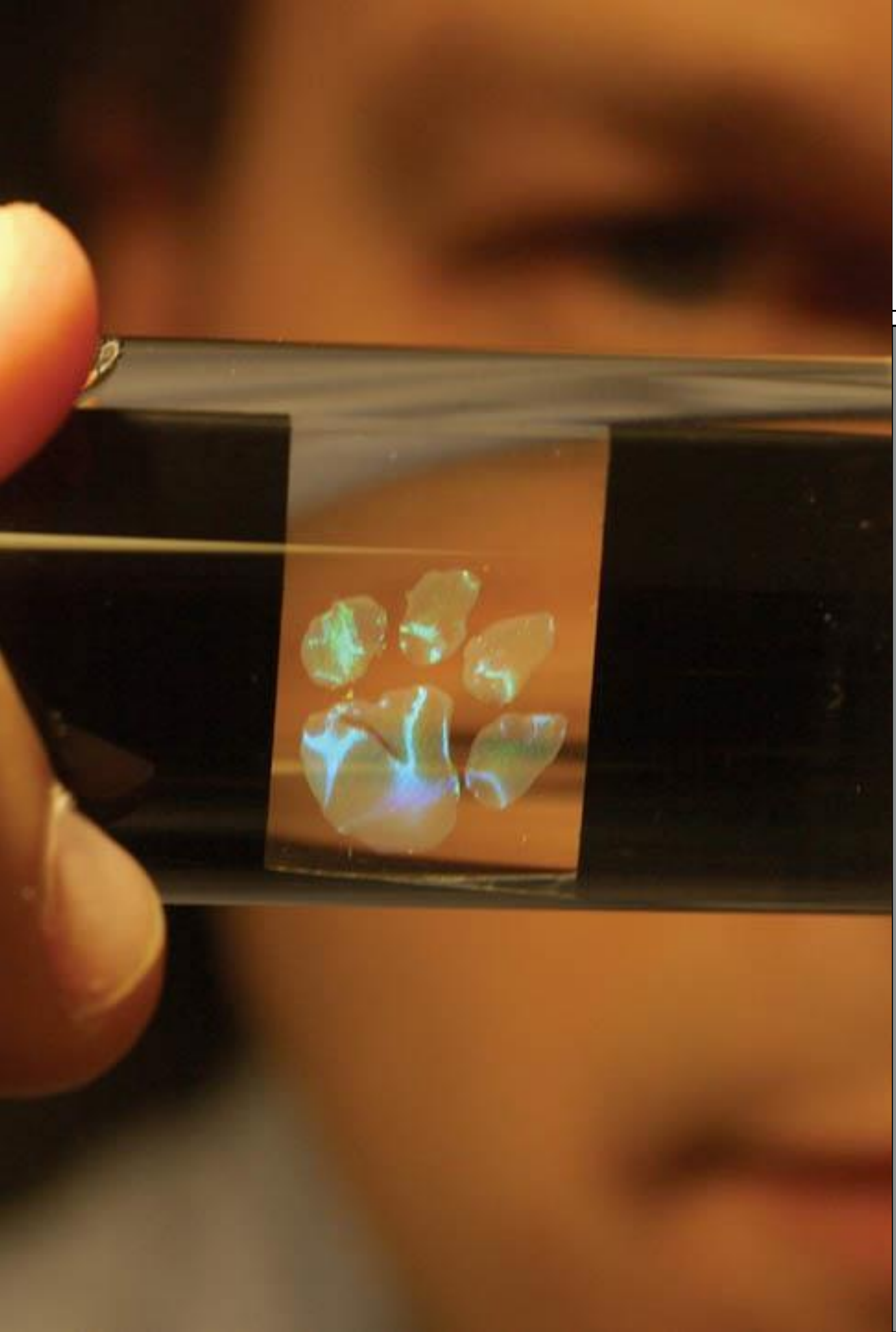


# New Bandwidth Sharing and Pricing Policies to Achieve a Win-Win Situation for Cloud Provider and Tenants

Haiying Shen and **Zhuozhao Li**

Dept. of Electrical and Computer Engineering
Clemson University, SC, USA

# Outline

- Motivation and Introduction
- Our Proposed Price Model
  - Payment of tenants
  - Network bandwidth sharing policies
  - Analysis of our price model
- Performance Evaluation
- Conclusions

