An Economical and SLO-Guaranteed Cloud Storage Service across Multiple Cloud Service Providers

Guoxin Liu and Haiying Shen

Presenter: Haiying Shen

Associate professor

Department of Electrical and Computer Engineering,
Clemson University, Clemson, USA

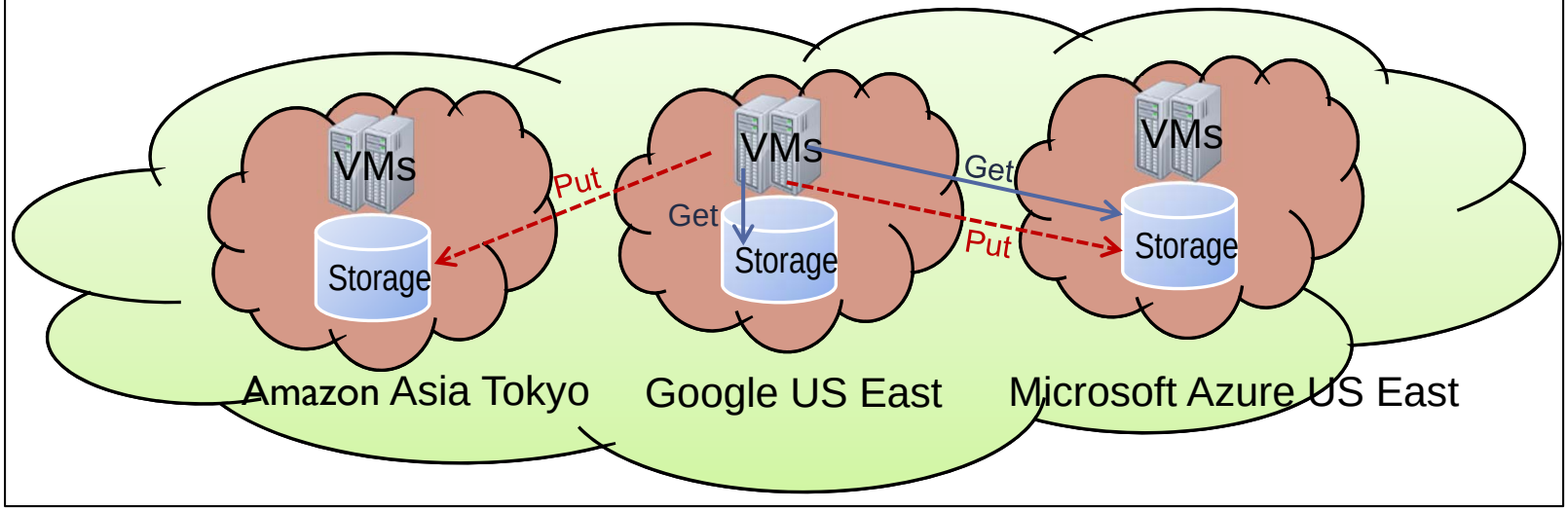
Outline

- Introduction
- Related work
- Problem Statement
- Economical and SLO-guaranteed Service
- Evaluation
- Conclusion

Cloud Service Providers (CSPs)

Geo-Distributed Storage over CSPs

- Use different CSPs
  - Objective: minimize payment cost
- Non-trivial



Geo-Distributed Storage over CSPs

- Cloud service broker
  - Collects resource usage requirements from customers
  - Generates data allocation over multiple clouds
    - Data storage and Get request allocation
- Reduce cost by leveraging different pricing policies
  - Different cloud providers, different datacenters of a cloud provider
  - Tiered pricing
  - Location of the destination datacenter
  - Pay-as-you-go price > reservation price
- Problem:
  - Input: customer data and request rates
  - Output: data storage & request allocation and resource reservation

Outline

- Introduction
- Related work
- Problem Statement
- Economical and SLO-guaranteed Service
- Evaluation
- Conclusion

