A Theory of Cloud Bandwidth Pricing for Video-on-Demand Providers

Presented by Xuanjia Qiu

Feb. 16, 2012

- Authors: Di Niu, Chen Feng, Baochun Li (all from Dept. of ECE, UoT)
- Publication: INFOCOM 2012

Background

- Network access in current cloud computing is offered with usage-based pricing, with no bandwidth capacity guarantee.
- Not suitable for bandwidth-intensive application, such as VoD.
- Supposing cloud providers offer bandwidth capacity guarantee, except letting each VoD provider reserves bandwidth capacity from the cloud providers, are there any better solutions?

Idea

- Introducing a broker to collect demands from many VoD providers. It is hopeful that jointly booking can reduce cost compared with individually booking, so that each VoD providers can enjoy a discount and the broker can get profit.
  - Demands of multiple applications may be statistically multiplexed to save bandwidth reservation.
  - Anti-correlation (i.e., negative correlation) : a relationship in which one value increases as the other decreases.
  - Similar to Ethernet.

Problem

- Payment:
  - The broker pays to cloud providers according to a fixed pricing strategy, and decides the amount of reservation
  - The pricing policy between the broker and VoD providers are to be discussed, which affects (1) the profit and the amount of reservation of the broker and (2) the cost/enjoyed discount and the utility of VoD providers

