Harnessing the Power of Multiple Cloud Service Providers: An Economical and SLA-Guaranteed Cloud Storage Service

Guoxin Liu and Haiying Shen
Department of Electrical and Computer Engineering, Clemson University
({guoxinl, shenh}@clemson.edu)

Background and Motivation

Cloud storage (e.g., Amazon S3, Microsoft Azure and Google Cloud Storage), as an emerging commercial service, is becoming increasingly popular. For a cloud customer's application, the data access latency is negatively proportional to the customer's income. In order to maximize profits, cloud customers must provide low data Get/Put latency and high availability to their clients while minimizing the total payment cost to the Cloud Service Providers (CSPs). Since different CSPs provide different storage service prices, customers tend to use services from different CSPs instead of a single CSP to minimize their payment cost (cost in short). However, the technical complexity of this task makes it non-trivial to customers, which calls for the assistance from a third-party organization.

Current solutions

Latency minimization: It allocates customer data into a datacenter with the shortest latency. However, It does not consider the Get/Put capacity limitation of a cloud service provider's datacenter.

Cost minimization: It selects the datacenters with the cheapest cost in the pay-as-you-go-manner. However, It does not consider the Get/Put SLA requirements (deadline and maximum percentage of requests violating deadlines) and the complex pricing policies.

Our approach

Broker: A broker collects resource usage requirements from many customers, makes resource requests to multiple clouds, pays the CSPs for the resources actually consumed as a customer, and charges its customers as a CSP.

ES^3: Economical and SLA-guaranteed cloud Storage Service for brokers.

