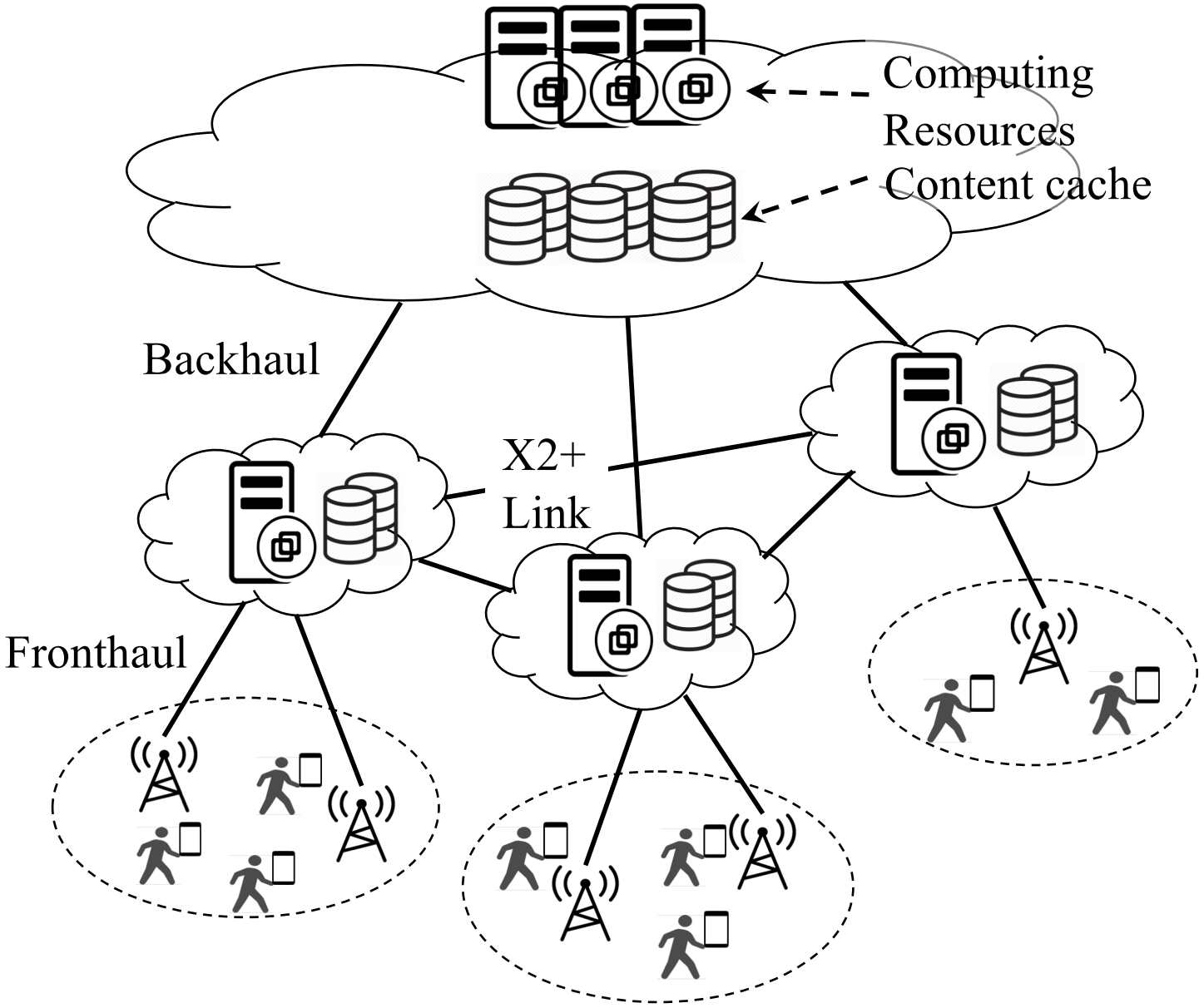
Joint Optimization of Data Caching and Processing for Mobile Edge Computing

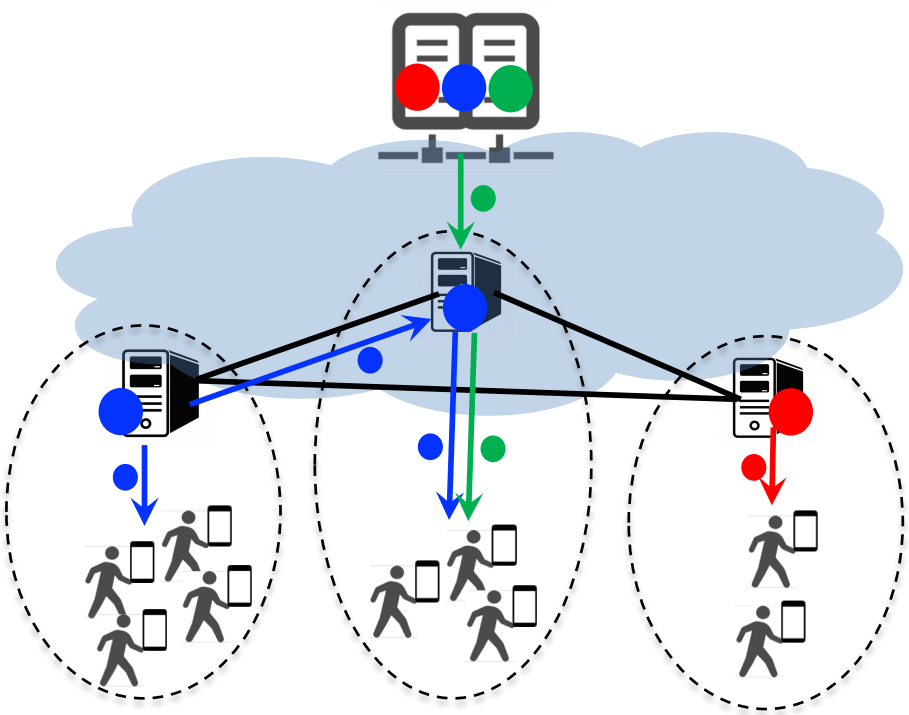
Xun Shao
Kitami Institute of Technology
2019-3-6

