Model Draft

Yu Wu*, Chuan Wu*, Francis C.M. Lau*
*Department of Computer Science, The University of Hong Kong,
Email: {ywu,cwu,fcmlau}@cs.hku.hk

Abstract

I. Model Overview

The model consists of a set of geographically diverse cloud clusters F, a set of videos O and a set of clients D. ($D^{(f)}$ denotes the consolidated viewer group within domain of cloud cluster f ($f \in F$).) Without loss of generality, I will first assume all the videos have unit length. (Extend this when the model is done)

A. Alphabet Soup

1) Each cloud cluster is assigned a storage capacity, S_f.
2) Each cloud cluster has a bandwidth capacity, μ_f. Here I plan to borrow the idea from our ICDCS paper's model *i.e.*, bandwidth will be abstracted into a VM instance, which will actually provide the bandwidth.
3) $x_{jf}^{(o)}$ denotes the variable indicating whether the request for video o issued from viewer j will be directed to cloud cluster f.
4) $y_f^{(o)}$ denotes the variable indicating whether to store a copy of video o at the cloud cluster f.
5) c_f denotes the storage cost of cloud cluster f.
6) v_f denotes the transferring cost from viewer j's location to cloud cluster f.
7) R_{jf} denotes the transferring latency from viewer j's location to cloud cluster f. It can be assigned with value of RTT between these two geographical regions.

There should be a mapping function to map user j to a location f. For simplicity, we can denote it as $D^{-1}(j)$. In that way, we can represent $x_{jf}^{(o)}$, c_{jf} and R_{jf} as $x_{D^{-1}(j)f}^{(o)}$, $c_{D^{-1}(j)f}$ and $R_{D^{-1}(j)f}$ respectively. For clearness, I will use j afterwards.

B. Objective function

To minimize the operation cost, based on the premise that the expected average global latency should be bounded below some tolerant value.

$$min \sum_{o \in O} \sum_{f \in F} y_f^{(o)} \times c_f + \sum_{o \in O} \sum_{f \in F} \sum_{j \in D^{(f)}} x_{jf}^{(o)} \times v_f$$

C. Constraints

1) *Storage*
$\sum_{o \in O} y_f^{(o)} \leq S_f$, $\forall f \in F$

2) *VM capacity*
$\sum_{o \in O} \sum_{j \in D} x_{jf}^{(o)} \leq \mu_f$, $\forall f \in F$

3) *Placement*
$\sum_{f\in F} y_f^{(o)} \geq 1$, $\forall o \in O$

4) *Latency guarantee*
$\frac{\sum_{o\in O}\sum_{f\in F}\sum_{j\in D(f)} x_{jf}^{(o)} \times R_{jf}}{|D|} \leq R_{threshold}$, where $R_{threshold}$ is an input into the system.

5) Variable constraint
$y_f^{(o)} \in \{0,1\}, x_{jf}^{(o)} \in \{0,1\}$
6) $x_{jf}^{(o)} \leq y_f^{(o)}$, $\forall j \in D$

II. Alternative LP (Relaxation)

Obviously, the optimization in Sec. I is an integer problem. Here we want to make an intuitive relaxation. The reason is two-fold. First, the number of users makes the optimization problem too large to solve. Second, we want to transform the original problem into a more tractable one.

A. Consolidate users

As what we have assumed, at any time, each user can at most view one video. So we can treat the users within one specific region f ($f \in F$) as one, which will make our optimization much slimmer. Based on that, we are able to eliminate all the variables $x_{jf}^{(o)}$ and consolidates them as one viewer. Suppose the total user set at time slot T is represented as D_T, so the viewer set in region f at that time slot is $D_T^{(f)}$. We introduce a new variable $\alpha_{jf}^{(o)}$, which denote the portion of request for video o issued from the aggregate user j to cloud cluster f. To note that, $\alpha_{jf}^{(o)}$ is a fractional variable. Due to our charging mode, the storage deployment is scheduled at a larger time scale while the VM rental is done at a smaller one, i.e. T. From above, the original ILP has only one type of integer variable $y_f^{(o)}$ and the original optimization problem is changed to,

$$min \sum_{o\in O}\sum_{f\in F} y_f^{(o)} \times c_f + \sum_{o\in O}\sum_{f\in F}\sum_{j\in F} \alpha_{jf}^{(o)} \times D_T^{(f)} \times v_f$$

B. Constraints

1) *Storage*
$\sum_{o\in O} y_f^{(o)} \leq S_f$, $\forall f \in F$

2) *VM capacity*
$\sum_{o\in O}\sum_{j\in F} \alpha_{jf}^{(o)} \times D_T^{(f)} \leq \mu_f$, $\forall f \in F$

3) *Placement*
$\sum_{f\in F} y_f^{(o)} \geq 1$, $\forall o \in O$

4) *Latency guarantee*
$\frac{\sum_{o\in O}\sum_{f\in F}\sum_{j\in F} \alpha_{jf}^{(o)} \times D_T^{(f)} \times R_{jf}}{|D|} \leq R_{threshold}$, where $R_{threshold}$ is an input into the system.

5) Variable constraint
$y_f^{(o)} \in \{0,1\}, \alpha_{jf}^{(o)} \in [0,1]$
6) $\alpha_{jf}^{(o)} \leq y_f^{(o)}$, $\forall j, f \in F$
7) $\sum_{f\in F} \alpha_{jf}^{(o)} = 1$

C. How to solve?

If we relax the variable $y_f^{(o)}$, the optimization problem in Sec. II-A is a problem with complicating constraints and we can utilize an efficient dual decomposition to solve it.