Design Details

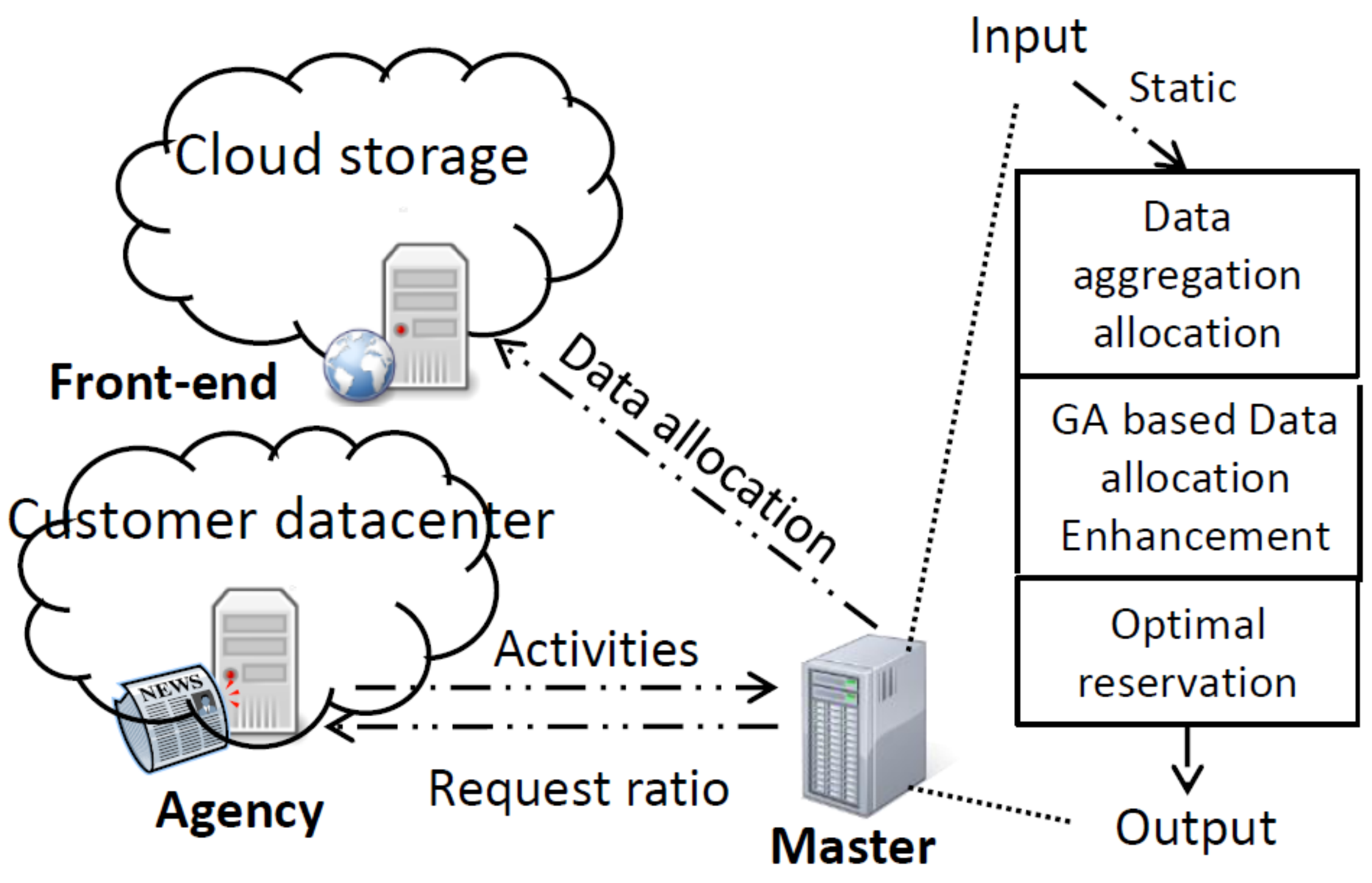

Figure 1 Overview of ES^3.

Scheme 1: Data Allocation and Resource Reservation.

- Storing the data in the datacenter that has the cheapest unit price for its dominant cost (e.g., Get, Put or Storage) can reduce the cost greatly.
- Storing the Storage-intensive data in the datacenter that results in the largest aggregate data storage size can reduce the cost greatly based on the tiered pricing policy.
- Storing the data in the datacenter with the largest reservation benefit if the data is Get-/Put-intensive, in order to minimize the reservation cost.

After the data allocation, the reservation of Gets/Puts in each datacenter is determined according to [2].

Scheme 2: GA-based Data Allocation Adjustment.

Genetic Algorithm (GA): A heuristic method that mimics the process of natural selection.

- Gene: Data allocation of a data item.
- Parents: Global optimal with minimum total cost and Storage/Get/Put optimal with minimum Storage/Get/Put cost.
- Crossover: Hatching between global-optimal and sub-optimal solutions.
- Mutation: Approach to global optimal.

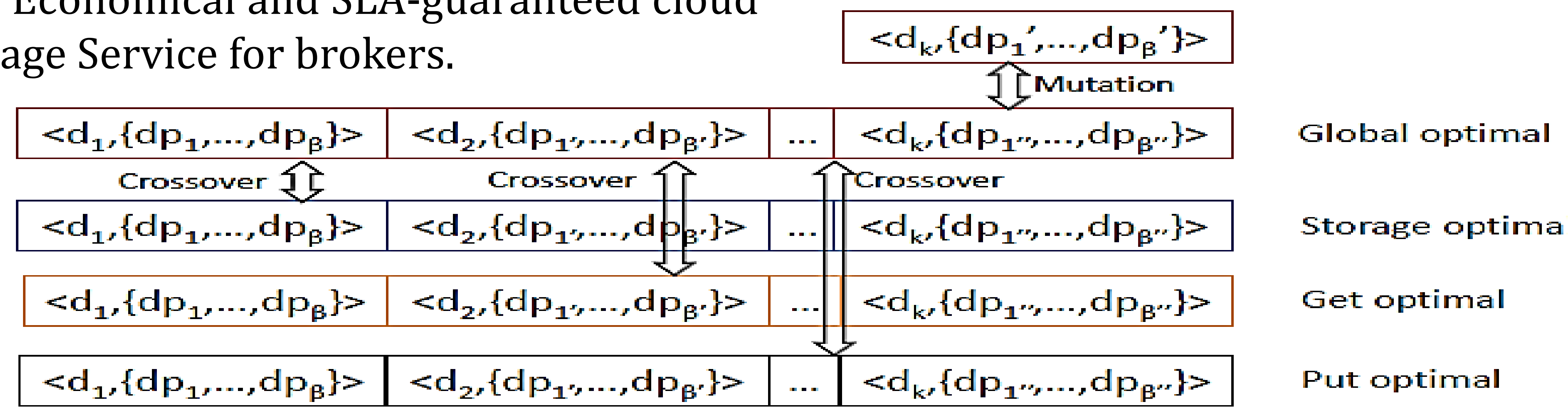
$<d_k,\{dp_1',...,dp_\beta'\}>$
Mutation
$<d_1,\{dp_1,...,dp_\beta\}>$
$<d_2,\{dp_{1'},...,dp_{\beta'}\}>$
...
$<d_k,\{dp_{1''},...,dp_{\beta''}\}>$
Global optimal
Crossover
Crossover
Crossover
Storage optimal
Get optimal
Put optimal

Figure 2 GA-based data allocation adjustment.