14th IEICE ICN Workshop

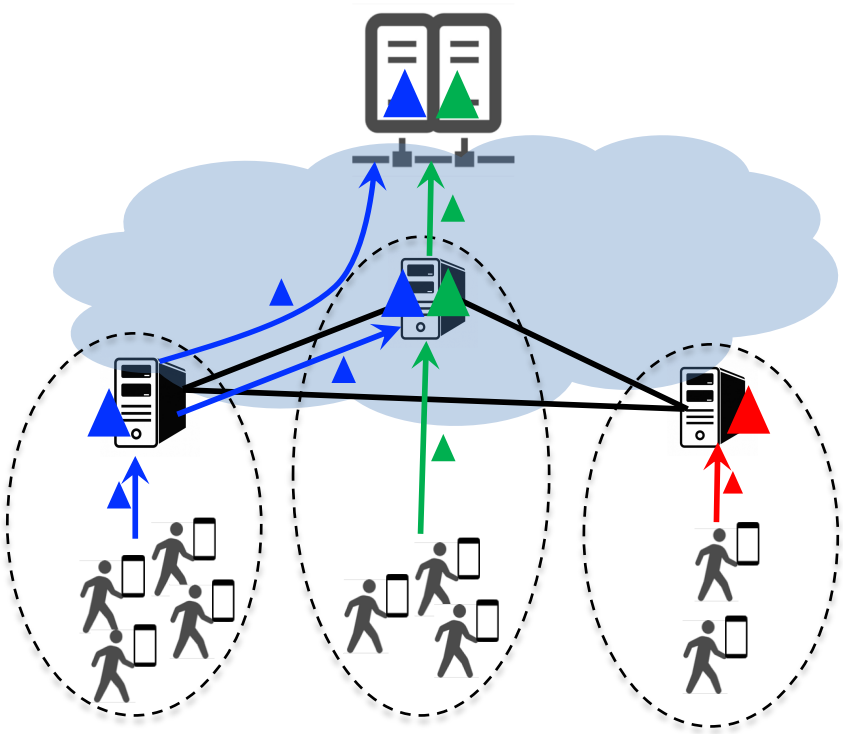
Mobile edge computing architecture



Current mobile edge services: either data caching or task offloading oriented

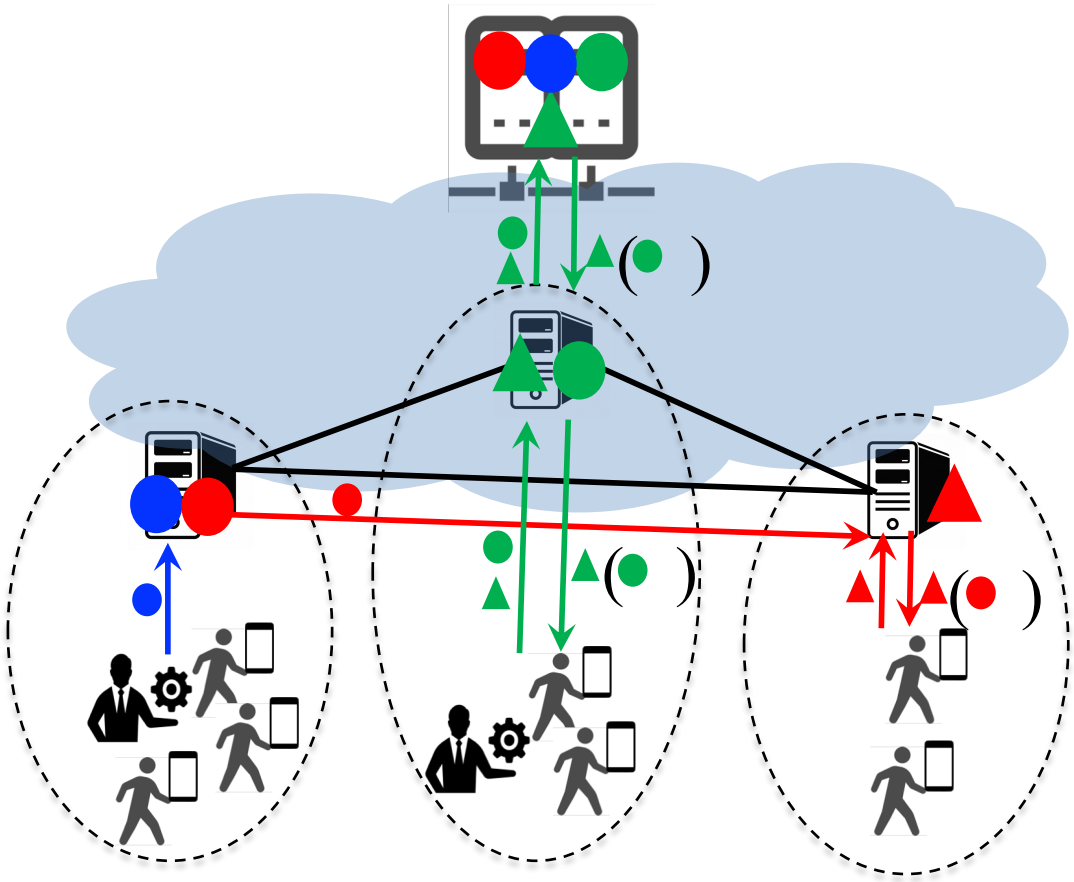


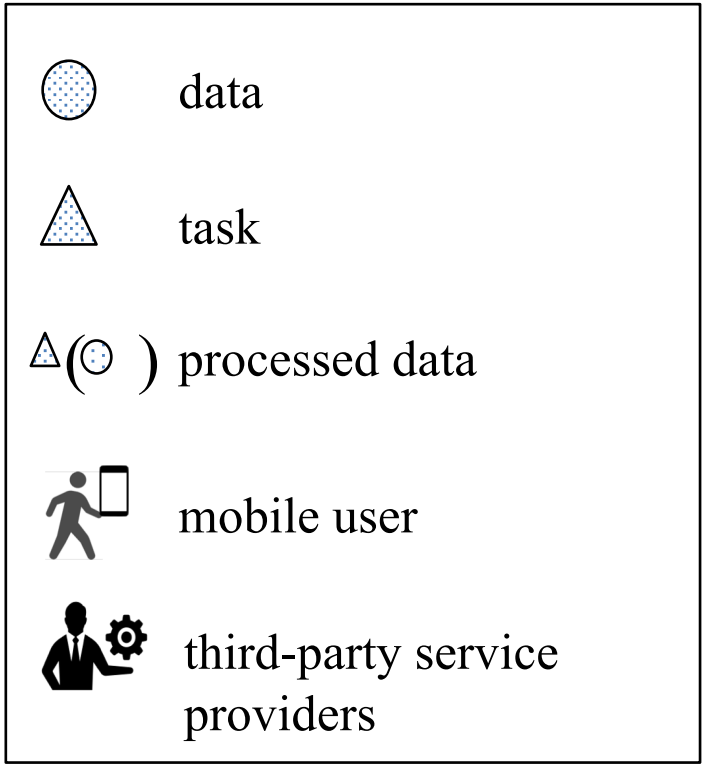
(a) Data caching oriented



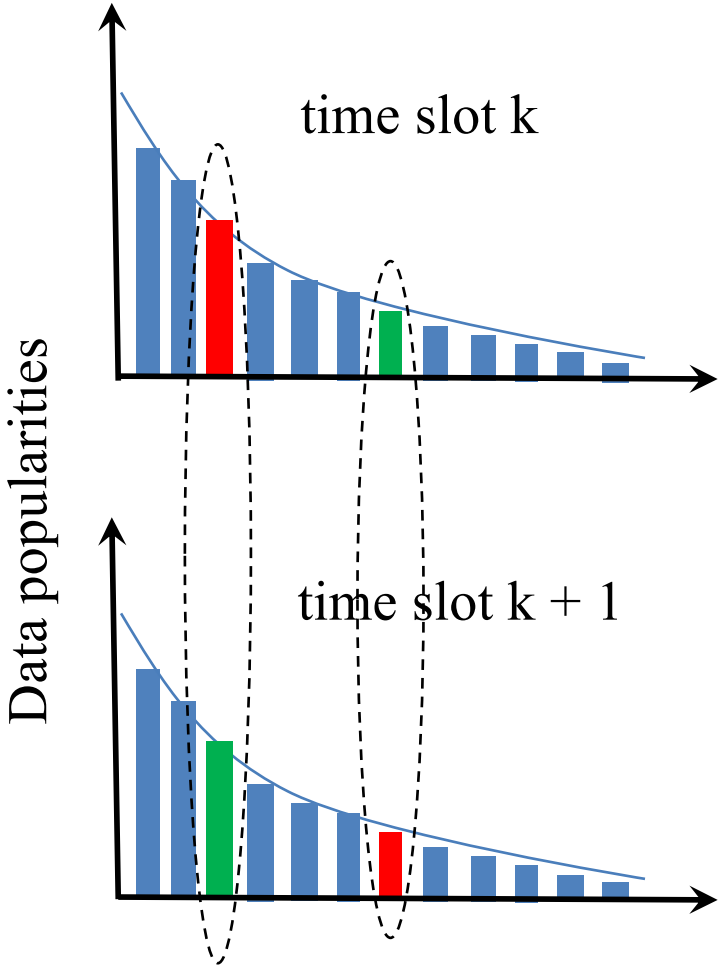
(b) Task offloading oriented

Our proposal: joint data caching and processing system

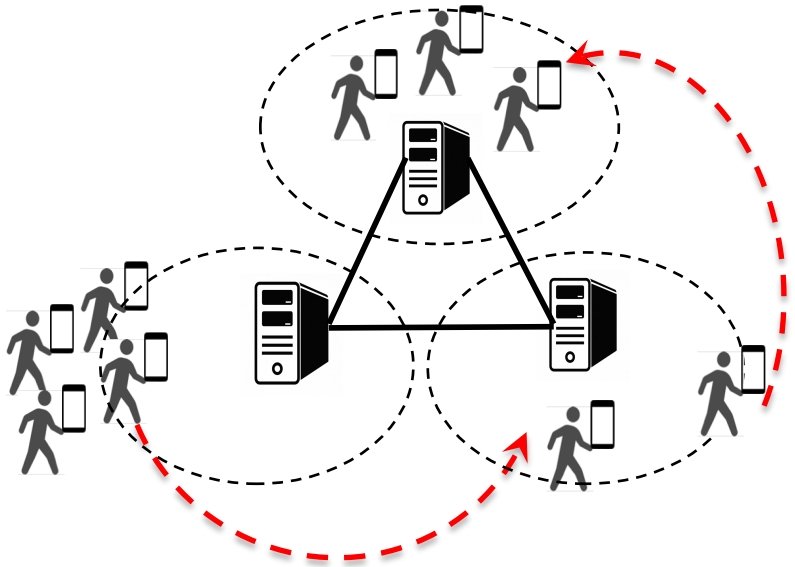




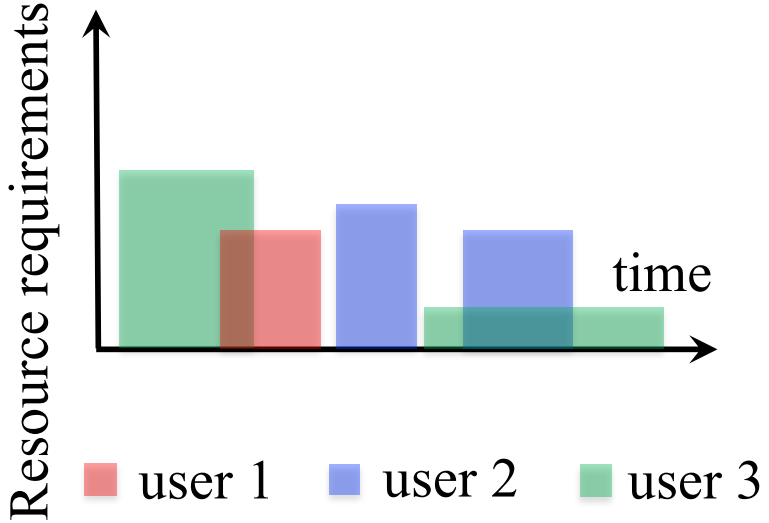
Main challenges: dynamics



(a) Data popularity dynamics

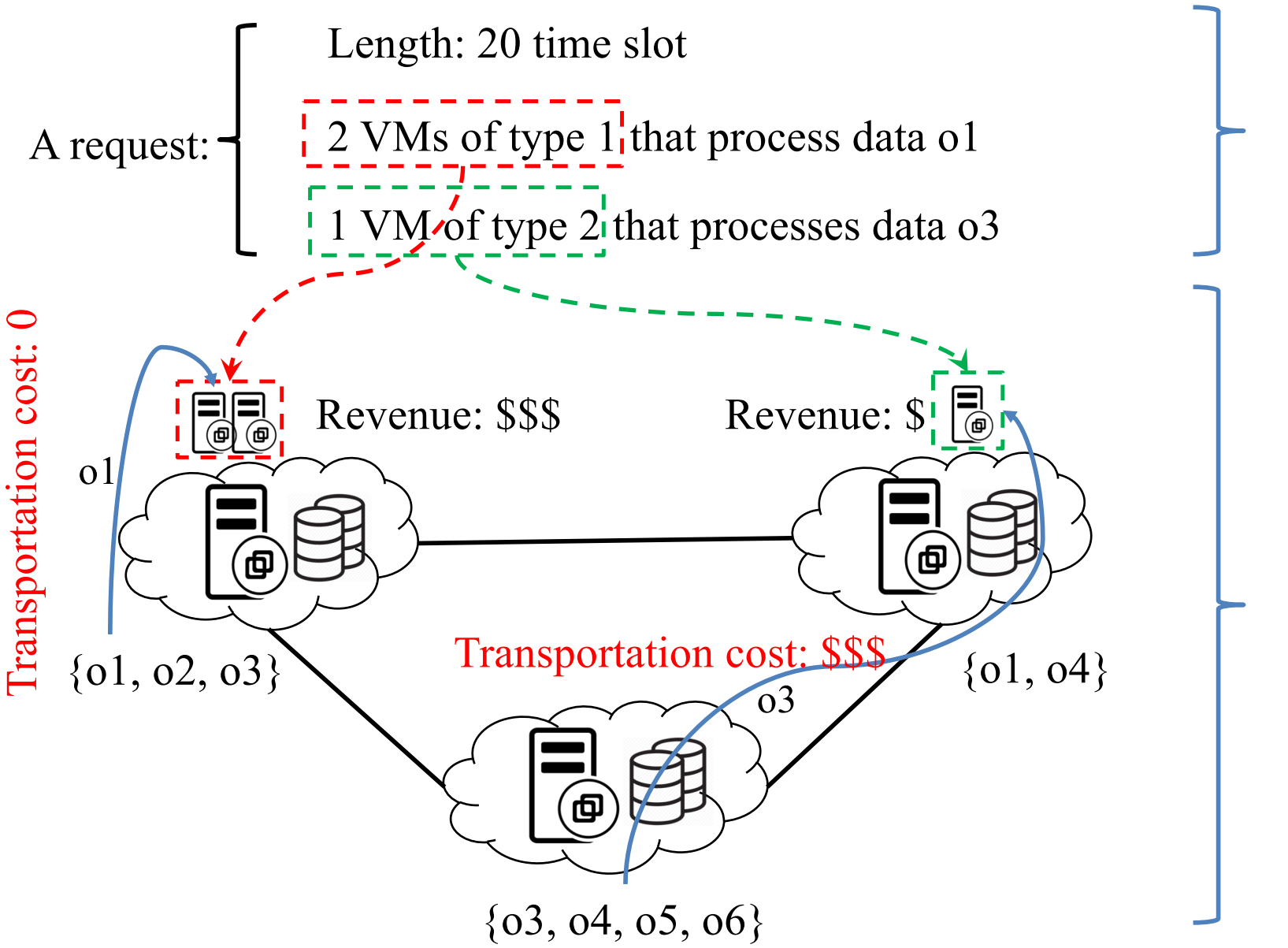


(b) User distribution dynamics

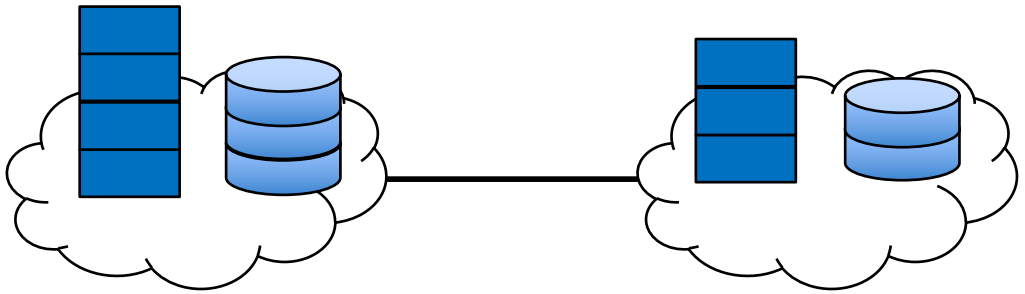


(c) Request dynamics

A detailed scenario



Making online decisions (1/9)



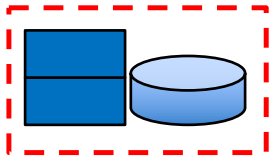
Cached data: {o1, o2, o3}

Cached data:{o1, o4}

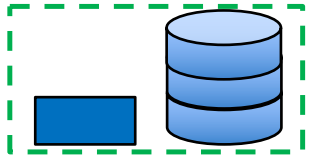
request k:

- 3 time slots
- 2 VMs of type 1
- process data o3

