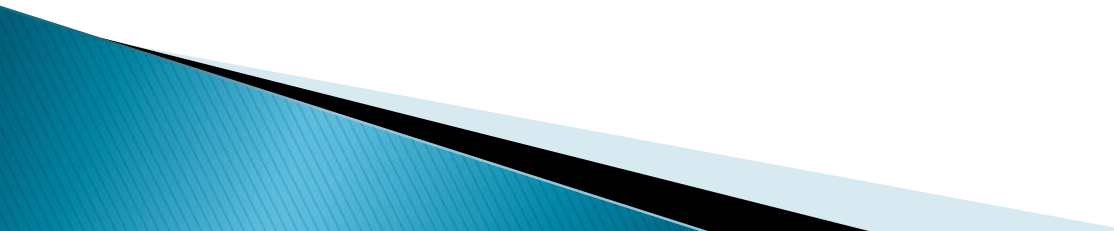


Kingfisher: Cost-aware Elasticity in the Cloud

Presented by
Zhaoyi Luo
CS 755
University of Waterloo

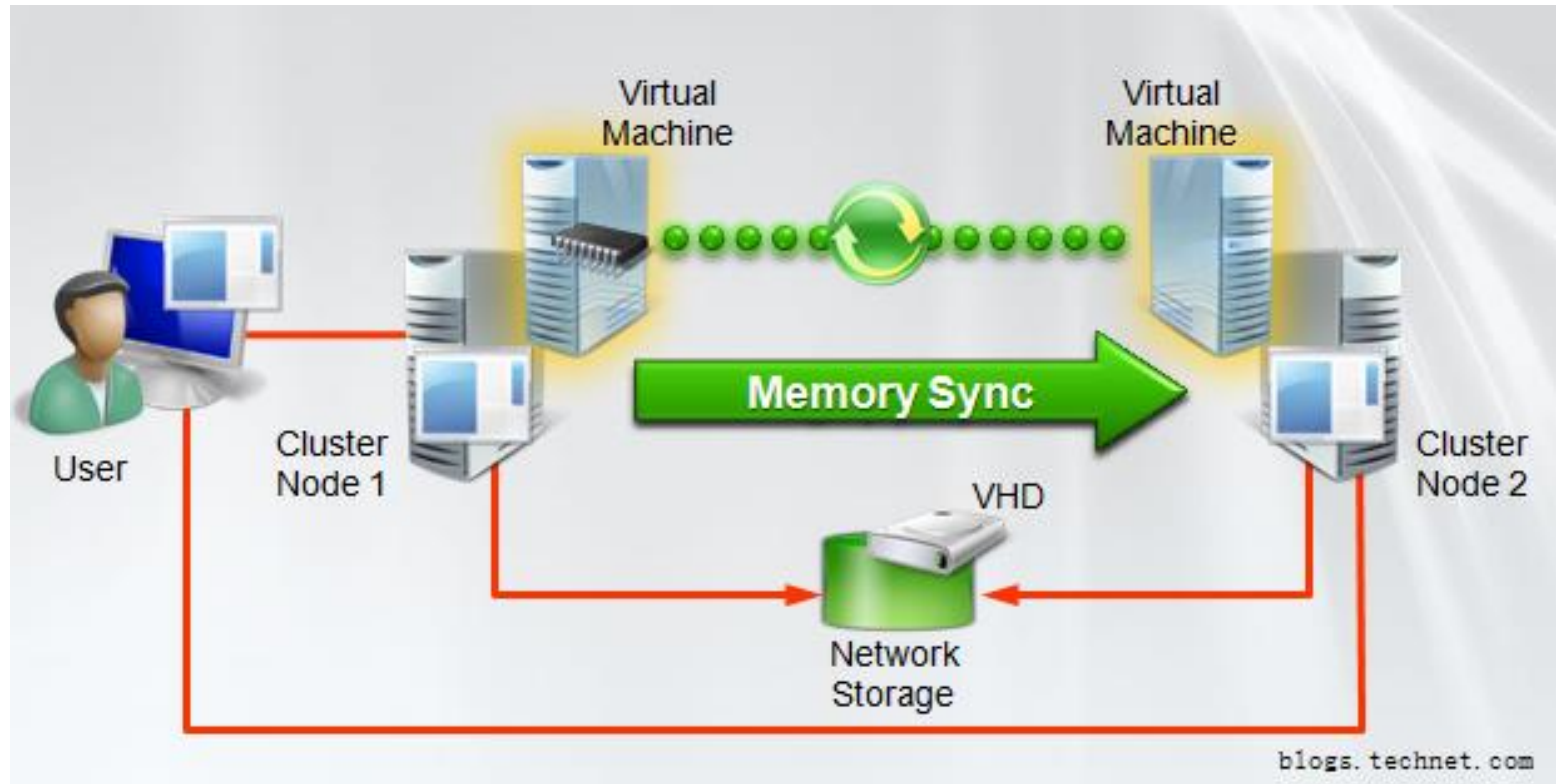
Outline

- ▶ Background
 - Cloud computing & Virtualization
 - Replication & Migration
 - ▶ Kingfisher Infrastructure Overview
 - ▶ Cost-aware Elasticity Algorithms
 - Rental Cost-aware Provisioning
 - Transition Cost-aware Provisioning
 - ▶ Experimental Study
 - ▶ Conclusion & Questions
- 

Background

- ▶ Cloud computing & Virtualization
 - Servers are deployed in cloud as virtual machines
 - Flavor of VMs (small, medium, large)
 - The rental cost of virtual machine is not linear
- ▶ Replication & Migration
 - Replication. e.g. “small” → double “small”
 - Local resizing. e.g. “small” → “large” on the same machine.