# Motivation

- Cloud Computing
  - Integrates computation storage and network resources in a virtualized environment
  - CPU and memory storage are guaranteed and charged in current IaaS
  - Bandwidth is an essential resource in Clouds, but not guaranteed in current cloud computing platforms

- Problems
  - Unpredictable network performance
  - Require us to develop explicitly allocation policies of network resources in cloud computing

CLEMSON UNIVERSITY

# Introduction

- Three Requirements (FairCloud [Sigcomm'12])
- Minimum Guarantee
  - Essential for predictable network performance
- High Utilization
  - Maximizing network utilization in the presence of unsatisfied demands
- Network Proportionality
  - Allocated network resources are proportional to tenants' payments
- Tradeoffs
  - Tradeoff between these requirements
  - Hard to meet all the three requirements simultaneously

# Introduction(cont.)

- Utility-the difference between the total cost and total revenue
- Utility of cloud provider

$$U_c = \sum_i p_i N_{v_i} - bB^a - \sum_{t_i} F_c(H_{t_i}) - cK$$

- Utility of tenant $t_i$

$$U_{t_i} = g_{t_i} B^a_{t_i} - \sum_i p_i N_{v_i} - F_{t_i}(H_{t_i})$$

where $g_{t_i}$ is the earned utility of each allocated bandwidth unit

| | |
|---|---|
| $N_{v_i}$ | Number of sold VMs, $p_i$ is the price of VM i |
| $B^a$ | The allocated bandwidth to all tenants, $b$ unit bandwidth price is charged by tier-1 ISPs |
| $M_{v_i}$ | The min-guarantee bandwidth for VM $v_i$ |
| $H_{t_i}$ | The unsatisfied bandwidth for tenant $t_i$ |
| $F_c(H_{t_i})$ | Utility loss of the cloud provider due to reputation degradation and potential revenue loss |
| $F_{t_i}(H_{t_i})$ | Utility loss of tenant $t_i$ due to unsatisfied bandwidth from cloud provider |
| $K$ | $cK$ is the reservation cost of reserving bandwidth $K$ |

# New Bandwidth Price Model

- Develop a policy that can flexibly meet the three requirements and achieve the ultimate goal
- Ultimate Goal
  - To increase the benefits of the cloud provider and tenants, that is, to increase the utility of both cloud provider and tenants
- Transforms the competitive environment to a cooperative environment
  - Competitive environment: tenants compete for bandwidth allocation
  - Cooperative environment: tenants cooperate to share bandwidth
- Notations
  - $D_l$ total bandwidth demand from VMs on link $l$
  - $C_l$ bandwidth capacity of link $l$
  - $v_j$ VM $j$

# Congested and Uncongested Links

- Congested links
  - Defined only when $D_l > C_l$ rather than $D_l \geq C_l$
  - Because when $D_l = C_l$, the link can exactly satisfy each VM's demand
  - When $D_l = C_l$, the link is fully utilized, which satisfies the expectation of the cloud provider
  - $D_l = C_l$, ideal situation for both cloud provider and tenants
- Obviously, we expect to achieve $D_l = C_l$ on the link when using our price model.
- Uncongested links
  - $D_l < C_l$
  - $D_l = C_l$