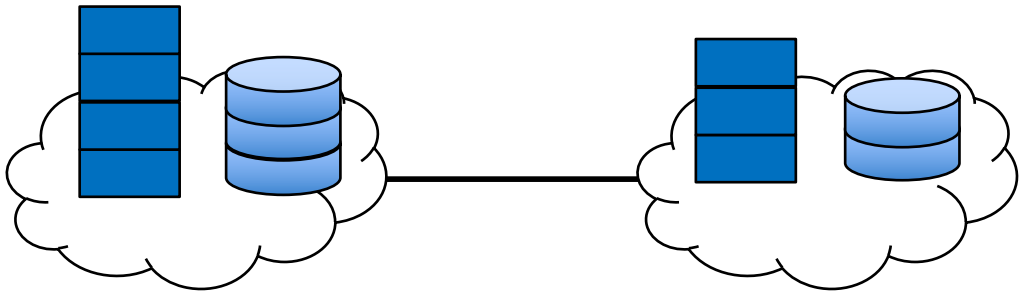
VM of type 1:
$3/time slot



VM of type 2:
$5/time slot



Making online decisions (2/9)



Cached data: {o1, o2, o3}

Cached data:{o1, o4}

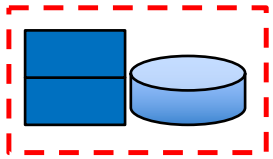
request k:
- 3 time slots
- 2 VMs of type 1
- process data o3

Choice 1: Assemble all VMs in location 1

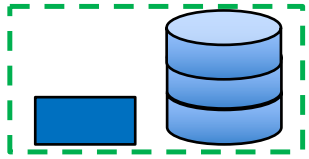
Choice 2: Assemble one VM in each location

Choice 3: Reject the request (redirect it to the cloud)

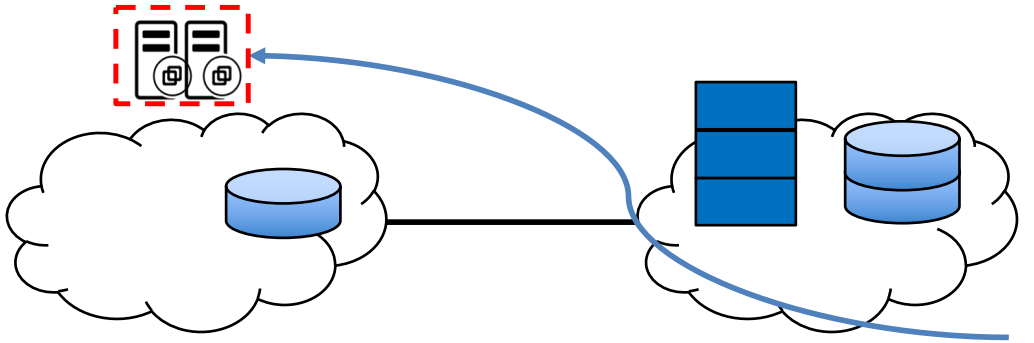
VM of type 1:
$3/time slot



VM of type 2:
$5/time slot



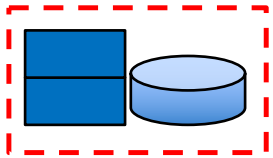
Making online decisions (3/9)



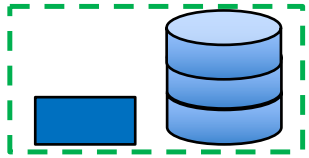
request k:
- 3 time slots
- 2 VMs of type 1
- process data o3

Assemble all VMs in location 1
- revenue: 3 * 2 *3
- transportation cost: cost(o4)

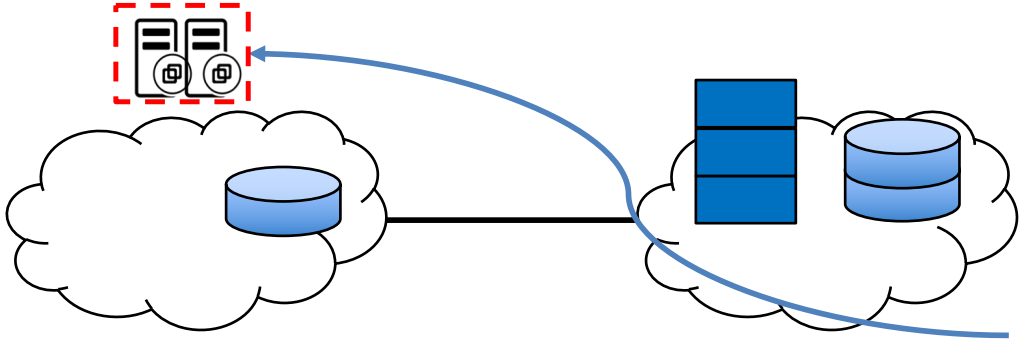
VM of type 1:
$3/time slot



VM of type 2:
$5/time slot



Making online decisions (4/9)