 - Live migration: e.g. “small” → “large” to another machine.

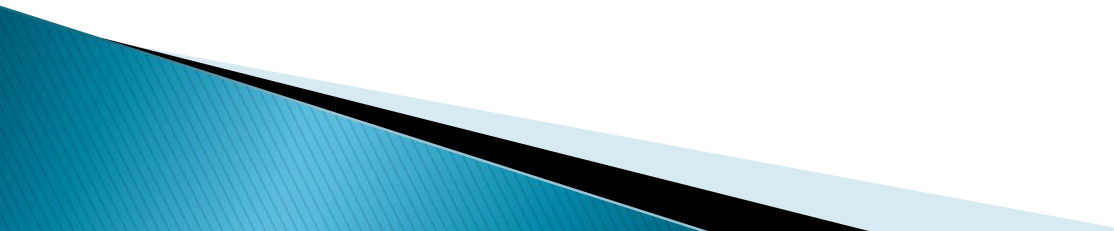
Live migration



Goal: Cost-aware Provisioning

- ▶ Rental Cost-aware Provisioning
 - Given the estimated peak workload
 - What type and how many servers needed
- ▶ Transition Cost-aware Provisioning
 - Using replication, migration and resizing
 - What mechanism to use to reduce transition latency

Outline

- ▶ Background
 - Cloud computing & Virtualization
 - Replication & Migration
 - ▶ Kingfisher Infrastructure Overview
 - ▶ Cost-aware Elasticity Algorithms
 - Rental Cost-aware Provisioning
 - Transition Cost-aware Provisioning
 - ▶ Experimental Study
 - ▶ Conclusion & Questions
- 

Infrastructure Overview

- ▶ Cost-aware elasticity algorithms are implemented in Capacity Planner and Orchestration Engine