Related Work

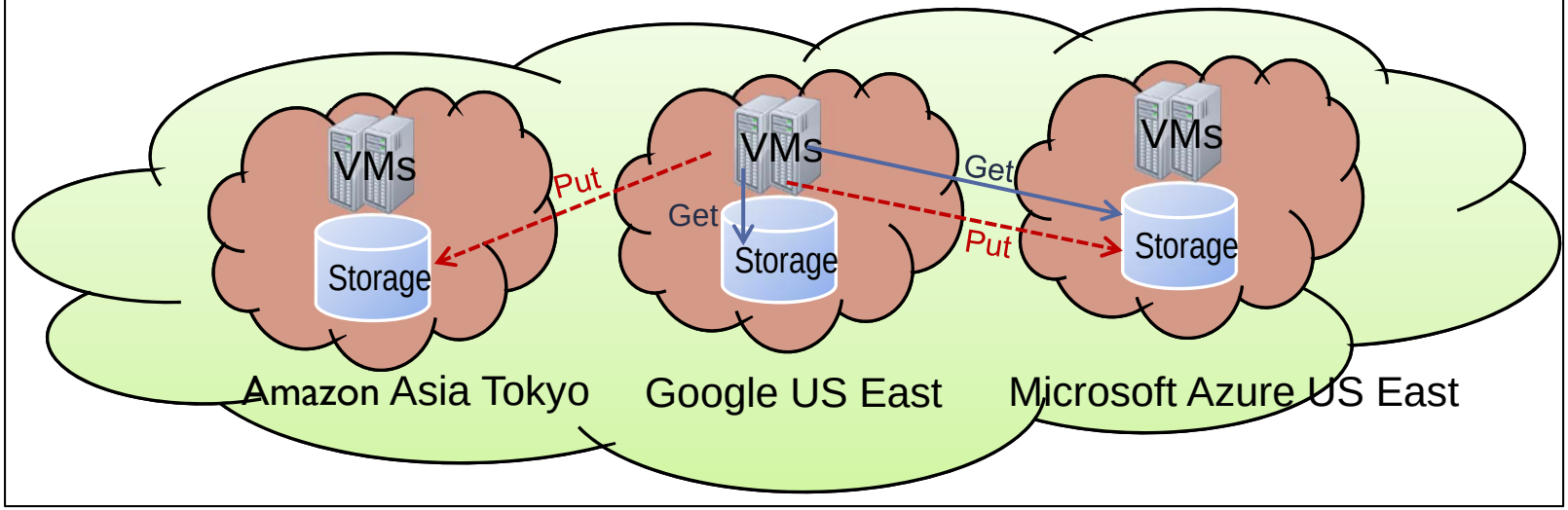
- Storage services over multiple clouds
  - Data availability, data retrieval latency
- Cloud/datacenter storage payment cost minimization
  - Adaptively assign data with different sizes to different storage services to minimize the cost for storage
- Pricing models on clouds
  - Dynamic pricing models
    - Adaptive learning auctions
    - Model concave gain
- Cloud service SLO guarantee
  - Guarantee service latency SLO and achieve high throughput
    - Caching and scheduling

Fully utilize all the pricing policies

Both cost minimization and SLO guarantee

SLO Guaranteed with Cost Minimization

- ES^3: Economical and SLO-guaranteed cloud Storage Service
  - Geographically distributed cloud storage service for multiple customers over multiple clouds with SLO guarantee and cost minimization



Outline

- Introduction
- Related work
- Problem Statement
- Economical and SLO-guaranteed Service
- Evaluation
- Conclusion

Data Allocation and Resource Reservation

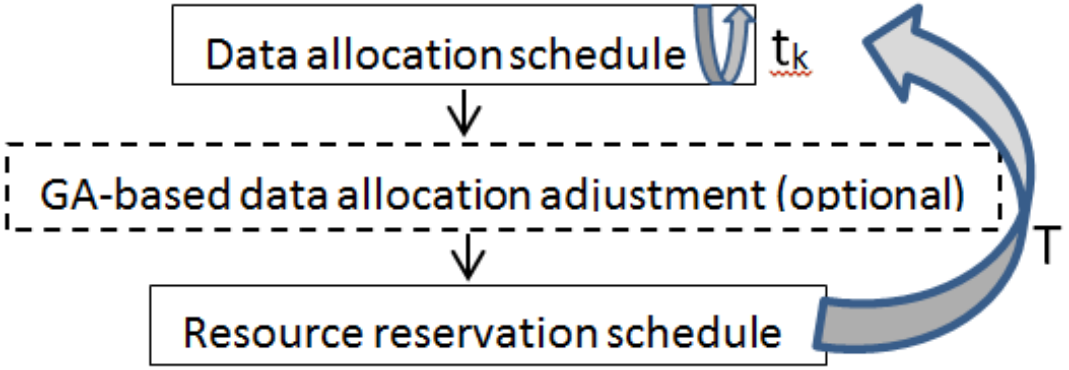
- Payment minimization objective
  - Get/Put: Minimize cost under both pay-as-you-go and reservation
  - Storage/Transfer: Minimize pay-as-you-go under tiered pricing model
- Constraints
  - Ensure Get/Put SLO for each tenant
  - Ensure data availability (maintain replicas over datacenters)
  - Avoid data congestion of a single datacenter
- NP-Hard