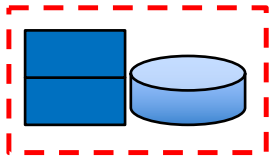
request k:
- 3 time slots
- 2 VMs of type 1
- process data o3

Assemble all VMs in location 1
- revenue: 3 * 2 *3
- transportation cost: cost(o4)

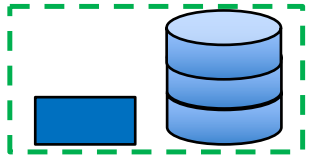
request k + 1:
- 3 time slots
- 1 VMs of type 2
- process data o3

Only to reject
- revenue: 0
- transportation cost: 0

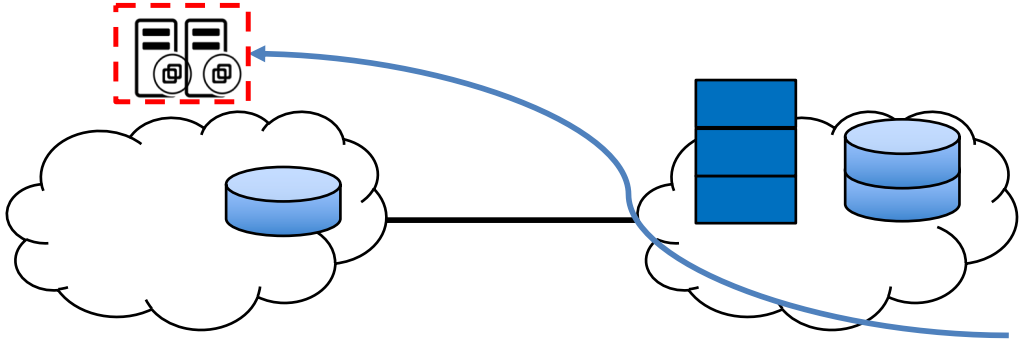
VM of type 1:
$3/time slot



VM of type 2:
$5/time slot



Making online decisions (5/9)



request k:
- 3 time slots
- 2 VMs of type 1
- process data o3

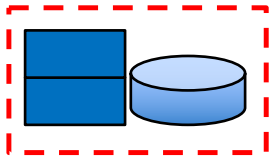
Assemble all VMs in location 1
- revenue: 3 * 2
- transportation cost: cost(o4)

request k + 1:
- 3 time slots
- 1 VMs of type 2
- process data o3

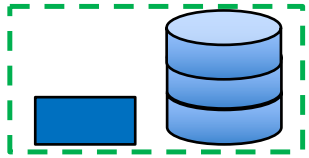
Only to reject
- revenue: 0
- transportation cost: 0

Total benefit: 3*2*3 – cost(o4)

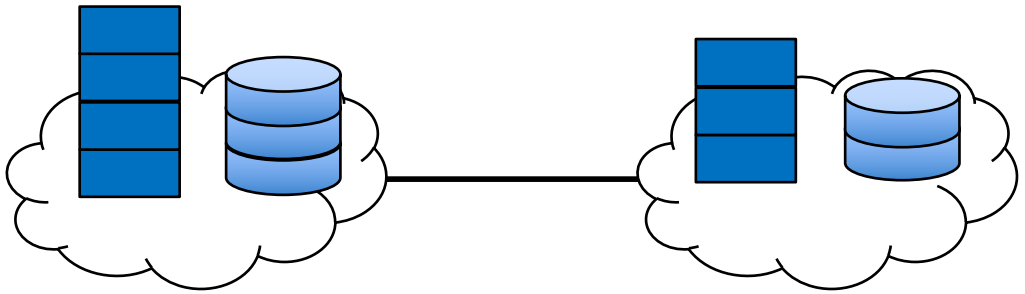
VM of type 1:
$3/time slot



VM of type 2:
$5/time slot



Making online decisions (6/9)



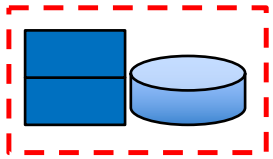
Cached data: {o1, o2, o3}

Cached data:{o1, o4}

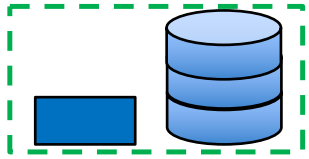
request k:
- 3 time slots
- 2 VMs of type 1
- process data o3

Reject
- revenue: 0
- transportation cost: 0

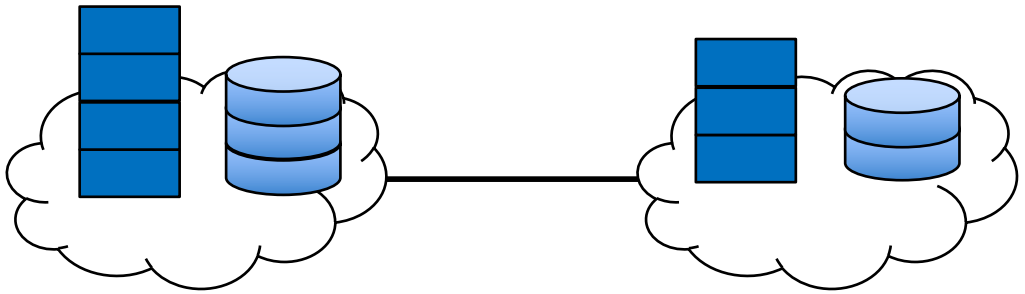
VM of type 1:
$3/time slot



VM of type 2:
$5/time slot



Making online decisions (7/9)



Cached data: {o1, o2, o3}

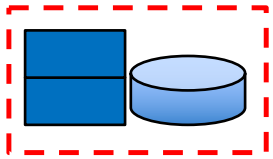
Cached data:{o1, o4}

request k:
- 3 time slots
- 2 VMs of type 1
- process data o3

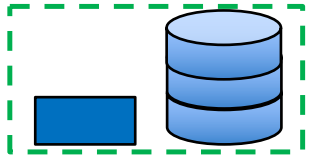
Reject
- revenue: 0
- transportation cost: 0

request k + 1:
- 3 time slots
- 1 VMs of type 2
- process data o3

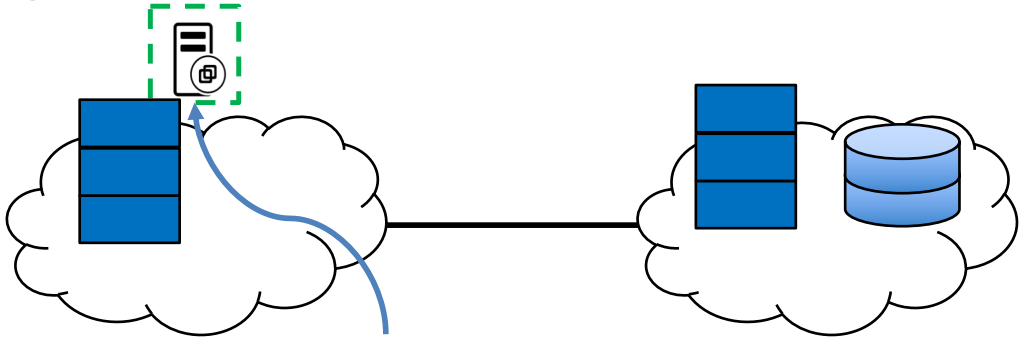
VM of type 1:
$3/time slot



VM of type 2:
$5/time slot



Making online decisions (8/9)



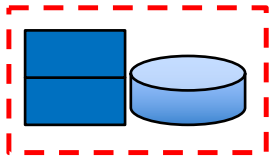
request k:
- 3 time slots
- 2 VMs of type 1
- process data o4

Reject
- revenue: 0
- transportation cost: 0

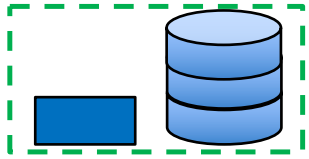
request k + 1:
- 3 time slots
- 1 VMs of type 2
- process data o3

Assemble the VM in location 1
- revenue: 3*5
- transportation cost: 0

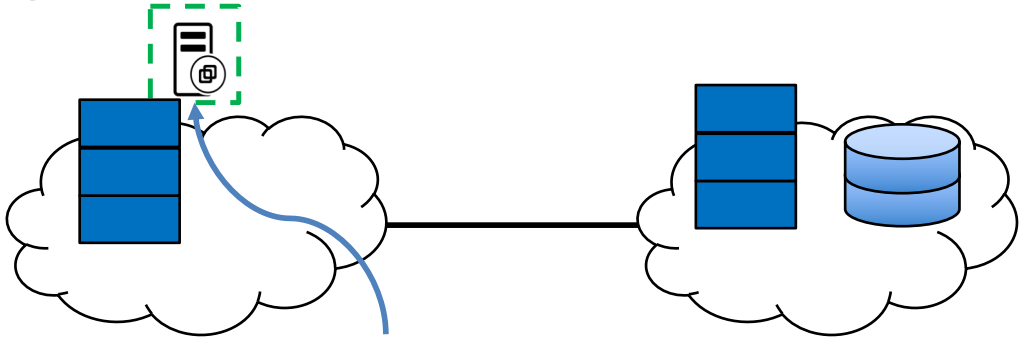
VM of type 1:
$3/time slot



VM of type 2:
$5/time slot



Making online decisions (9/9)



request k:
- 3 time slots
- 2 VMs of type 1
- process data o4

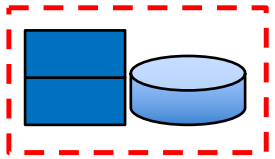
Reject
- revenue: 0
- transportation cost: 0

request k + 1:
- 3 time slots
- 1 VMs of type 2
- process data o3

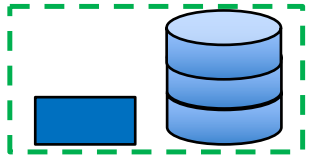
Assemble the VM in location 1
- revenue: 3*5
- transportation cost: 0