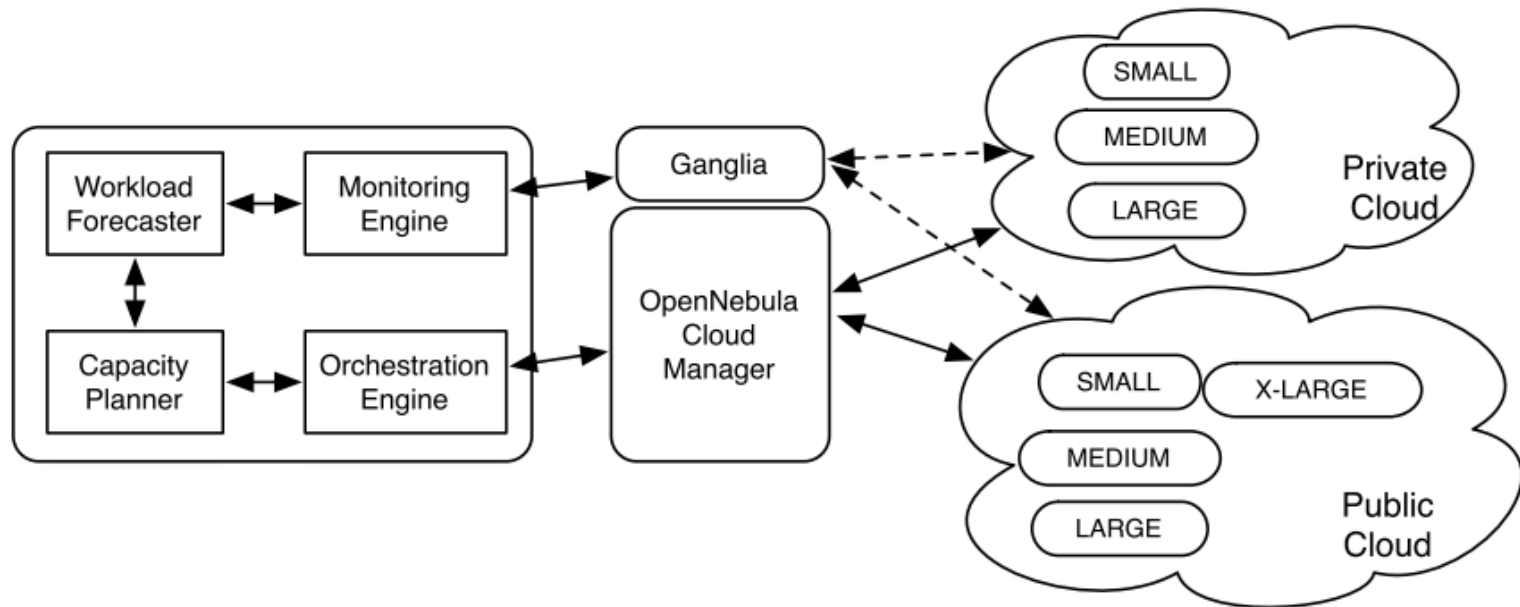


Fig. 1: Kingfisher Architectural Overview

Cost-aware Elasticity Algorithms

- ▶ Rental Cost-aware Provisioning
 - Step 1: Determine Server Capacities C_j
 - Step 2: Determine Server Configuration

$$\text{minimize } \sum_{j=1}^M n_j p_j \quad \text{s.t.} \quad \sum_{j=1}^M n_j C_j > \lambda$$

Where λ is the amount of user-required resources (request rate)

Cost-aware Elasticity Algorithms

- ▶ Transition Cost-aware Provisioning
 - Local resizing (minimal overhead)
 - Replication ($D/r + b$)
 - Live migration ($\omega R/r$)
 - Shutdown migration ($2D/r + b$)