# Payment of Tenants

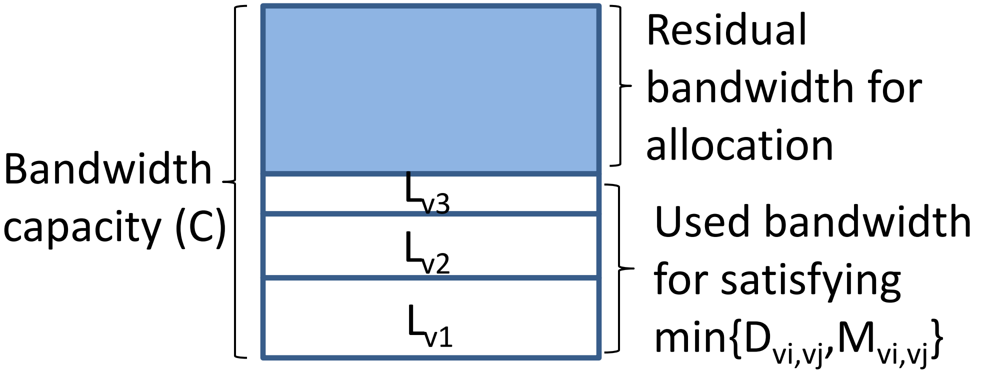
- A tenant's payment consists of three parts of bandwidth
  - $M_{t_i}$ the sum of minimum bandwidth guarantee of tenant $t_i$'s VMs
  - $B_{t_i}^{c}$ consumed bandwidth on congested links of tenant $t_i$
  - $B_{t_i}^{u}$ consumed bandwidth on uncongested links of tenant $t_i$
  - $P_{t_i}$ the total payment of tenant $t_i$
- $P_{t_i} = \alpha M_{t_i} + \beta B_{t_i}^{c} + \gamma B_{t_i}^{u}$
  - $\alpha$, $\beta$, $\gamma$ is the unit price of minimum guaranteed bandwidth, congested bandwidth and uncongested bandwidth, respectively
- We specify $\alpha > \beta > \gamma$

# Relationship between α, β, and γ

- $\alpha > \beta, \gamma$
  - Reserved bandwidth is more valuable than non-reserved bandwidth
  - Each tenant would try to buy more minimum bandwidth
    - More reserved but unused and wasteful bandwidth
    - Reduce cloud provider's profit
    - Poor network performance because of the competition for minimum bandwidth between tenants
  - Reserve no more bandwidth than tenants actually need
- $\beta > \gamma$
  - Tenants are incentivized to use uncongested links
  - Tenants are not encouraged to use congested links
  - Ideal situation $D_l = C_l$ (Network is fully utilized and bandwidth demands are satisfied) is defined as uncongested link
  - Specifically, $\beta = \gamma \frac{D_l}{C_l}$, the more congested, the more payment

# Network Bandwidth Sharing Policy

- Consider both min-guarantee and proportionality
- When $D_l \leq C_l$, the link is uncongested
  - Each VM achieves allocation exactly equal to what they request
- When $D_l > C_l$, the link is congested
  - Each VM first receives its min-guarantee
  - Receives its share on the residual bandwidth based on the weight of VM (proportional allocation policy)



# Proportional Allocation Policy

- Network proportionality, congestion proportionality, and link proportionality(FairCloud[Sigcomm'12])
- Weights of VMs are determined by the min-guarantees of VMs
- E.g. link proportionality
  - Assume two VMs communicating with each other
  - The weight of this pair of VMs on one link would be:

$$W_{v_i, v_j} = M_{v_i} \frac{M_{v_j}}{\sum M_{v_{i,k}}} + M_{v_j} \frac{M_{v_i}}{\sum M_{j,k}}$$

where $\sum M_{v_{i,k}}$ is the sum of the min-guarantees of all VMs that $v_i$ communicates with through this link

# Analysis of Our Price Model

- Utility of the cloud provider

$$U_c = \sum_i p_i N_{v_i} - bB^a - \sum_{t_i} F_c(H_{t_i}) - cK$$

$$= \sum_{t_i} \{((\alpha - c)M_{t_i} + \beta B_{t_i}^c + \gamma B_{t_i}^u) - b(B_{t_i}^c + B_{t_i}^u) - F_c(H_{t_i})\}$$