Total benefit: 3*1*5

VM of type 1:
$3/time slot



VM of type 2:
$5/time slot



Problem formulation

To optimize the long-term benefit (i.e. the time-averaged benefit) for arbitrary request sequences. Formally:

$$\max \quad \overline{R} = \lim_{T\to\infty} (1/T) \sum_{\tau=0}^{T-1} R^{\tau}$$ To maximize the time-averaged revenue

$$\text{s.t.} \quad \overline{C} \leq L$$ Keep the time-averaged transportation cost below L

$$\sum_{A\in A^l} x_A^l \leq 1, \quad x_A^l \in \{0,1\}$$ Decision: to accept an allocation A ¥in A^l or not for a certain request l

$$\sum_{l:\tau\leq t^l<\tau+1} \sum_{A\in A^l} N_{A,i,k}^l g_{r,k} x_A^l \leq c_{i,r,t} \quad \forall i, r, t,$$

Resource constraints: an allocation could not exceeds the resource amount

Decompose the coupling of the transportation cost constraint

Introduce a virtual queue

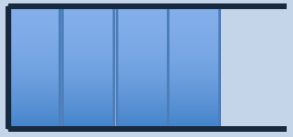
$$\overline{R} = \lim_{T\to\infty} (1/T) \sum_{\tau=0}^{T-1} R^{\tau}$$

$$\overline{C} \leq L$$

$$\sum_{l:\tau\leq t^l<\tau+1} \sum_{A\in A^l} N^l_{A,i,k} g_{r,k} x^l_A \leq c_{i,r,t} \quad \forall i, r, t,$$

$$\sum_{A\in A^l} x^l_A \leq 1, \quad x^l_A \in \{0,1\}$$

Arrival of epoch T -> C^T



Departure -> L

$$Q^{\tau+1} = Q^{\tau} + C^{\tau} - L$$

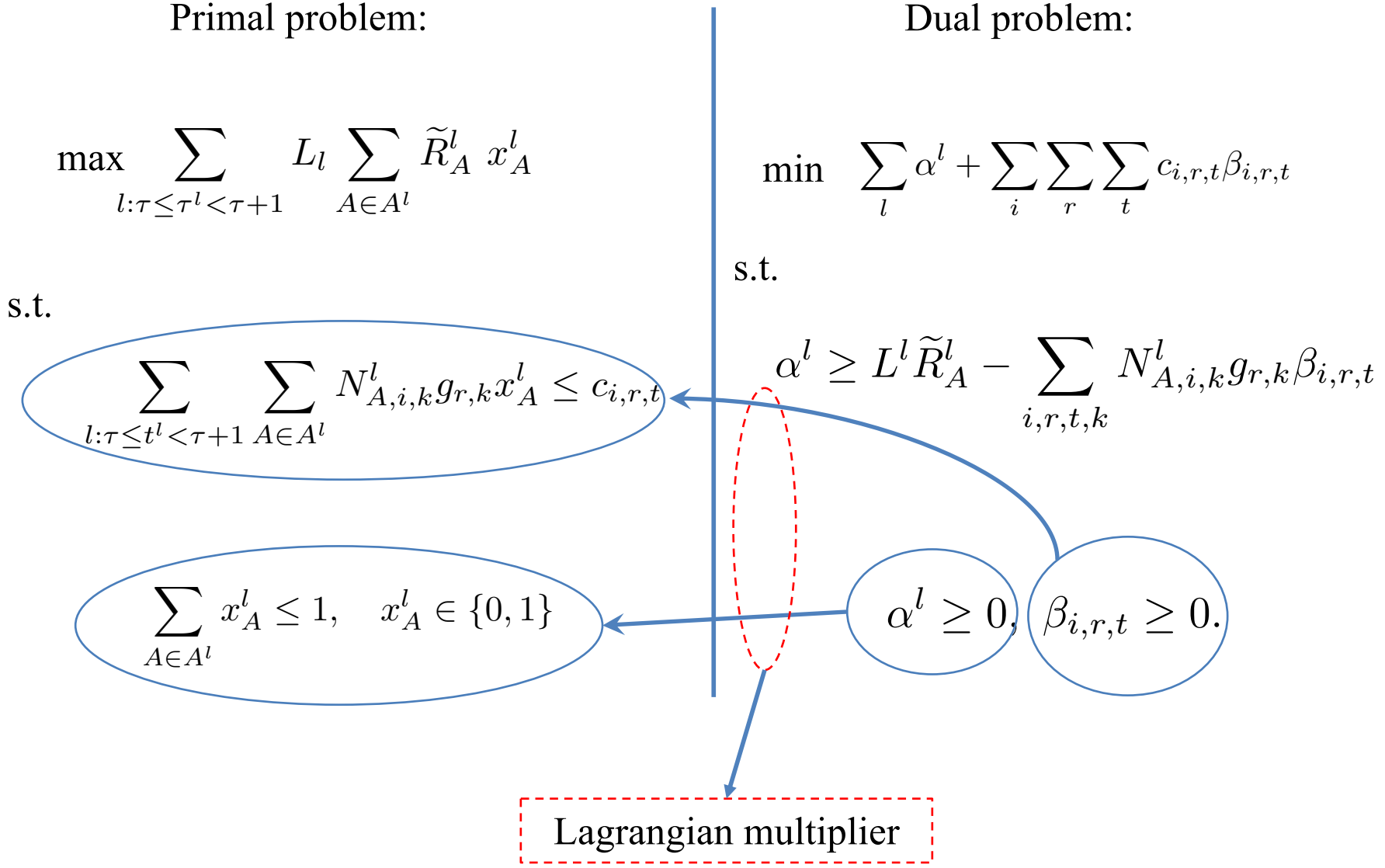
Almost equal

$$\max_{\boldsymbol{x}} V R^{\tau} - Q^{\tau} (C^{\tau} - L)$$

$$\sum_{l:\tau\leq t^l<\tau+1} \sum_{A\in A^l} N^l_{A,i,k} g_{r,k} x^l_A \leq c_{i,r,t} \quad \forall i, r, t,$$

$$\sum_{A\in A^l} x^l_A \leq 1, \quad x^l_A \in \{0,1\}$$

Construct the dual problem for each epoch



Primal-dual based resource allocation whenever a request arrives

Input:
The backlog of Q^τ
The cache configuration $\{S_i\}$
The request sequence (indexed by l)

Output: The computing resource allocation decisions

1: Initialize $\beta_{i,r,t} \leftarrow 0 \ \forall i, r, t$
2: For each request l
Compute the $A^* \in A^l$ such that

$$A^* = \underset{A \in A^l}{\operatorname{argmax}} \left(L^l \widetilde{R}^l_A - \sum_{i,r,t,k} N^l_{A,i,k} g_{r,k} \beta_{i,r,t} \right)$$

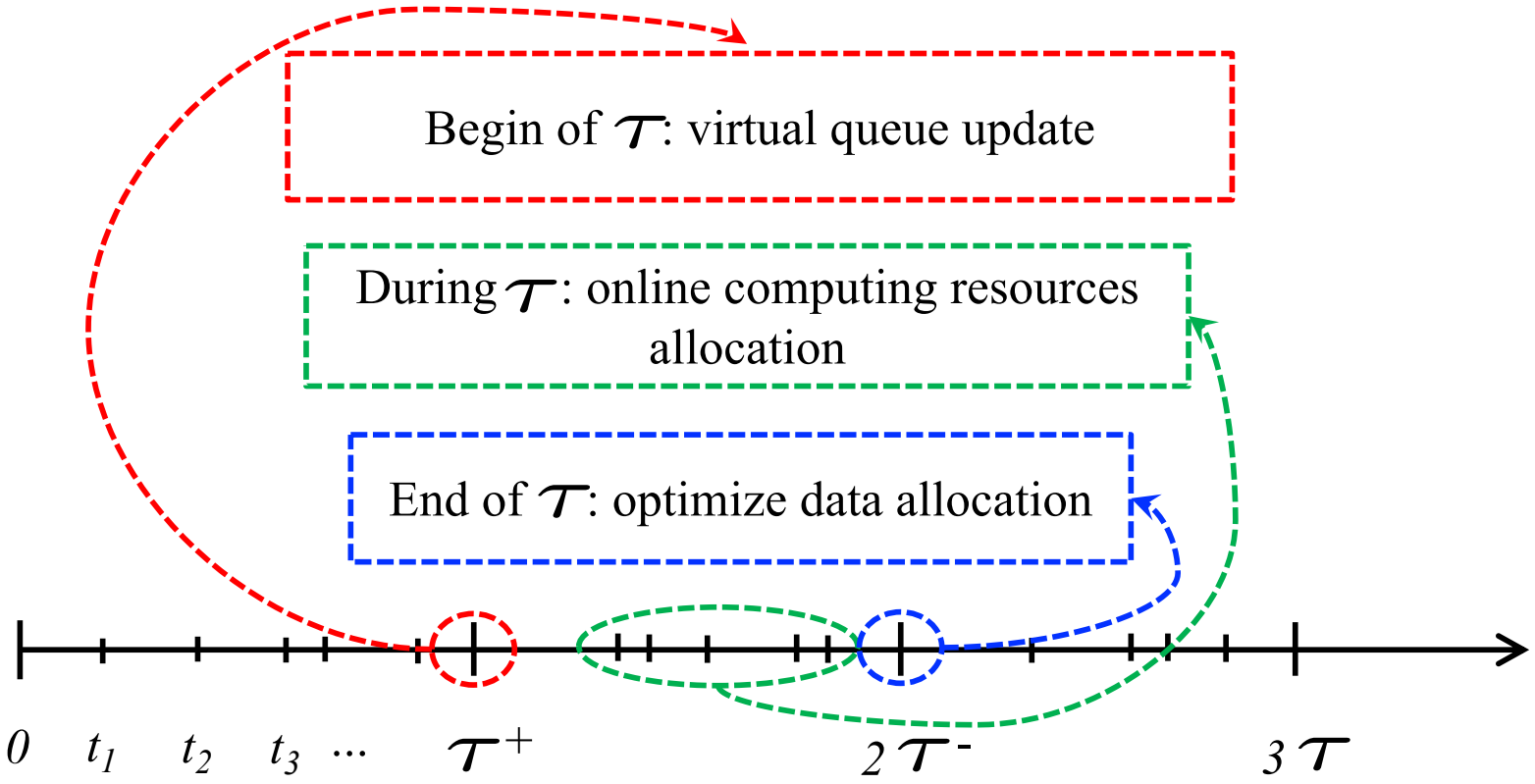
3: **if** $L^l \widetilde{R}^l_{A^*} - \sum_{i,r,t,k} N^l_{A^*,i,k} g_{r,k} \beta_{i,r,t} < 0$ **then**
4: Redirect the request to the cloud
5: **else**
6: Allocate resources to l as A^*
7: Update $\delta_{i,r,t}$ as