Experimental Results

We conducted trace-driven experiments on Clemson University's Palmetto Cluster [3]. We simulated 50 globally distributed datacenters in 25 cloud storage regions, and compared ES^3 with SPANStore [4], COPS [5], Cheapest, and Random.

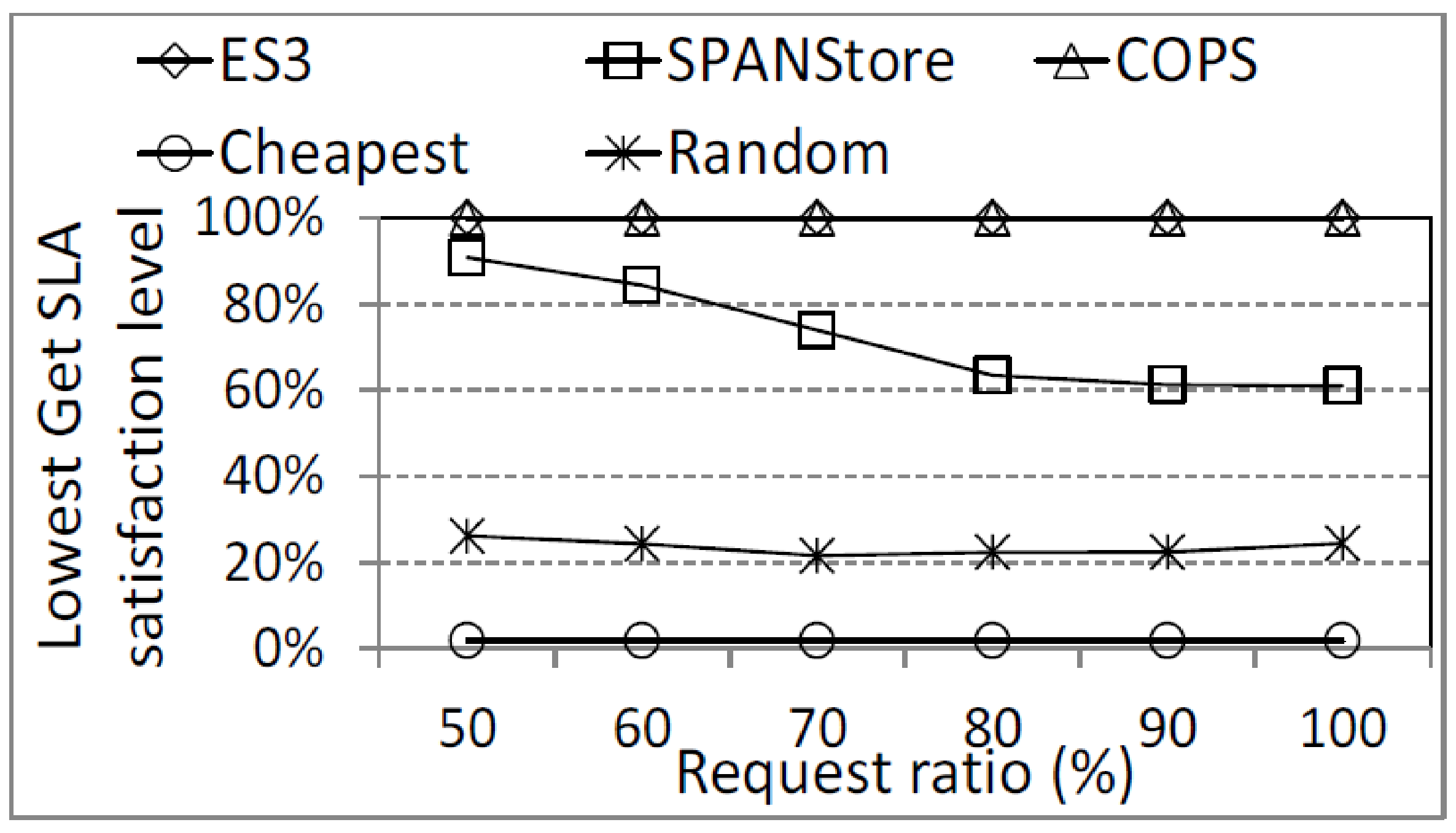

Figure 3 Get SLA guarantee.

Result: Supply a Get SLA satisfied service.