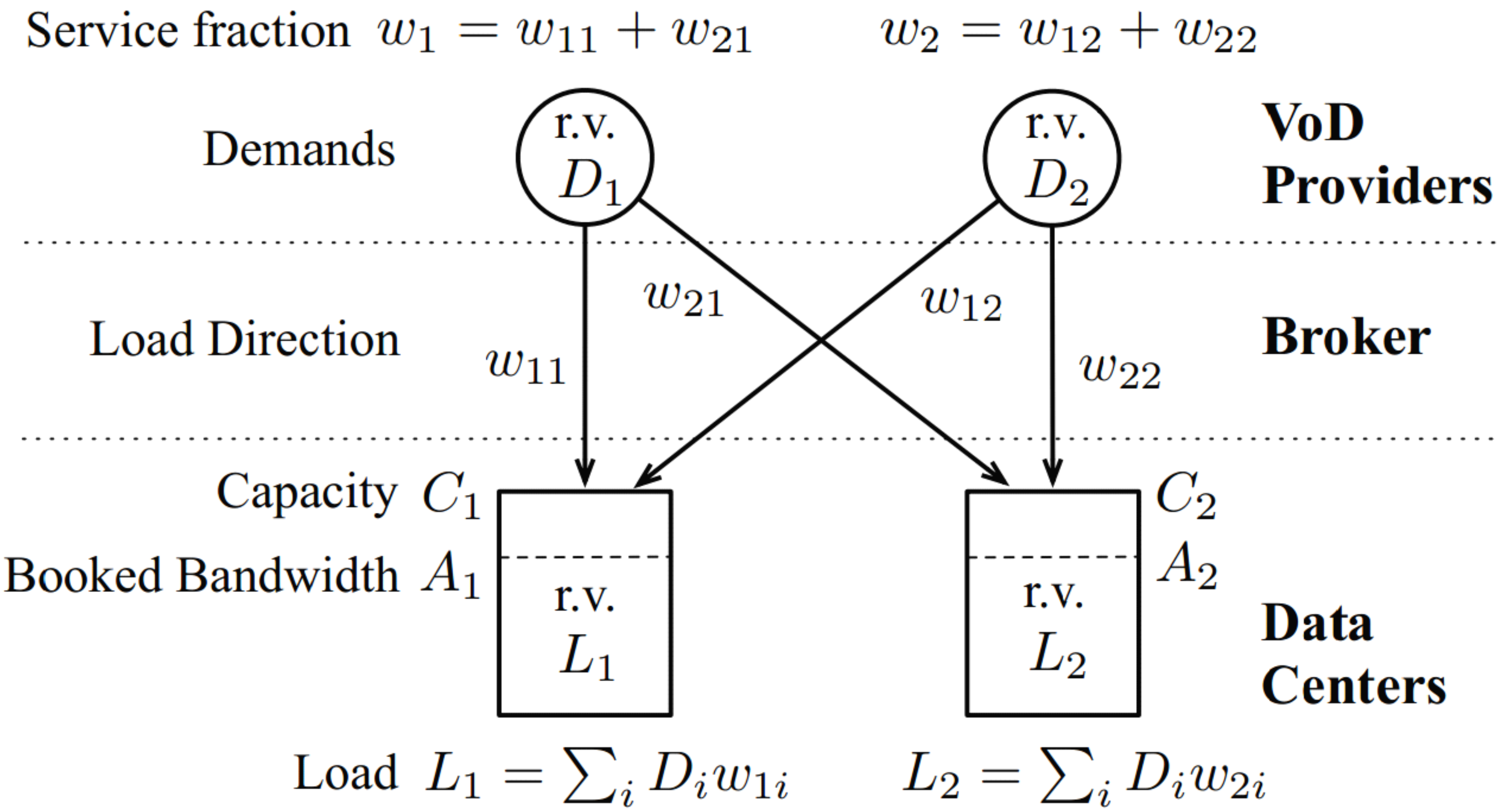
Problem

- Conflict of interest:
  - Cloud providers want to maximize resource efficiency (i.e., workload consolidation)
  - Brokers want to maximize profit.
  - VoD providers want to minimize cost.

Contribution of the paper

- Present a model:
  - Formulate the workload consolidation problem and give the optimal solution to the problem
  - Formulate the broker-profit-maximization probem, and answer under what pricing policy (i.e. good pricing region), the optimal value to the problem is equal to the workload consolidation problem
  - Show that in the free market where each VoD provider submits a pricing strategy to the broker, the submitted pricing strategies converge to a unique Nash equilibrium, which is the lower bound of the good pricing region

Model



Service fraction $w_1 = w_{11} + w_{21}$
$w_2 = w_{12} + w_{22}$
Demands
r.v. D_1
r.v. D_2
VoD Providers
Load Direction
w_{11}
w_{21}
w_{12}
w_{22}
Broker
Capacity C_1
Booked Bandwidth A_1
r.v. L_1
C_2
A_2
r.v. L_2
Data Centers
Load $L_1 = \sum_i D_i w_{1i}$
$L_2 = \sum_i D_i w_{2i}$

- Consider the demand and reservation in a short time slot, e.g., 10 minutes.
- VoD provider i's bandwidth demand (based on prediction) is a random variable D_i with mean μ_i and variable σ_i
- Optimizing variables
  - load direction matrix $W=[w_{si}]_{S\times N}$: Portion of i's demand D_i directed to and served by cloud provider s.

Workload consolidation problem

$$\min_{\mathbf{W}} \sum_s A_s$$
$$\text{s.t.} \quad A_s \le C_s, \quad \forall s,$$
$$w_i = 1, \quad \forall i.$$

Where

$$A_s := f_\epsilon(L_s)$$
$$L_s = \sum_i w_{si} D_i.$$
$$f_\epsilon(X) = \mathbf{E}[X] + \theta\sqrt{\mathbf{Var}[X]}, \quad \theta = F^{-1}(1-\epsilon)$$

Theorem 2: When $C_{\mathsf{sum}} \geq \boldsymbol{\mu}^{\mathsf{T}}\mathbf{1} + \theta\sqrt{\mathbf{1}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{1}}$, an optimal solution $[w_{si}^*]$ to (15) is given by

$$w_{si}^* = \alpha_s, \quad \forall i, \quad s = 1, \ldots, S, \tag{19}$$

where $\alpha_1, \ldots, \alpha_S$ can be any solution to

$$\sum_s \alpha_s = 1,\ 0 \leq \alpha_s \leq \min\left\{1, \frac{C_s}{\boldsymbol{\mu}^{\mathsf{T}}\mathbf{1} + \theta\sqrt{\mathbf{1}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{1}}}\right\},\ \forall s. \tag{20}$$

When $C_{\mathsf{sum}} < \boldsymbol{\mu}^{\mathsf{T}}\mathbf{1} + \theta\sqrt{\mathbf{1}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{1}}$, there is no feasible solution that satisfies constraints (16) to (18).

The maximum bandwidth saving of joint bandwidth booking over individual booking for each tenant is

$$\begin{aligned}\Delta B(\mathbf{W}^*) &= \textstyle\sum_i B_i - \sum_s A_s \\ &= \textstyle\sum_i(\mu_i + \theta\sigma_i) - \sum_s(\boldsymbol{\mu}^\top \mathbf{w^*}_s + \theta\sqrt{\mathbf{w^*}_s^\top \boldsymbol{\Sigma}\mathbf{w^*}_s}) \\ &= \theta(\boldsymbol{\sigma}^\top \mathbf{1} - \sqrt{\mathbf{1}^\top \boldsymbol{\Sigma}\mathbf{1}}) = \theta(\textstyle\sum_i \sigma_i - \sigma_M),\end{aligned} \tag{21}$$

which is θ times the gap between the sum of all demand standard deviations and the standard deviation of all demands combined. This confirms the belief that statistical multiplexing saves resource reservation.