$$\beta_{i,r,t} \leftarrow \beta_{i,r,t} \left(1 + \frac{\sum_k N^l_{P^*,i,k} g_{r,k}}{c_{i,r,t}} \right) + \frac{1}{e-1} \frac{\widetilde{R}^l_{A,i}}{\mathbf{R} c_{i,r,t}}$$

8: Set
$\alpha^l \leftarrow L^l \widetilde{R}^l_{A^*} - \sum_{i,r,t,k} N^l_{A^*,i,k} g_{r,k} \beta_{i,r,t}$
9: **end if**

Update the primal variables less than (*1 – 1/e*) times of the corresponding dual variables

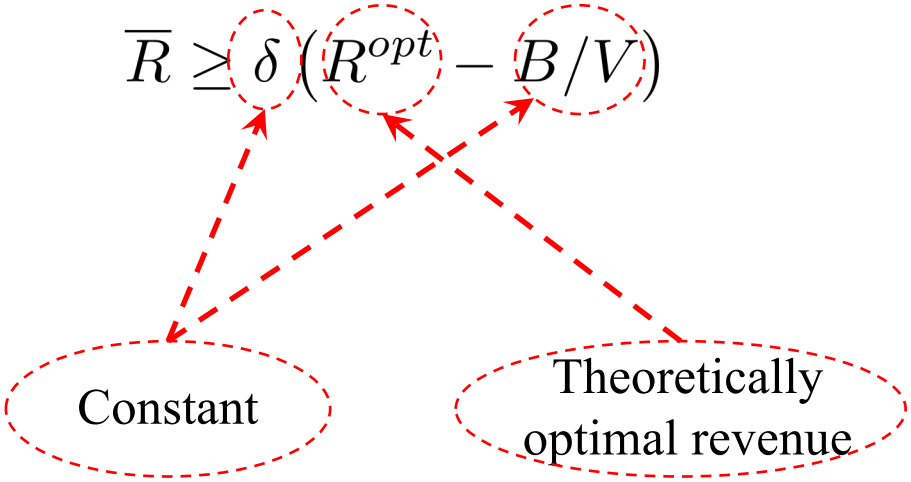
Primal-dual based resource allocation whenever a request arrives



Near-optimal performance of the proposed approach

Theorem 1. *For any $V \geq 0$, if the problem is feasible, then the drift-plus-penalty algorithm stabilizes the virtual queue, and*

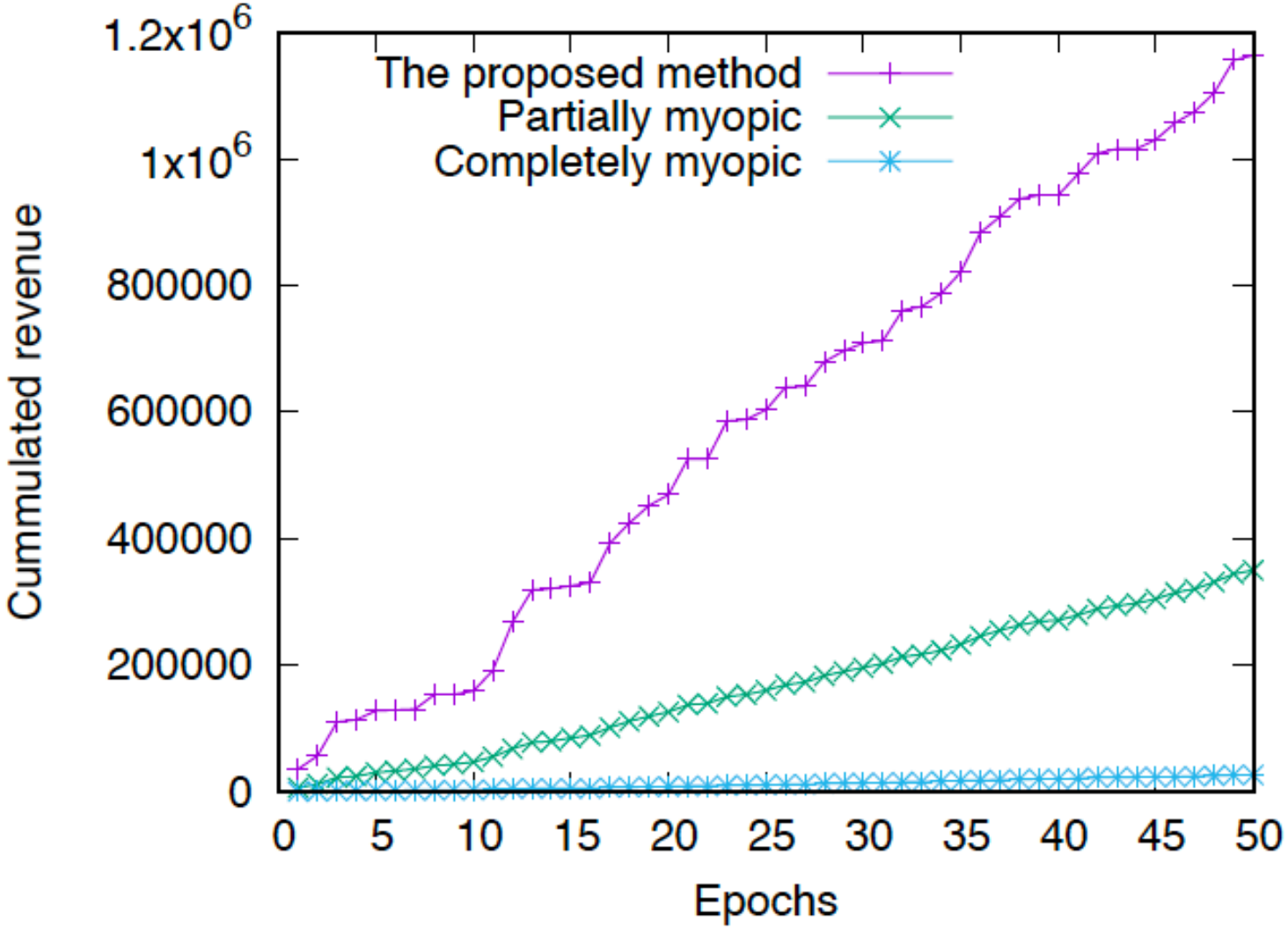
$$\overline{R} \geq \delta\left(R^{opt} - B/V\right)$$