Cost-aware Elasticity Algorithms

- ▶ Integer Linear Program (ILP) for rental/transition cost-aware problems

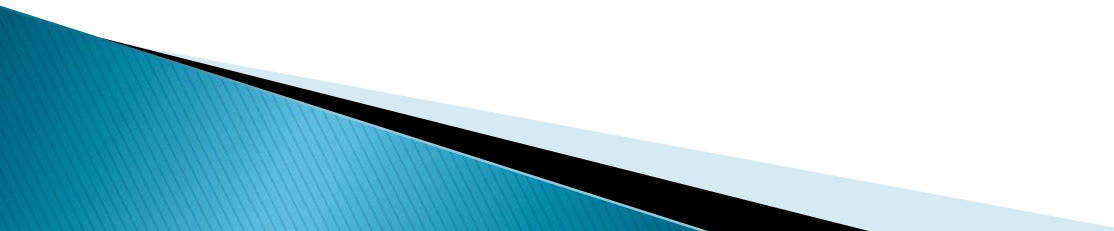
$$\min \sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^T p_{(j)} x_{ijk} \quad (1)$$

subject to the constraints

$$\sum_{i=1}^N \sum_{j=1}^M \sum_{k=1}^T x_{ijk} C_j \geq \lambda; \quad \sum_{j=1}^M \sum_{k=1}^T x_{ijk} = 1, \forall i \quad (2)$$

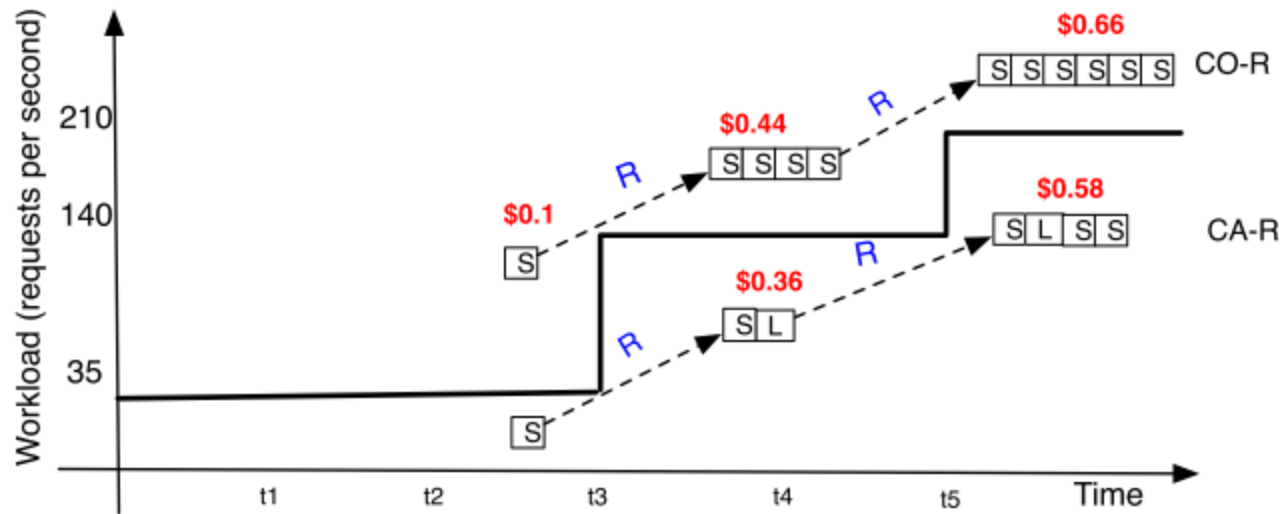
- ▶ 2-approximate greedy heuristic
 - sorts $x_{i,j,k}$ in increasing order of p_j/C_j and then finds the smallest list of $x_{i,j,k}$ s which satisfy Eq. (2).