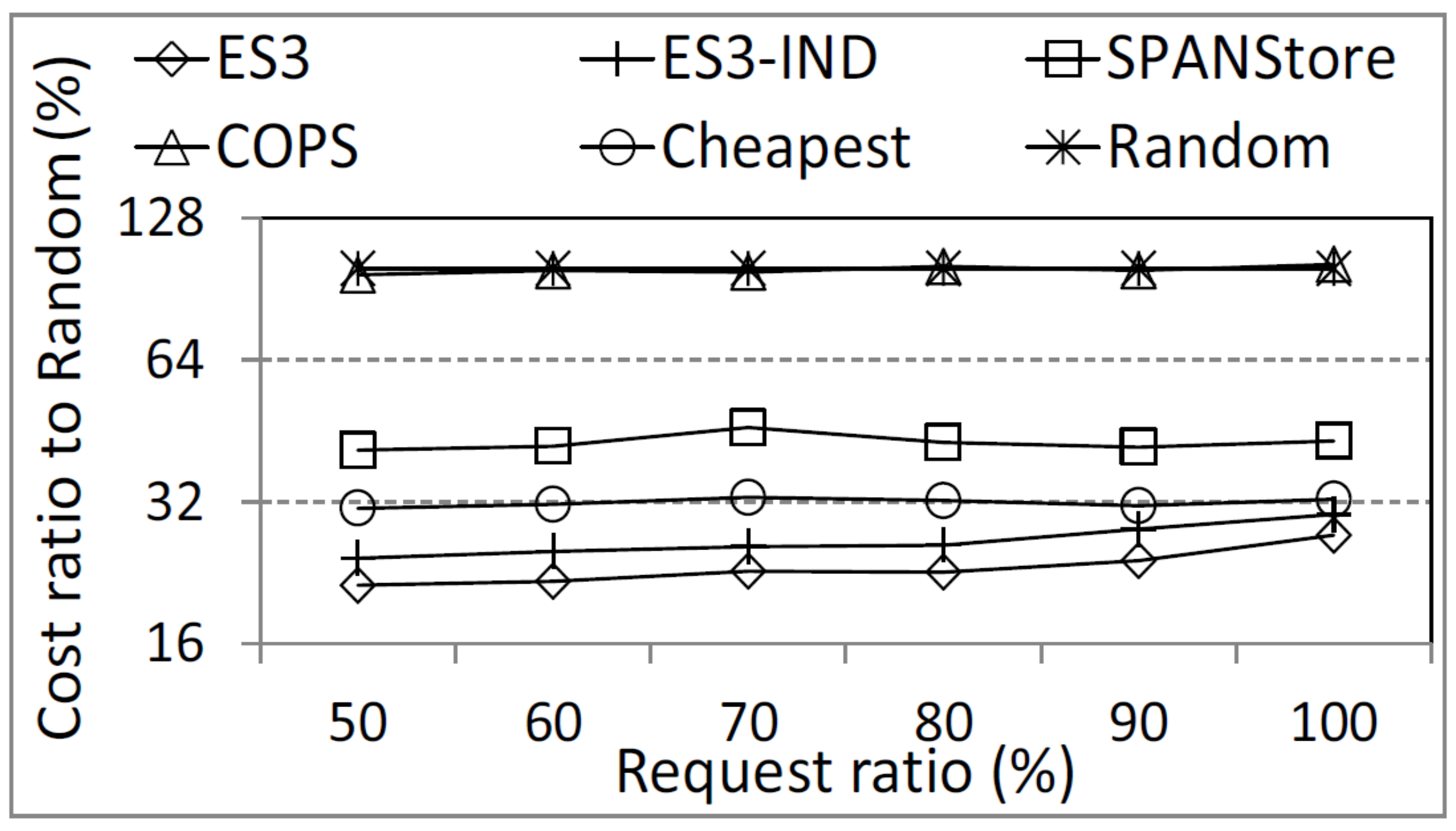

Figure 4 Cost minimization.

Result: Supply a cost minimization service.

References:

[1] G. Liu and H. Shen. Geographical Cloud Storage Service with SLA Guarantee over Multiple Cloud Providers. Technical report, Clemson University, 2014.

[2] Palmetto Cluster. http://citi.clemson.edu/palmetto/.

[3] Z. Wu, M. Butkiewicz, D. Perkins, E. Katz-Bassett, and H. V. Madhyastha. SPANStore: Cost-Effective Geo-Replicated Storage Spanning Multiple Cloud Services. In SOSP, 2013.

[4] 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.

Future Work

Develop dynamical Get redirection strategy under request burst to fully utilize Get reservation to reduce cost further.

Acknowledgments

U.S. NSF grants NSF-1404981, IIS-1354123, CNS-1254006, CNS-1249603, and Microsoft Research Faculty Fellowship 8300751.