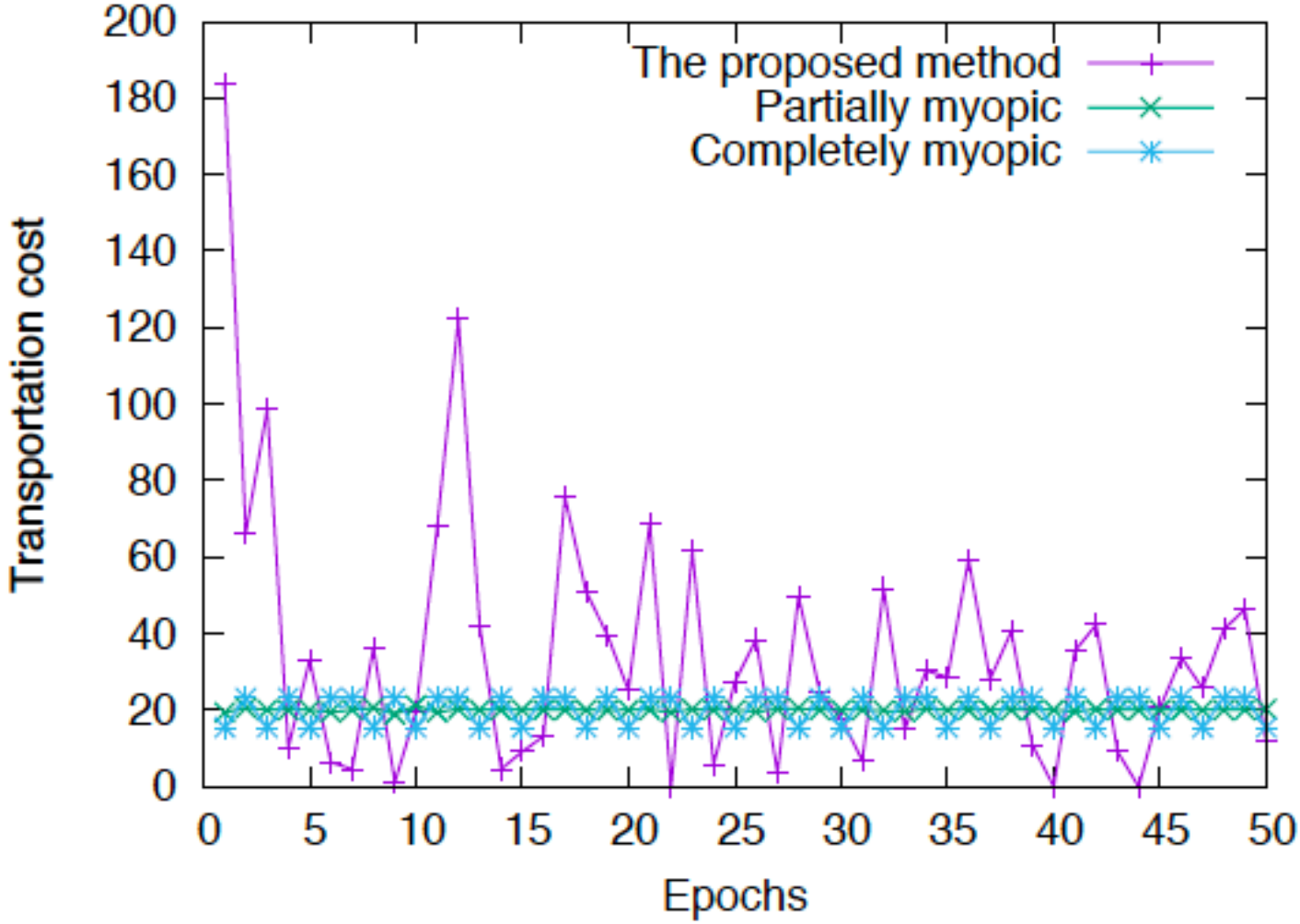
Experiment Settings

Symbol	Explanation
Number of edge computing locations	5
Topology of the locations	Full mesh
The transportation cost between locations	follows $U(10, 20)$
The transportation cost from edge to cloud	fixed to 100
Kinds of resources	3
The resources each location has	$[5000, 5000, 5000]$
Types of VMs	2 types of VMs; VM 1 is assembled by resources $[1, 2, 3]$, while VM 2 is assembled by resources $[3, 2, 1]$
Types of request	VM 1 or VM 2 with the same probability
The price for each type of VM	$[10, 20]$
Content popularity	Zipf distribution with exponent 0.6
Request arrivals	Poisson distribution; λ follows $U(1, 10)$
Length of requests	$U[0 : 0.25$ epoch$]$