Outline

- Introduction
- Related work
- Problem Statement
- Economical and SLO-guaranteed Service
- Evaluation
- Conclusion

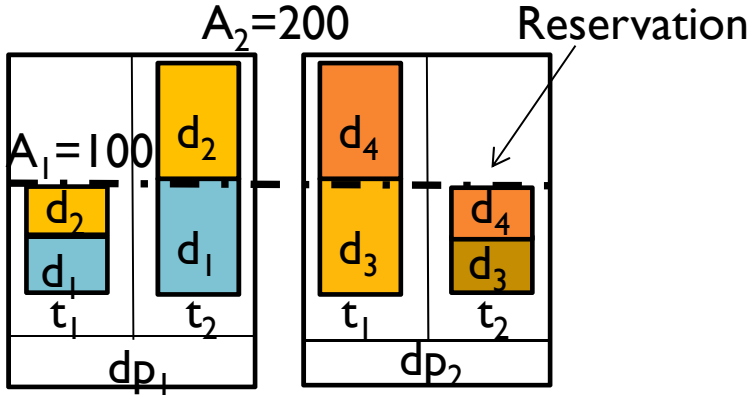
Economical and SLO-guaranteed Cloud Storage Service

- Cost minimization with SLO guarantee under pay-as-you-go model & reservation benefit maximization
  - Coordinated data allocation and reservation
- Cost minimization under reservation
  - GA-based data allocation adjustment
- Dynamic Get rate variation
  - Dynamic request redirection



Coordinated Data Allocation and Resource Reservation

- **A**={A1,A2,….An} as a list of the number of Gets in different t_k in T sorted in an increasing order
- Rule 1: Among several datacenter candidates to allocate a data item, we need to choose the datacenter that leads to the largest A1 increment after being allocated with the data item
- The optimal reservation is the Ai in **A** that generates the largest reservation benefit



Unbalanced data allocation

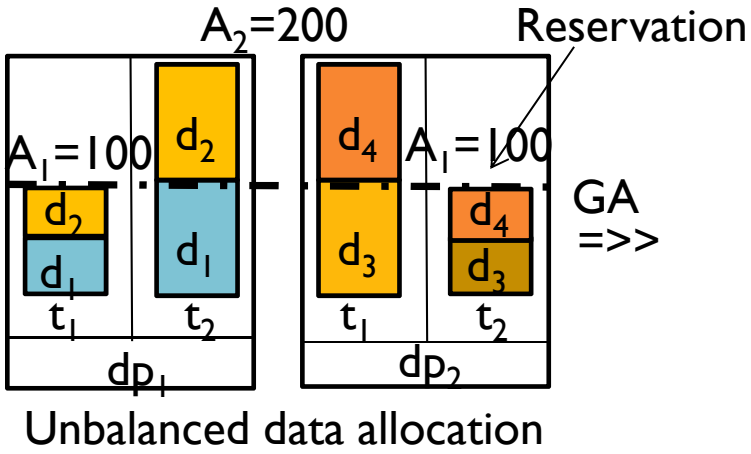
Data Allocation and Reservation

- Dominant cost
  - Cost >> sum of all other costs
- Intensive data
  - Storage/Get/Put intensive (storage/Get/Put cost domination)
  - Existence of high domination ratio
    - Storage intensive: Old webpages/photos/videos
    - Put intensive: Replicas for data availability
    - Get intensive: New popular news/photos/videos