  - Reduce unsatisfied demand rate

- Utility of the tenants

$$U_{t_i} = g_{t_i} B_{t_i}^a - \sum_i p_i N_{v_i} - F_{t_i}(H_{t_i}) = g_{t_i} B_{t_i}^a - (\alpha M_{t_i} + \beta B_{t_i}^c + \gamma B_{t_i}^u) - F_{t_i}(H_{t_i})$$

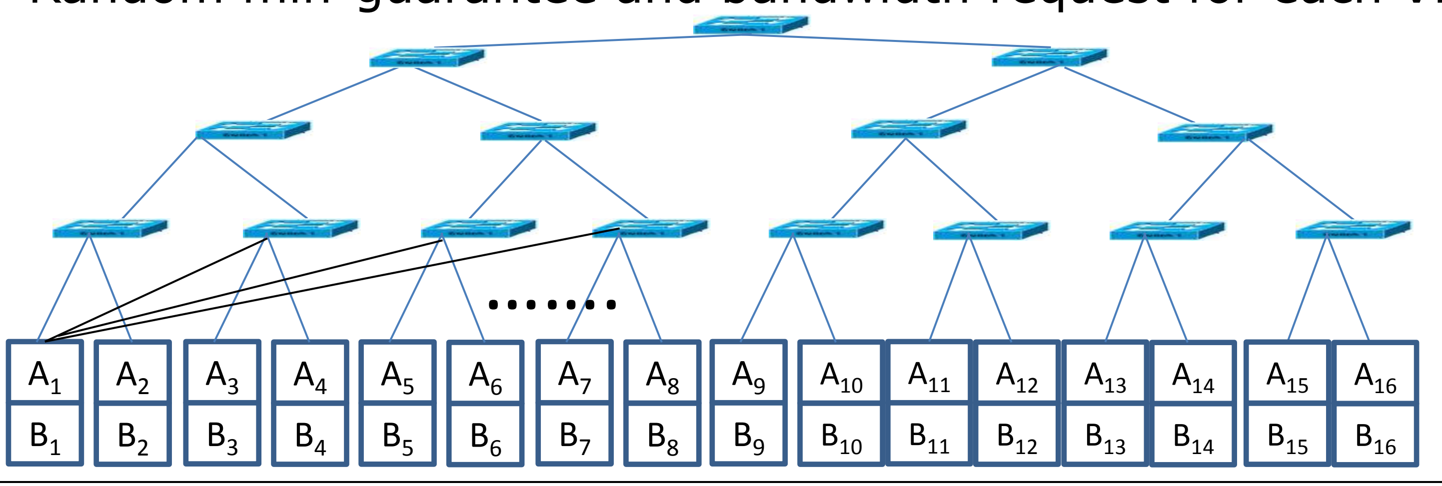
  - Reduce payment
  - Reduce unsatisfied demand rate
  - Receive more allocation → higher cost

# Volunteering

- $\beta B_{t_i}^{c} + \gamma B_{t_i}^{u}$
  - Encourage to use uncongested links
  - $\beta \propto \frac{D_l}{C_l}$ →reduce the total demand on the links →encourage to use less congested links
  - Congested link is changed to uncongested link→the unit price β is changed to γ →volunteer to reduce VMs' unimportant demands to make link uncongested

