODS

Girls Chasing *Boys*! (Stable Maching plus Linear Programming)



$$\max_{D^{(T)}} \sum_{u\in\mathbf{U}^{(T)}, r\in\mathbf{R}} H_{u,r} D^{(T)}_{u,r};\ \max_{P^{(T)}} \sum_{(s,v)\in\mathbf{K}^{(T)}} P^{(T)}_{(s,v)} e^{(T)}_{(s,v)};\ \min_{A^{(T)}} \sum_{(s,v)\in\mathbf{E}^{(T)}} \sum_{r\in\mathbf{R}} A^{(T)}_{(s,v),r} F_{(s,v),r}$$

RESULTS



HIGHTLIGHTS

- 44.8% users enjoy higher bitrate versions than the load-balanced redirection scheme
- 4.5x users enjoy the highest possible bitrate
- Mismatch rate reduced by over 42.2%
- $\sim$ 80% computing resource saved when the number of versions is 4.
  The higher the number is, the more resource our approach saves.

AFTERWORD

Transcoding and delivery can, and should be considered & done jointly.

§ A joint work with THU, HKU, Tencent.

§ Appeared in IEEE TMM 2015.

§ In serving WeChat video clip, QZone video...