Outline

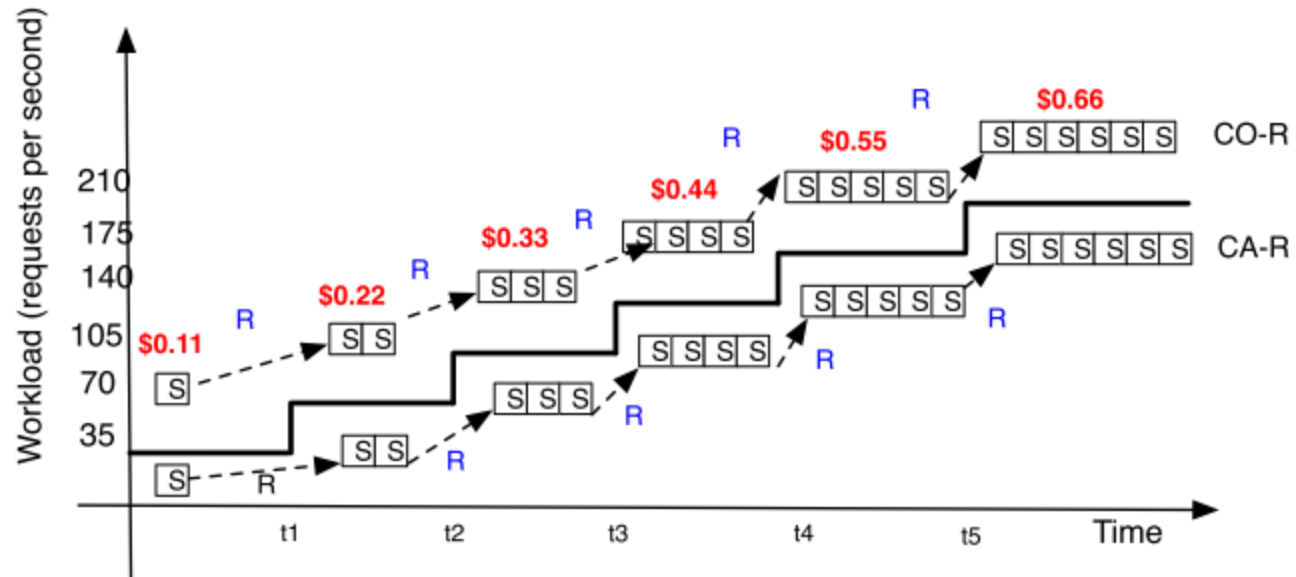
- ▶ Background
 - Cloud computing & Virtualization
 - Replication & Migration
 - ▶ Kingfisher Infrastructure Overview
 - ▶ Cost-aware Elasticity Algorithms
 - Rental Cost-aware Provisioning
 - Transition Cost-aware Provisioning
 - ▶ **Experimental Study**
 - ▶ Conclusion & Questions
- 

Experimental Study

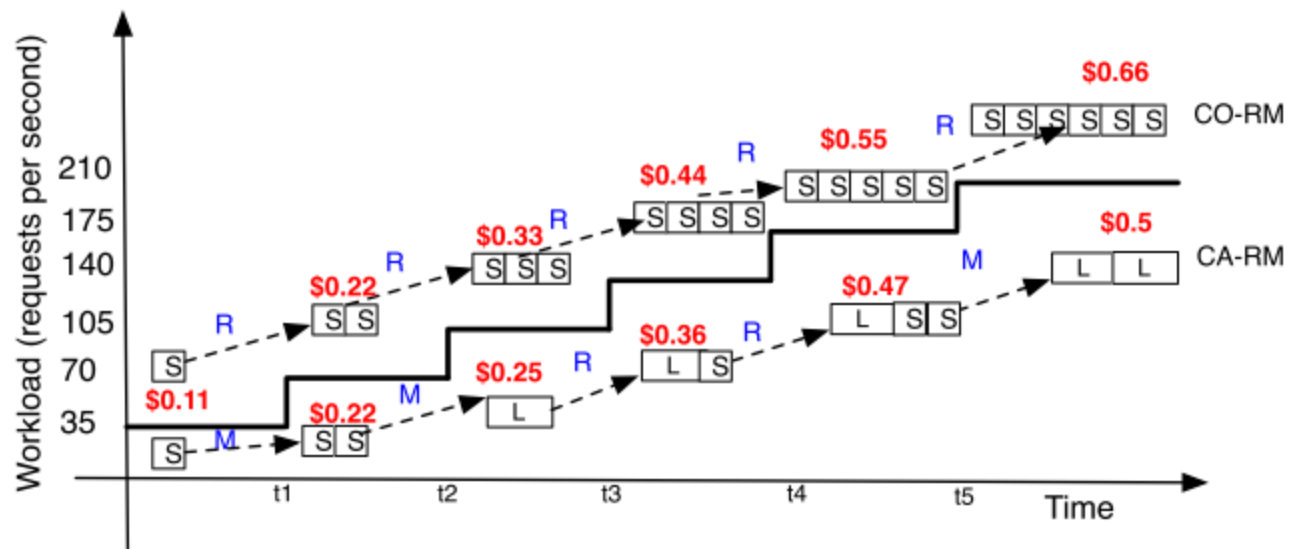
- ▶ Rental cost-aware study
 - Without migration
 - CO-R (cost-oblivious)
 - CA-R (cost-aware)
 - With migration
 - CO-RM
 - CA-RM
- ▶ Transition cost-aware study
 - TO-RM
 - TA-RM
- ▶ Small/large-jump workload