# Performance Evaluation

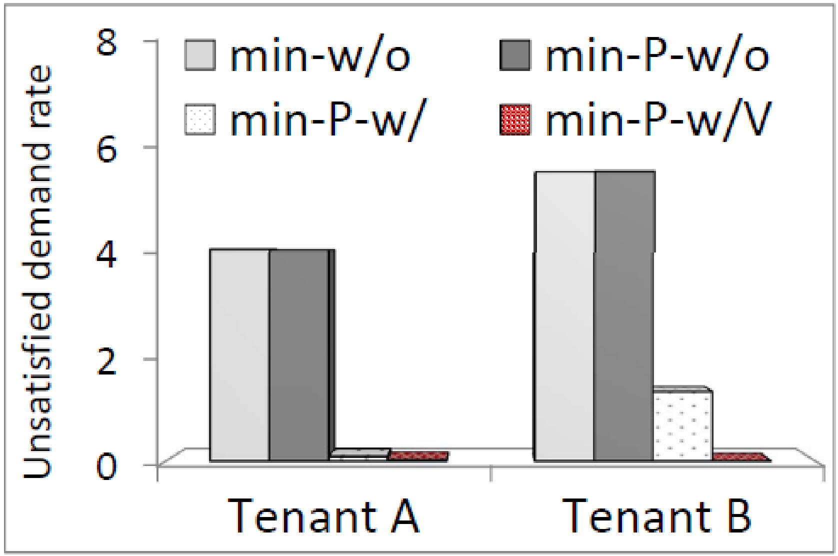
- Simulation and trace-driven experiments
  - A tree topology with 16 servers and two tenants A and B
  - Each server has one VM from each tenant
  - Each server connects to four switches
  - $A_i \leftrightarrow A_{i+8}, i = 1,2,\dots,8$ and $B_i \leftrightarrow B_j, i \neq j$
  - $\alpha = 1,\ \gamma = 0.3,\ \beta = \frac{D_l}{C_l}\gamma$
  - Random min-guarantee and bandwidth request for each VM

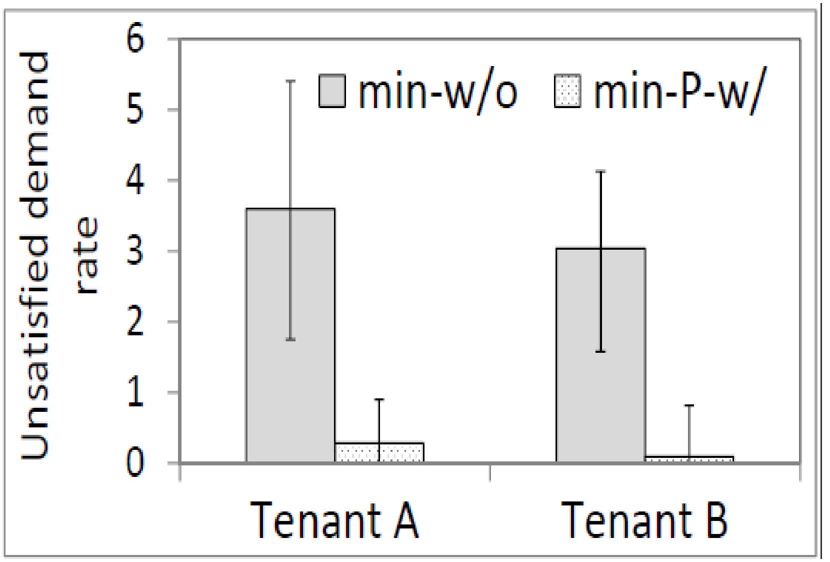


# Comparing Methods

| | |
|---|---|
| min-w/o | Proportional allocation policy (PS-P, FairCloud [Sigcomm'12]) |
| min-P-w/o | First min-guarantee and proportionality allocation policy (i.e., our allocation policy) **without** using the least congested links price model. |
| min-P-w/ | First min-guarantee and proportionality allocation policy (i.e., our allocation policy) **with** using the least congested links price model. |
| min-P-w/V | min-P-w/ in which tenants further are incentivized to volunteer to reduce unimportant demands. |

