1 Modeling of the P2P service migration problem

We suppose there are M videos, and N ISPs. There are one on-premise server and one cloud node in each ISP.

Notation definition:

C_s^j: storage capacity of the on-premise server at the $j-th$ ISP

C_u^j: upload bandwidth capacity of the on-premise server at the $j-th$ ISP

h_j: charging rate for storage on the cloud at the $j-th$ ISP

k_j: charging rate for upload bandwidth on the cloud at the $j-th$ ISP

s_m: storage of $m-th$ video

$x_m^j = \{0,1\}, m = 1, ..., M$: $x_m^j = 1$ if the placement of the $m-th$ video is on the on-premise server at the $j-th$ ISP; $x_m^j = 0$ otherwise;

$y_m^j = \{0,1\}, m = 1, ..., M$: $y_m^j = 1$ if the placement of the $m-th$ video is on the cloud at the $j-th$ ISP; $y_m^j = 0$ otherwise;

r_m^j: request rate of the $m-th$ video from the $j-th$ ISP, i.e., the bandwidth demand is $s_m r_m^j$.

R_{ji}^m: percentage of requests from j for video m is routed to on-premise server i

T_{ji}^m: percentage of requests from j for video m is routed to cloud i

1.1 Optimization of the problem

$min \sum_{m=1}^{M} \sum_{j=1}^{N} \sum_{i=1}^{N} (s_m r_m^j T_{ji} k + s_m h) y_m^j - \alpha \sum_{m=1}^{M} \sum_{j=1}^{N} s_m r_{jj}$ (maximize local traffic, i.e., minimize delay)

subject to:

$y_m^j = \{0,1\}, \forall j = 1, ..., N, \forall m = 1, ...M$

$x_m^j = \{0,1\}, \forall j = 1, ..., N, \forall m = 1, ...M$

$\sum_{i=1}^{N} (R_{ji}^m + T_{ji}^m) = 1, \forall j = 1, ...N, \forall m = 1, ..., M$

$0 \leq R_{ji}^m \leq x_m^i$

$0 \leq T_{ji}^m \leq y_m^i$

$\sum_{m=1}^{M} s_m x_m^j \leq C_s^j, \forall j$ (on-premise server's storage constraint)

$\sum_{m=1}^{M} \sum_{j=1}^{N} s_m r_m^j R_{ji}^m \leq C_u^i, \forall i = 1, ...N$ (on-premise server's upload bandwidth constraint)

Note:

known values: C_s^j, C_u^j, h_j, k_j, s_m, r_m^j

optimization variables: r_m^j, x_m^j, y_m^j, R_{ji}^m, T_{ji}^m

This is a one-time optimization of placement of videos across cloud and on-premise server. We can further do optimization over time using Lyapunov technique.

2 Lyapunov Optimization

2.1 Modeling requests and replication of blocks as 2 queues

For the first queue, we model the requests as arrival process, the serving of blocks as departure process. replication of block. For the second queue, we model downloading of blocks as arrival process, and eviction of block as departure process.

2.2 Model buffer level as queue

We intend to model the live-streaming system.
We model cloud as node 0. The uploading rate is modeled as departure rate and downloading as arrival rate. The uploading bandwidth of peers varies over time (calculated by fluid model). The buffer level is modeled as queue length. The uploading bandwidth is a random variable over time.

The constraint is buffer level should always be larger than 0.

For the cloud, we simply minimize the average uploading rate over time. While for the dedicated server, in order to reflect the 95-percentile rule, we not only minimize the average uploading rate, but also minimize the variance of uploading rate (hint: for same avarege uploading rate, when the variance of uploading rate is smaller, the charge based on 95-percentile rule is smaller). And the variance can be expressed by $E[X^2] - (E[X])^2$.

We may minimize the average (and variance of) buffer level simultaneously.