Broker-profit-maximization problem

$$\max_{\mathbf{W}} R(\mathbf{W}) = \sum_i P_i(w_i) - \sum_s A_s$$

$$\text{s.t.} \quad A_s \leq C_s, \quad \forall s.$$

where

(1) Pricing strategy $P_i(.)$ is a concave function on [0,1] with $P_i(0)=0$.

(2) Cloud provider charges $1 for every unit bandwidth reservation.

Observation

Brokers always have incentive to operate, but may deny demand.

Theorem 3: Broker profit maximization (22) and cloud workload consolidation (15) have a same optimal solution (19), if and only if

$$P_i'(1) \geq \mu_i + \theta \cdot \frac{\sigma_{iM}}{\sigma_M}, \quad \forall i, \tag{26}$$

where σ_{iM} is the covariance between D_i and $\sum_i D_i$ given by (12) and σ_M is the standard deviation of $\sum_i D_i$ given by (13). Furthermore, if $P_i'(1) < \mu_i + \theta\sigma_{iM}/\sigma_M$ for some i, then $\mathbf{w}^* \neq \mathbf{1}$.

Good pricing region

Corollary 1: In a good pricing policy $\{P_i(\cdot) : i = 1, \ldots, N\}$, each $P_i(\cdot)$ must satisfy $\forall w_i \in [0, 1]$,

$$(\mu_i + \theta \cdot \frac{\sigma_{iM}}{\sigma_M})w_i \leq P_i(w_i) \leq (\mu_i + \theta\sigma_i)w_i. \quad (32)$$

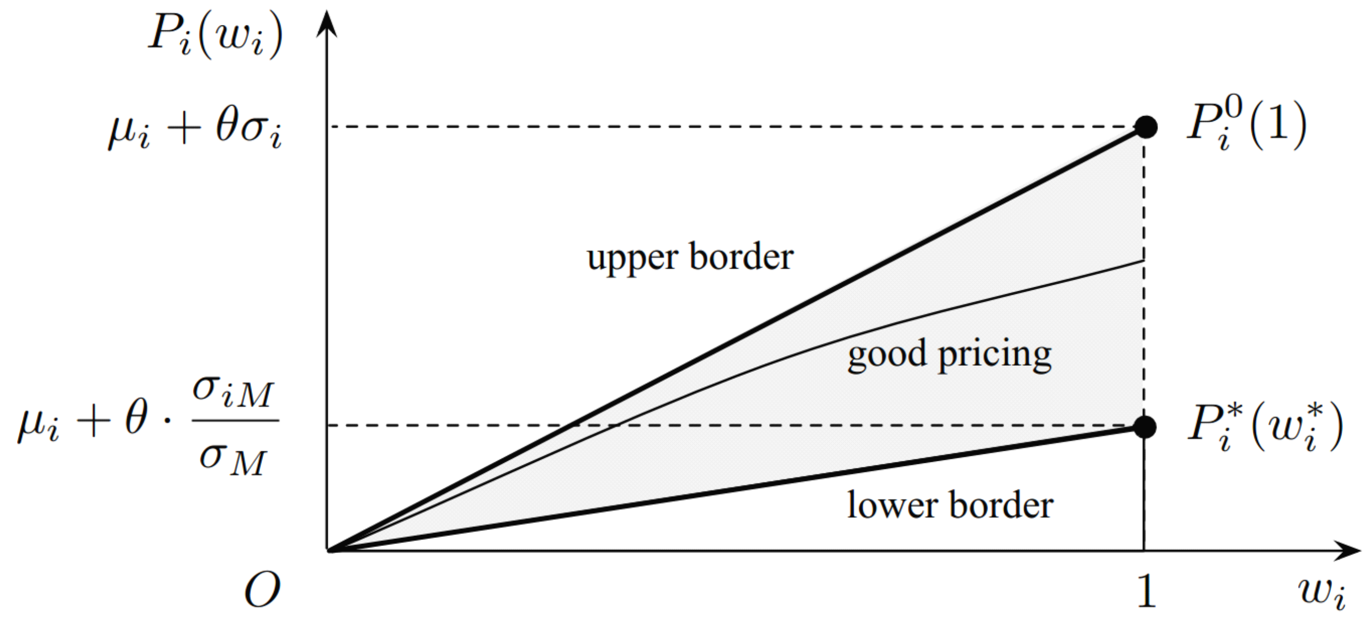


Fig. 2. The region of $P_i(\cdot)$ in a good pricing policy $\{P_i(\cdot)\}$. $P_i(\cdot)$ is between $P_i^*(\cdot)$ and $P_i^0(\cdot)$, and satisfies $P_i'(1) \geq P_i^{*\prime}(1)$.