# Unsatisfied Demand Rate



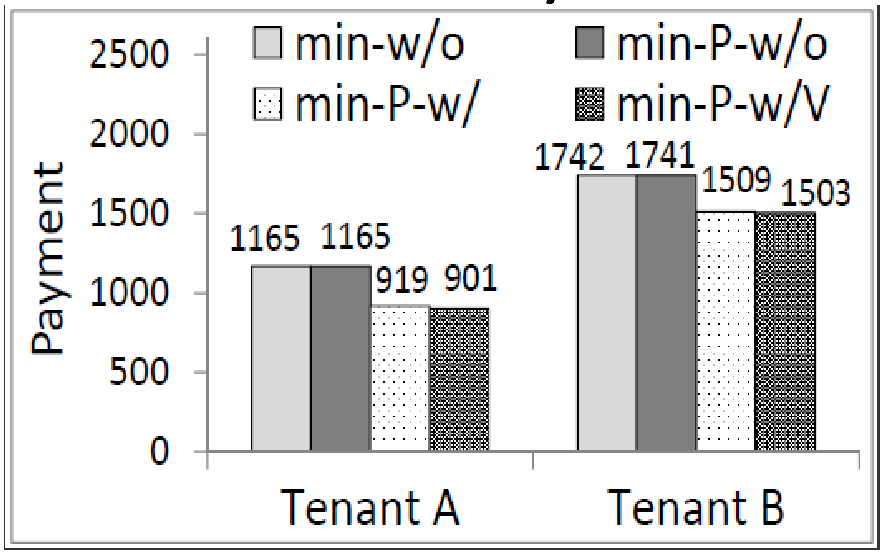

Simulation

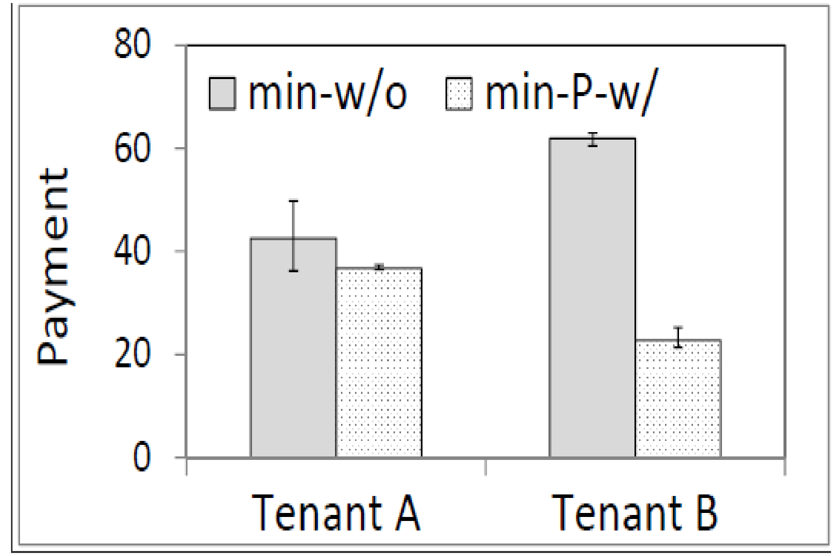
Conclusion:
min-P-w/V<min-P-w/<min-P-w/o< min-w/o

Trace-driven experiment

Conclusion:
min-P-w/<min-w/o

# Payment of Each Tenant




Simulation

Conclusion:
min-P-w/V<min-P-w/<min-P-w/o<
min-w/o

Trace-driven experiment

Conclusion:
min-P-w/<min-w/o

Reason: encourage to use less congested links
and uncongested links

# Conclusions

- New Bandwidth Pricing Model
  - Incentivizes tenants to use uncongested links and constrain congestion
  - Encourages tenants to increase their utility
  - Increases network utilization and reduces unfulfilled bandwidth demands
- Future Work
  - Consider rewarding tenants for reducing demand to maintain the uncongested link states

*Thank you!*

*Questions & Comments?*


**Zhuozhao Li**

**zhuozhl@clemson.edu**

**Ph.D. Candidate**

**Pervasive Communication Laboratory**

**Clemson University**