Data Allocation and Reservation

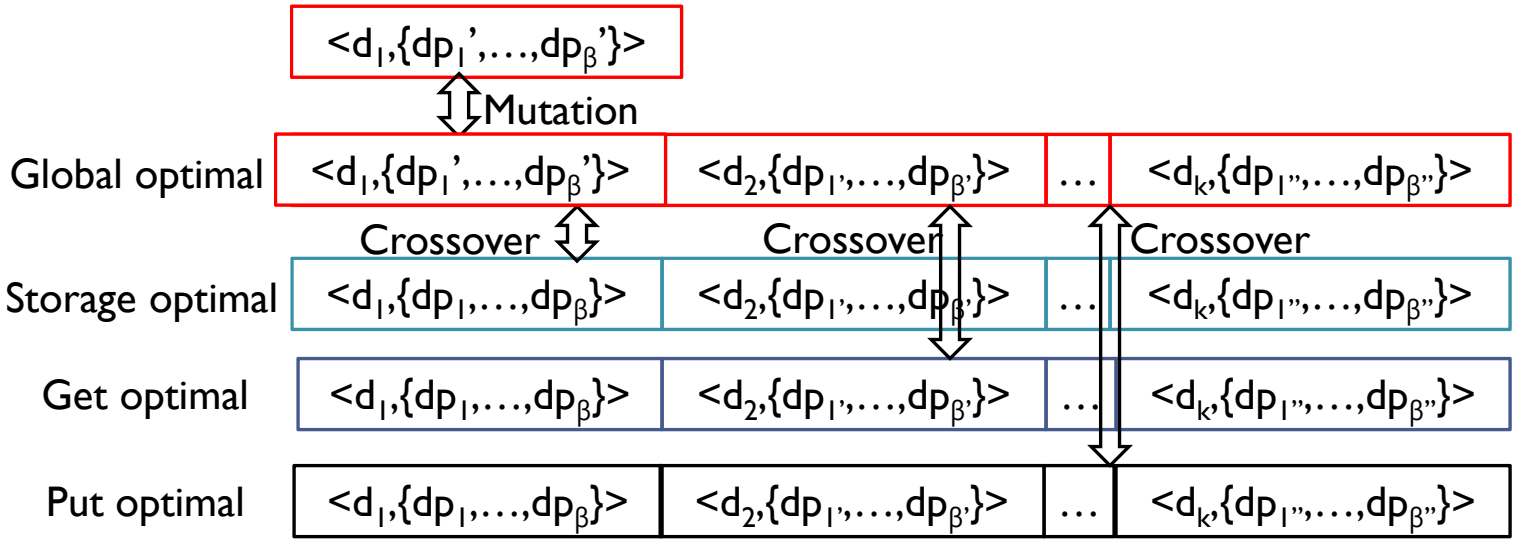
- SLO guarantee
  - Latency and capacity aware cloud service selection
- Cost efficiency
  - Get/Put intensive data
    - Minimum unit price & follow Rule 1
  - Storage intensive data
    - Minimum unit price and maximum aggregation size

Is it optimal?



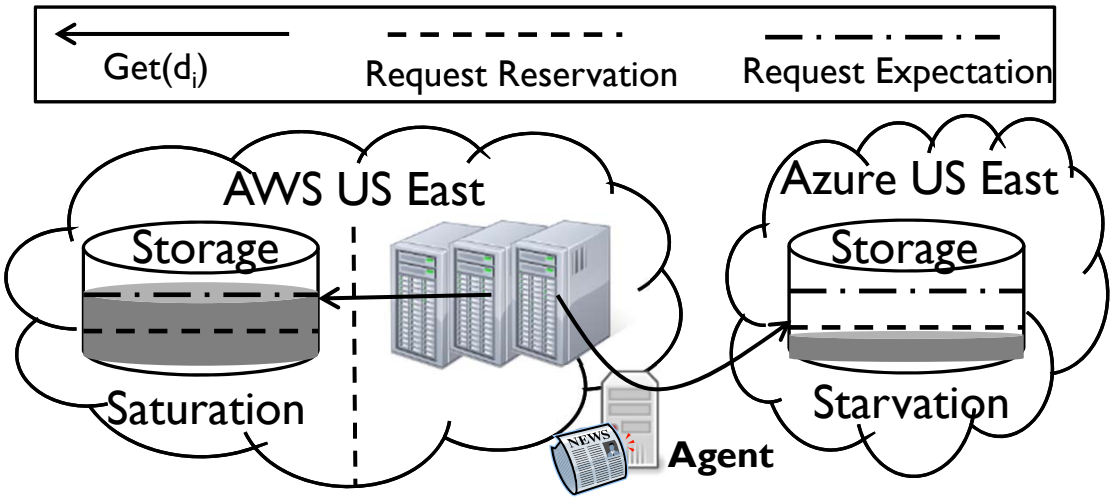
GA-based Data Allocation

- Genetic algorithm (GA): mimics the process of natural selection
  - Gene: Data allocation of a data item
  - Crossover: Between global-optimal and sub-optimals
  - Mutation: Approach to global optimal



Dynamic Request Redirection

- Observation: Get rate variation
- Solution
  - Data availability: Multiple replicas
  - Saturation datacenter ➔ Starvation datacenter
    - Saturation: Usage over reservation
    - Starvation: Usage under reservation