Discussion about free markets

In a free market where each selfish tenant (VoD provider) competes for service by submitting a pricing strategy $P_i(\cdot)$, what will $\{P_i(\cdot)\}$ eventually look like?

- VoD provider i's utility function

$$U_i[P_1(\cdot), \ldots, P_N(\cdot)] = \begin{cases} -P_i(w_i^*), & \text{if } w_i^* = 1, \\ -\infty, & \text{if } w_i^* < 1, \end{cases}$$

Unique Nash equilibrium in free markets

Theorem 4: If tenants have utility (10) and the broker decides $\mathbf{W}^*$ by maximizing its profit via (22), then $\{P_i(\cdot)\}$ will converge to a **unique Nash equilibrium** $\{P_i^*(\cdot)\}$, where

$$P_i^*(w_i) = (\mu_i + \theta \cdot \frac{\sigma_{iM}}{\sigma_M})w_i, \quad 0 \leq w_i \leq 1, \tag{37}$$

where σ_{iM} and σ_M are given by (12) and (13), respectively.

Theorem 1: In a free market, $\{P_i(\cdot)\}$ will converge to a **unique Nash equilibrium** $\{P_i^*(\cdot)\}$, where $w_i^* = 1$ and

$$P_i^*(w_i^*) = \mu_i + \theta\sigma_i\rho_{iM}, \tag{11}$$

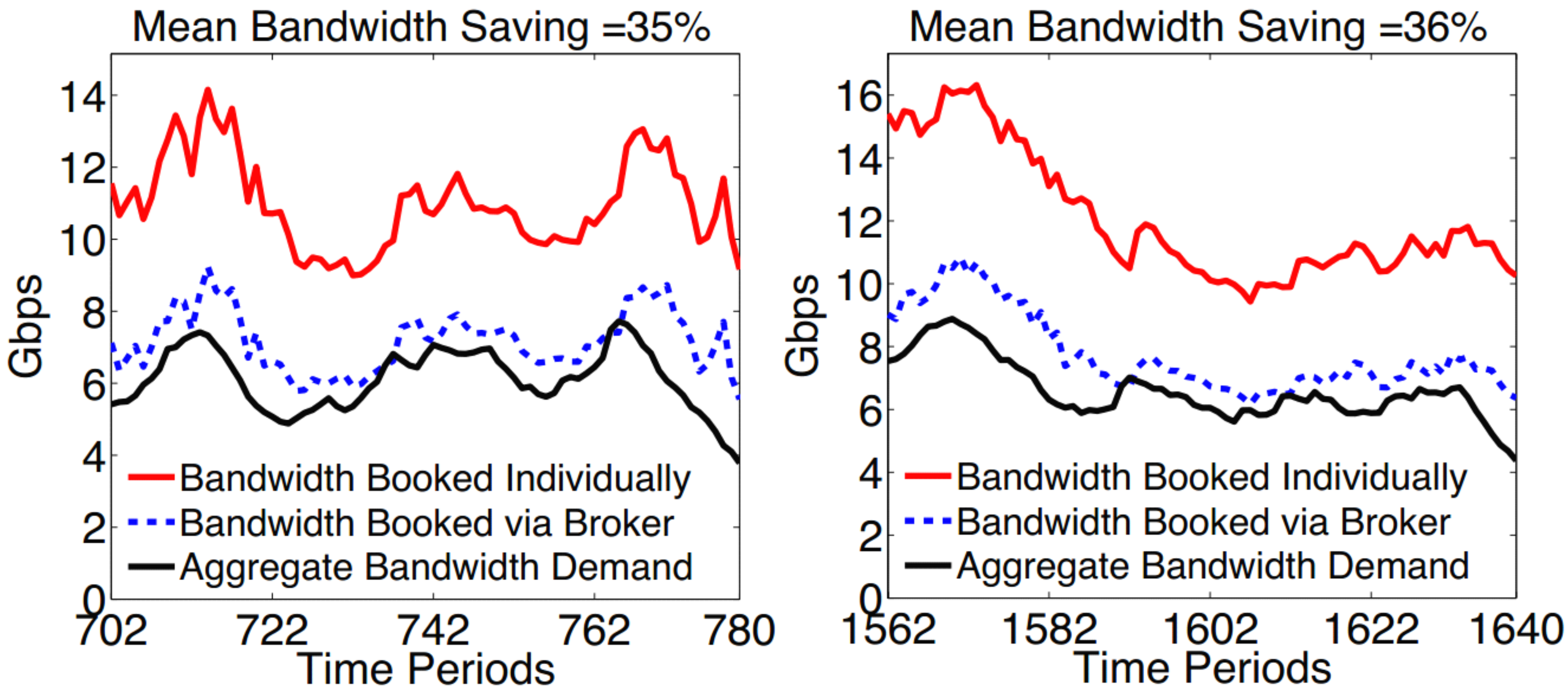
$\rho_{iM} \in [-1, 1]$ being the correlation coefficient between D_i and $\sum_i D_i$.