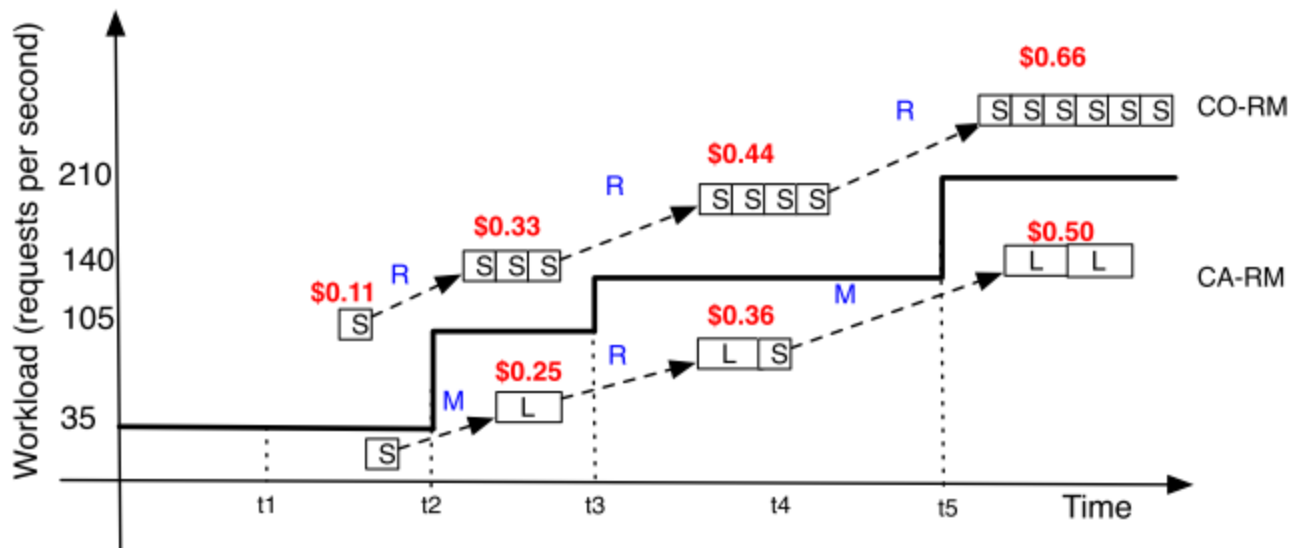
(a) *large-jump* workload



(b) *small-jump* workload



(a) small workload increase



(b) large workload jumps

Experimental Study

► Performance Study

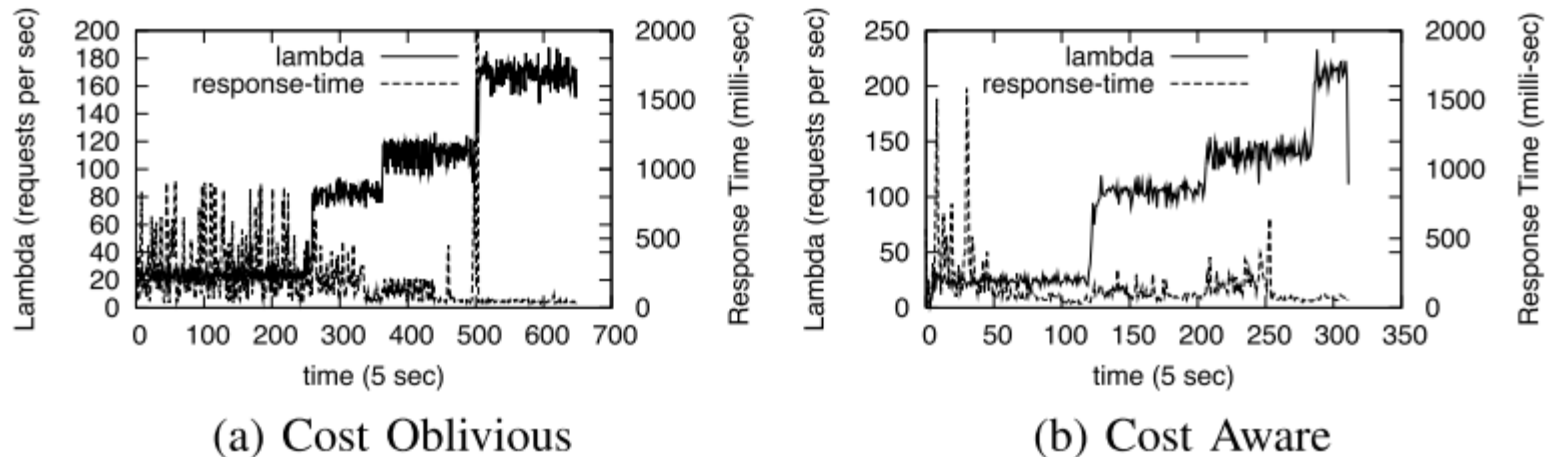