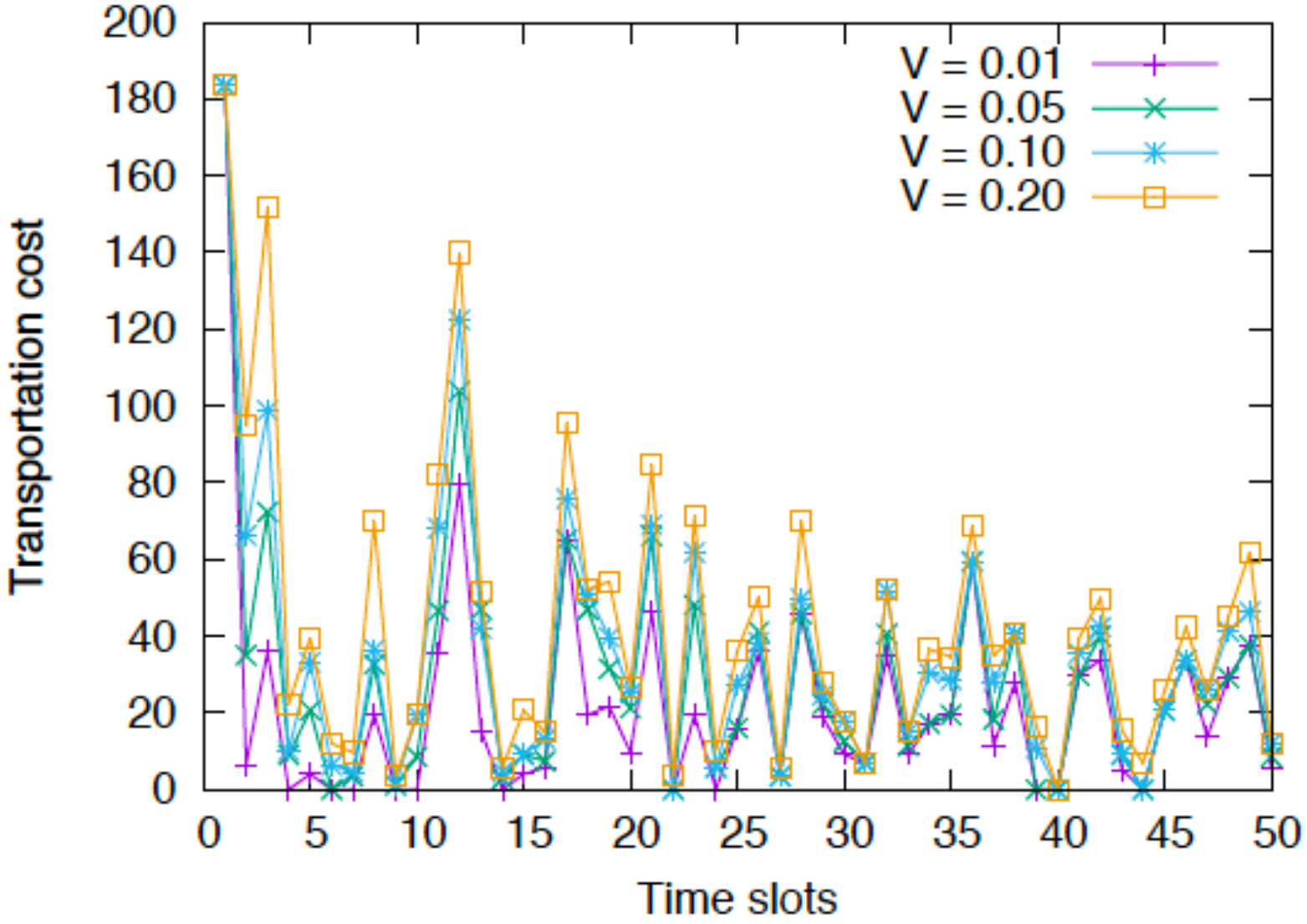
Cumulated revenue



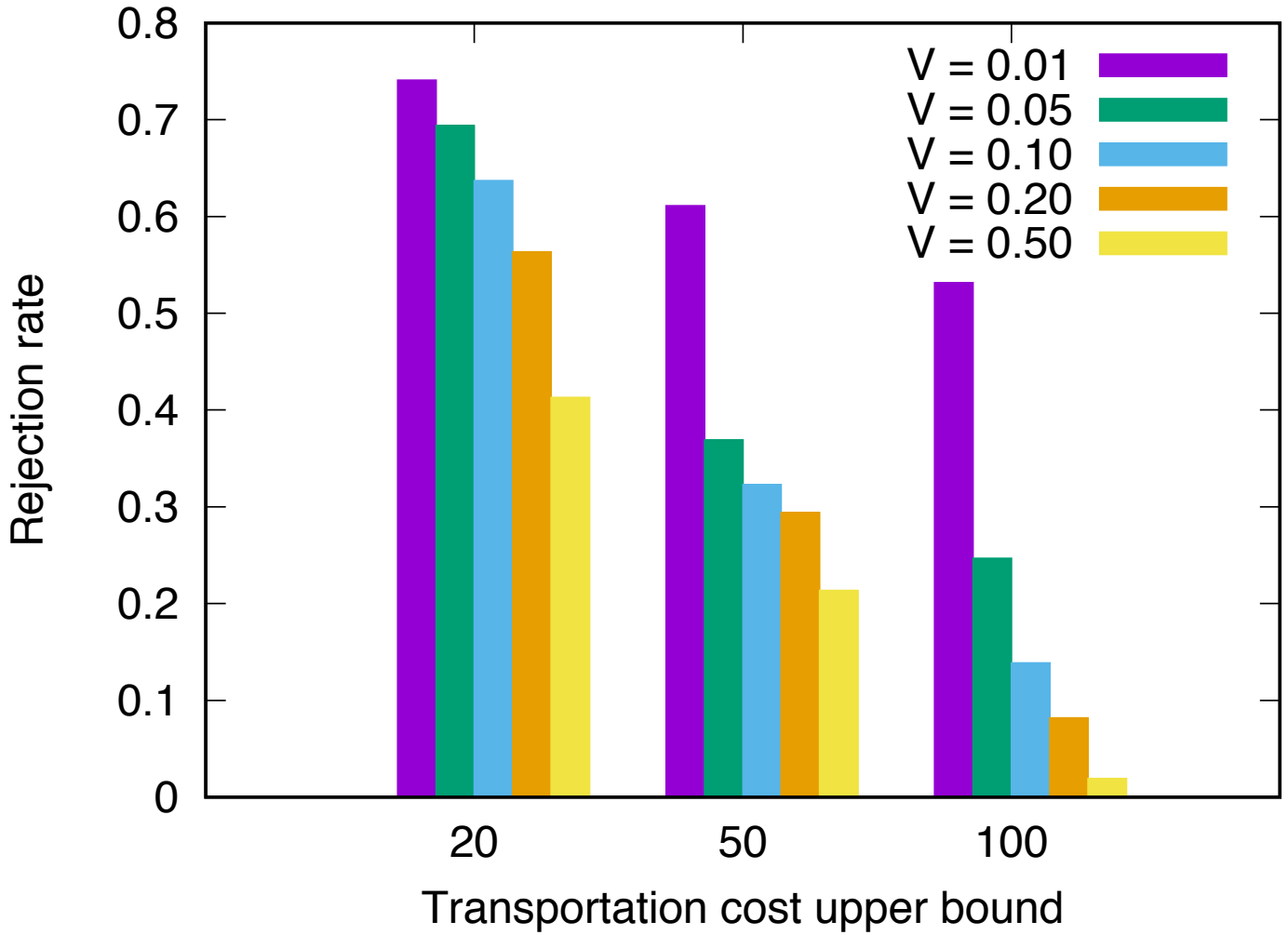
Transportation cost



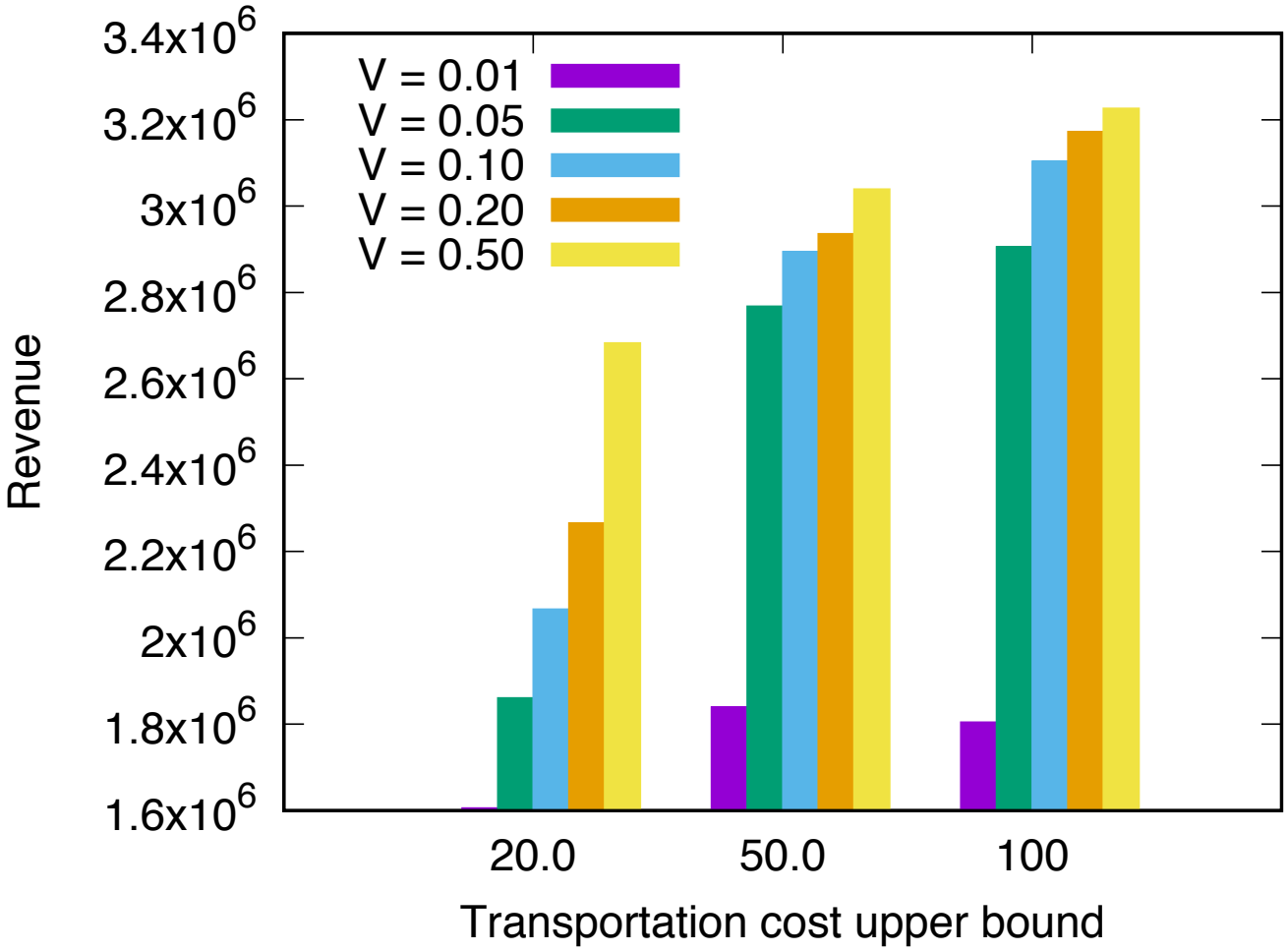
The violation of the transportation cost constraint



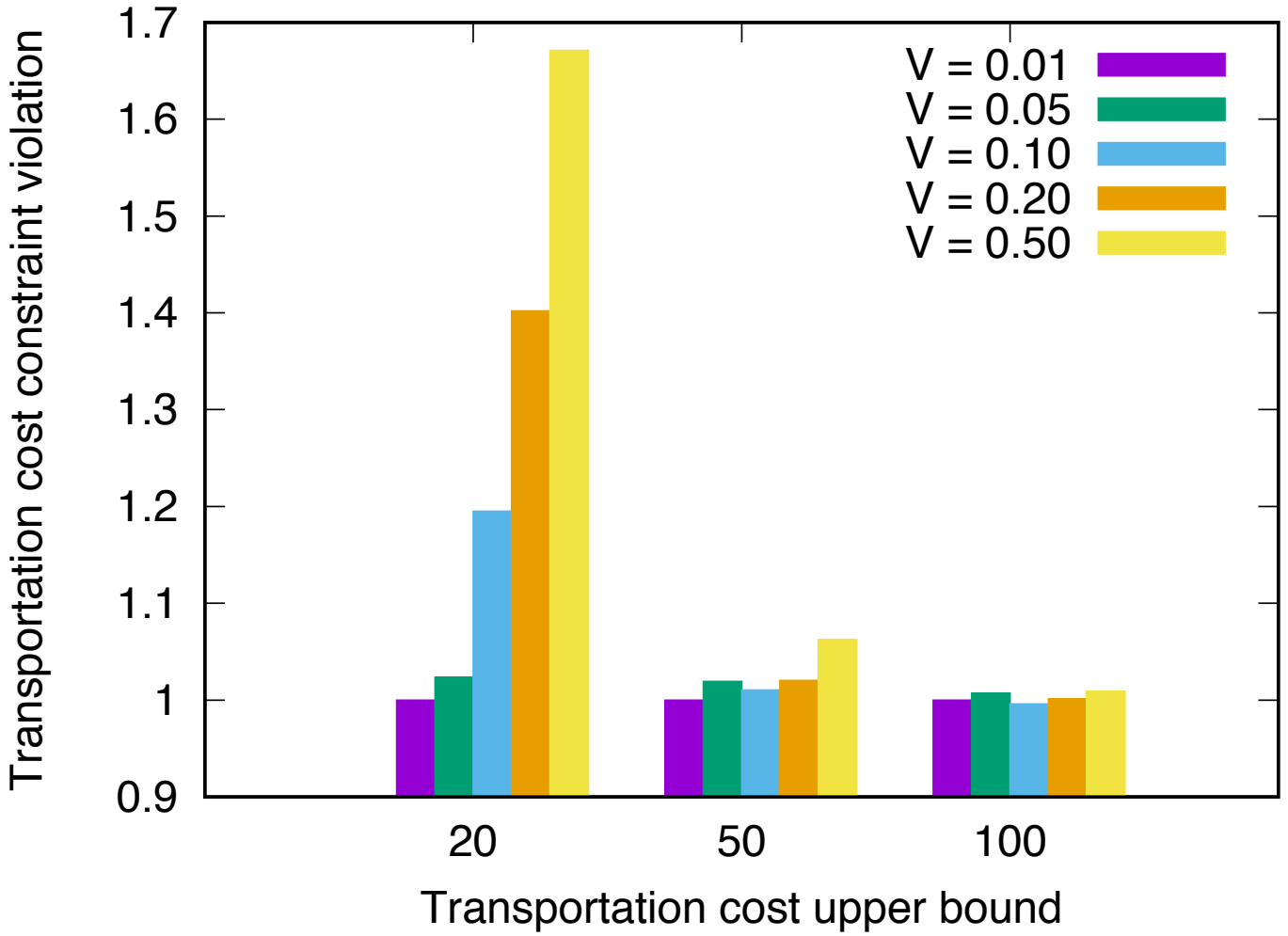
The trade off between rejection rate and the transportation cost upper bound



The trade off between time-averaged revenue and the transportation cost upper bound



The trade off between the transportation cost constraint violation and the transportation cost upper bound



Conclusions

- We proposed a joint data caching and processing framework for MEC
- We analyze the main challenges of the joint optimization of caching and processing
- We propose an online approach to solve the joint optimization problem without knowledge of the future
- The proposed approach achieves proven performance guarantees