Outline

- Introduction
- Related work
- Problem Statement
- Economical and SLO-guaranteed Service
- Evaluation
- Conclusion

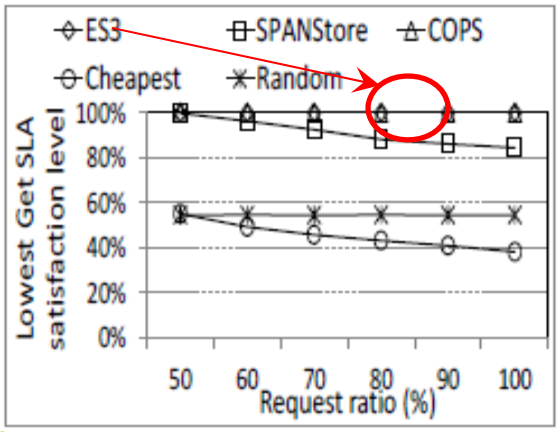
Evaluation of ES^3

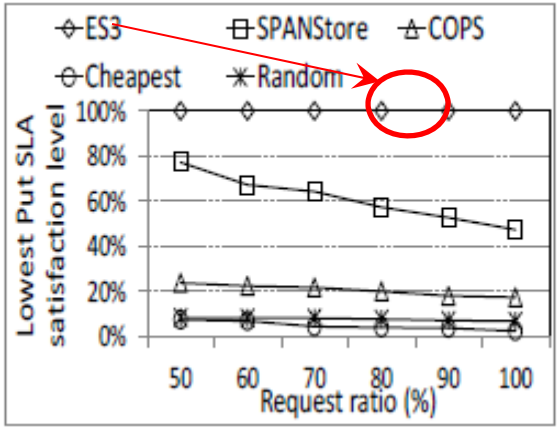
- Simulated CSPs: 25 regions
  - Amazon S3, Microsoft Azure, and Google cloud storage
- Simulated customers
  - 52 Cloud customers
- Real deployment
  - One customer: Amazon EC2 US East & West regions
- Comparison
  - COPS [9]: shortest latency
  - SPANStore [10]: latency guaranteed and unit cost minimization
  - Cheapest: Unit cost minimization
  - Random: Random CSP region selection

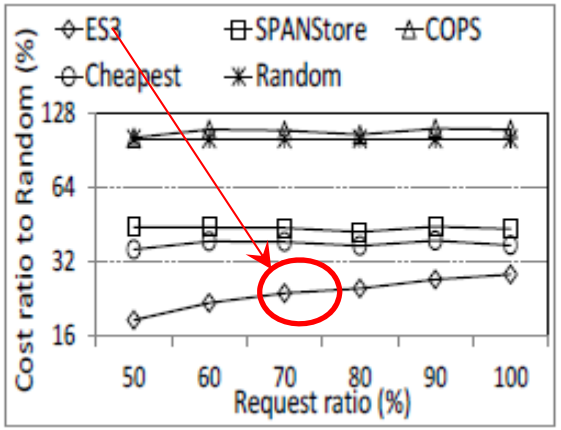
[9] W. Lloyd, M. J. Freedman, M. Kaminsky, and D. G. Andersen. Dont Settle for Eventual: Scalable Causal Consistency for Wide-Area Storage with COPS. In Proc. of SOSP, 2011.
[10] Z. Wu, M. Butkiewicz, D. Perkins, E. Katz-Bassett, and H. V. Madhyastha. SPANStore: Cost-Eective Geo-Replicated Storage Spanning Multiple Cloud Services. In Proc. of SOSP, 2013.

Evaluation (cont.)

- Effect of ES^3
  - Due to capacity and latency awareness
    - ES^3 supplies get-SLO and put-SLO guaranteed service
  - Due to the comprehensive pricing policy awareness
    - ES^3 generates the minimum payment cost to CSPs







Outline

- Introduction
- Related work
- Problem Statement
- Economical and SLO-guaranteed Service
- Evaluation
- Conclusion

Conclusion

- Multi-cloud Economical and SLO-guaranteed cloud Storage Service (ES3)
  - Coordinated data allocation and reservation
  - GA-based data allocation adjustment
  - Dynamic request redirection
- Effectiveness:
  - Minimize payment cost and achieve the SLO of each tenant
- Future wok:
  - Dynamically create and delete data replicas in datacenters to fully utilize the Put reservation

Thank you!
Questions & Comments?

Haiying Shen, Associated Professor

shenh@clemson.edu

Pervasive Communication Laboratory

Clemson University