Fig. 5: Application Performance during cost-aware and cost-oblivious provisioning for *large-jump* workload.

- Live migration is preferred whenever feasible

Experimental Study

► Transition cost-aware study

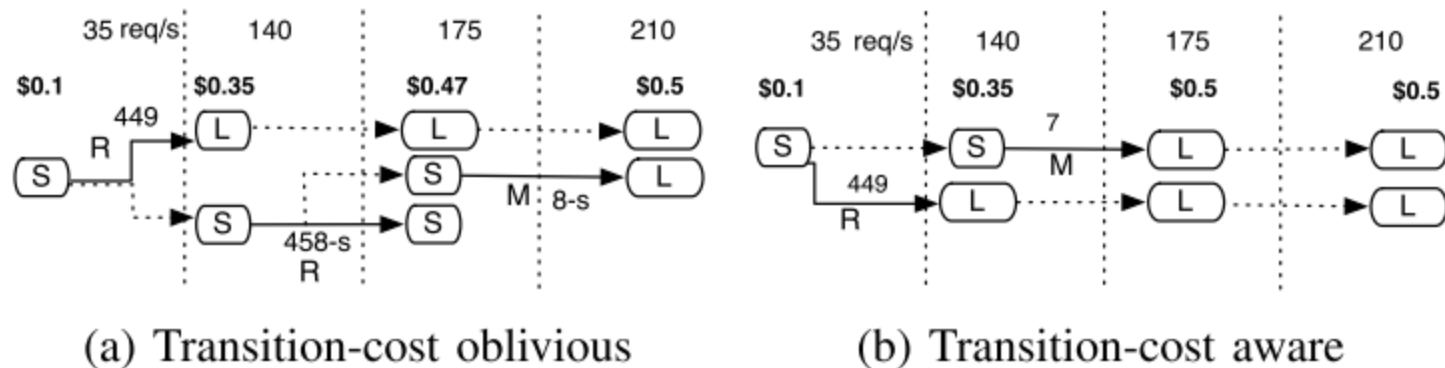
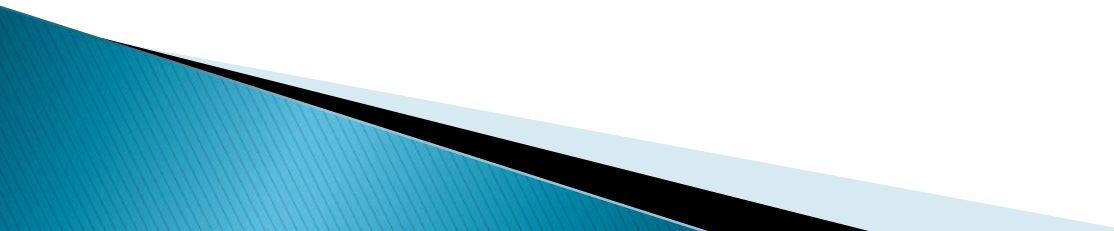