The prices at equilibrium are the lower border of the good pricing region.

Simulation

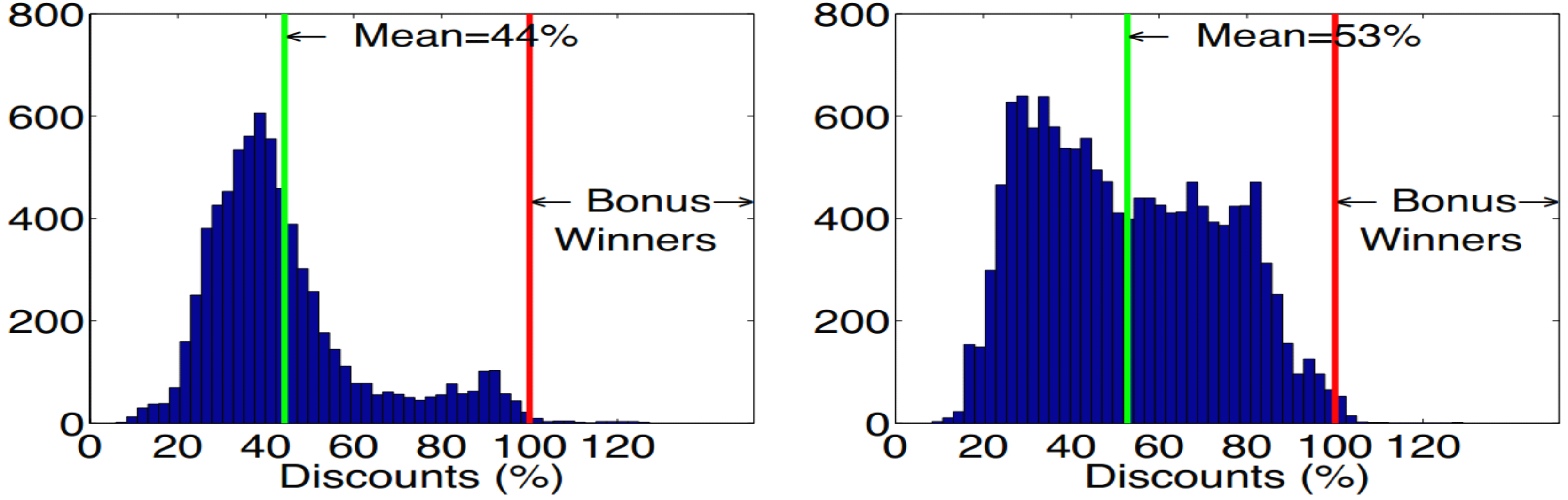
- Use trace of UUSee
  - Let each video channel represents a VoD provider.
- Relatively simple, only show
  - Aggregate bandwidth reservation saving
  - Discount enjoyed by the VoD providers

Aggregate bandwidth reservation saving



(a) $\epsilon = 5\%$, $\epsilon' = 5.06\%$ (b) $\epsilon = 1\%$, $\epsilon' = 1.27\%$

Discount enjoyed by the VoD providers



(a) t =702—780, 91 channels (b) t =1562—1640, 176 channels

Some VoD providers are risk-neutralizers that earn bonus for having demand negatively correlated to the market.

Summary and insight

- Idea: statistical multiplexing
  - Leverage the fact that VoD demand is fractionally splittable into video requests, which can be optimally directed to different clouds and statistically mixed toward workload consolidation.
  - The idea of "View-Upload Decoupling (VUD) design" in the paper "Modeling and Analysis of Multichannel P2P Live Video Systems" (by Di Wu et. al, published on INFOCOM 2010) is also statistical multiplexing -- bandwidth capacity contributed by all peers forms a resource pool and multiplexed. (difference: fixed resource amount+ minimize risk v.s. fixed risk + minimize resource amount)
- Possible extension
  - Although there are multiple clouds in the model, in fact, their role is almost the same as a single cloud, because their prices are the same and constant ($1). What if clouds compete against each other, or offer different QoS and charge with different pricing strategies?

Thank you