Fig. 6: Comparison of a transition-cost aware system with a transition-cost oblivious system.

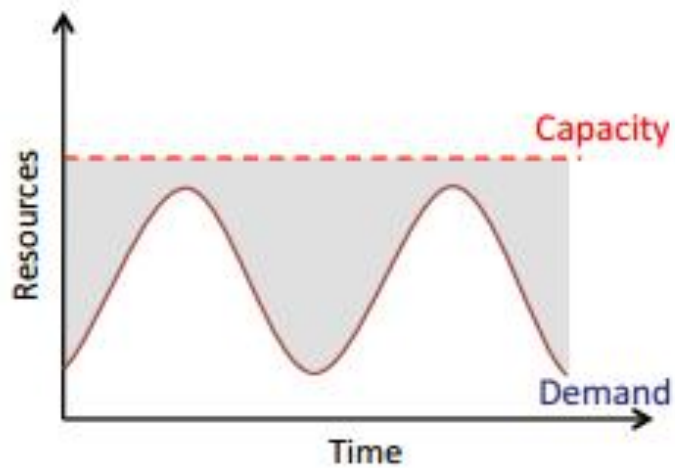
Conclusion

- ▶ Optimize the selection of virtual server to minimize **rental cost**
 - ▶ Leverage multiple mechanisms to minimize **transition cost**
 - ▶ Formalize the problem as **Integer Linear Program** and solve it by greedy-like heuristic
 - ▶ Experiments are conducted to show efficacy
- 

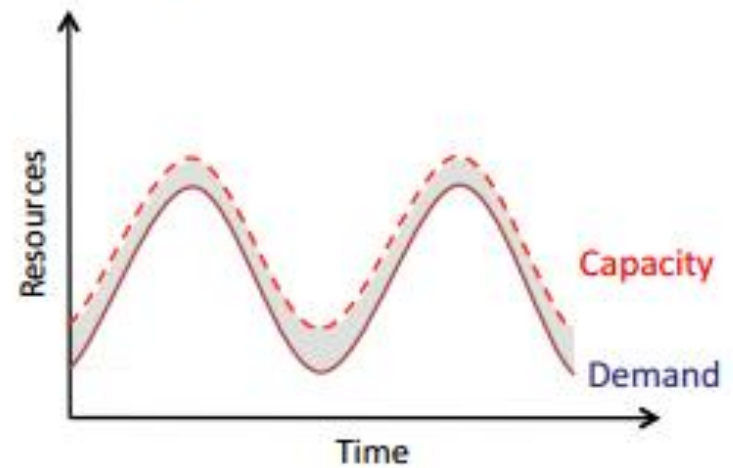
Questions?

- ▶ Is the scenario based on a IaaS, PaaS or SaaS cloud?
 - ▶ What truth/what assumptions they made might undermine their work?
 - Perfect forecaster.
 - Cost estimation of live migration.
 - Integer Linear Program. 2-approximate heuristic.
 - ▶ How to simulate “cost-oblivious” in experiment?
- 

Static data center



Data center in the cloud



 Unused resources