Cost-aware Cloud Storage Service Allocation for Distributed Data Gathering

Catalin Negru, Florin Pop, Mariana Mocanu, Valentin Cristea
Computer Science and Engineering Department
University *Politehnica* of Bucharest, Romania
Emails: catalin.negru@cs.pub.ro, florin.pop@cs.pub.ro,
mariana.mocanu@cs.pub.ro, valentin.cristea@cs.pub.ro

Anca Hangan, Lucia Vacariu
Automation and Computer Science Faculty
Technical University of Cluj-Napoca, Romania
Emails: anca.hangan@cs.utcluj.ro,
lucia.vacariu@cs.utcluj.ro

***Abstract*—In today cyber-infrastructures, large datasets are produced in real-time by different sources geographically distributed. These data must be acquired and preserved for further use in knowledge extraction. In the context of multi-cloud environments, the cost-efficient storage service selection is a challenge. There are plenty of Cloud storage providers offering multiple options so, it is crucial to select the best solution in terms of cost and quality of service that meet customers requirements. Due to its multi-objective nature, the process of optimal service selection becomes a difficult problem. In this paper, we study the multi-objective optimization problem for storage service selection. We start from a real world case scenario and build our mathematical model for the optimization problem. Then we propose an aggregated linear programming technique to find a near optimal solution for the service selection problem.**

***Index Terms*—Cost optimization; Linear Programming; Data Storage; Cloud Computing; Datacenters.**

I. Introduction

Nowadays, large datasets are produced by different sources and devices, such as: sensor networks, water resources management platforms, scientific experiments, WWW. Moreover, both data sources and data produced by them, increase exponentially [1], [2]. Basically, we face with a data deluge, and size has surpassed the capabilities of computation [3].

Generated data must be acquired and preserved for further use in knowledge extraction. Cloud storage services are attractive as they present a series of important features such as on-demand capacity, virtualized storage infrastructure, over an overlay network. Also, a SLA (Service Level Agreement) contract guarantees a minimum level of performance for the service. So, users can buy storage capacity and pay for the used services.

The cost for used services plays a key role when buying Cloud storage services. Therefore, the cost-efficiency of management operations is mandatory [4]. Each Cloud storage service, offered by a different provider, is characterized by specific features, limitations and prices. There are plenty of Cloud storage providers offering multiple options so, it is crucial to select the best solution in terms of cost and quality of service that meet customers requirements. Due to its multi-objective nature, the process of optimal service selection become a difficult problem.

In the context of multi-cloud environment, the cost-efficient service selection problem arises, which is particularly challenging for cyber-infrastructures. These systems refer to research environments, designed to handle different operations on data (e.g. data acquisition, storage, visualization and processing) distributed over Internet. Therefore, huge volumes of data are, at various speeds, which have to be stored in before analysis and process [5].

In this paper we extend the previous results published in [6], which considers a multi-objective optimization problem for storage service selection with budget constraints. The problem is as follow. We have different data sources, distributed geographically on a wide area, each of them producing data with different frequency. We need to store this data on cloud providers, then to process the data on datacenters. We solved the problem of providers allocation with flexible budget constraints and now we propose a continuous linear programming solution for data sources allocation to cloud providers. Our model considers that we already have the storage cost for each cloud provider and the transfer cost for each established link between data sources and storage spaces. Based on this model we can develop further an API-based service that compute in real-time these costs using our aggregated linear programming technique.

The paper is structured as follows. Section II presents similar approaches on cost-aware cloud resource allocation. Section III describes the problem model and our solution. Finally, Section IV presents the results obtained using an iterative linear programming solver. The paper ends with conclusions.

II. Related Work

The problem of cost-optimal Cloud storage service selection is very recent and interesting research topic. Many authors proposed different approaches in order to solve this problem.

The authors of [7] propose a selection strategy that use fuzzy inference or Dempster-Shafer theory of evidence. This strategy is complemented by of a game theoretic approach in order to promote truth-telling ones among service providers. The effectiveness of the solution is demonstrated with empirical evidence through properly crafted simulation experiments.

In [8] an algorithm that can select optimal provider subset for data placement among a set of providers in a multi-cloud storage architecture based on IDA is presented. This is designed to achieve good trade-off among storage cost, algorithm cost, vendor lock-in, transmission performance and data availability. Using parameters from cloud providers the authors demonstrate that it is efficient and accurate to find optimal solutions in reasonable amount of time.

In [9] the authors propose a data distribution service that can identify target cloud providers with sets of resources that are capable of hosting the data payload when we need to migrate data from a host cloud network to a target cloud provider in order to leverage cost, security, redundancy, consolidation, or other advantages. The data distribution service can receive the data payload from the host cloud network, and transport the data payload to a selected target cloud provider via the set of dedicated communication channels.

The authors of [10] introduce a predictive approach to identify the cloud availability zone that maximizes satisfaction of an incoming request against a set of requirements. The prediction models are built from historical usage data for each availability zone and are updated as the nature of the zones and requests change. Simulation results show that their method successfully predicts the unpublished zone behavior from historical data and identifies the availability zone that maximizes user satisfaction against specific requirements.

Another work that deals with cost-aware selection of cloud storage services in presented in [11]. The authors present a system, method and computer program product for allocating shared resources. Upon receiving requests for resources, the cost of bundling software in a virtual machine (VM) image is automatically generated. Software is selected by the cost for each bundle according to the time required to install it where required, offset by the time to uninstall it where not required. A number of VM images having the highest software bundle value (i.e., highest cost bundled) is selected and stored, e.g., in a machine image store. With subsequent requests for resources, VMs may be instantiated from one or more stored VM images and, further, stored images may be updated selectively updated with new images.

III. Problem Model and Proposed Solution

In our previous work [6] we introduced the following optimization problem, under a total budget constraint B and a request Req:

$$\begin{aligned} \min_j \left\{ \sum_{i=1}^{n} A(i,j) \times c_{ij} \times Data(L_i) \right\}, \quad j = 1 \dots m; \\ \min \left\{ \sum_{j=1}^{m} D_j \times c_{jD} \times Data(C_j) \right\}; \\ \sum_{j=1}^{m} b_j \times Data(C_j) \leq B. \end{aligned} \tag{1}$$

with the following bounds:

$$\begin{aligned} Data_{min} \leq Data(L_i) \leq Data_{max}, \quad i = 1 \dots n; \\ \sum_{j=1}^{m} D_j \times Data(C_j) \leq DataCapacity(D). \end{aligned} \tag{2}$$

were: n is the number of data locations - a set of geographically distributed sources of data; L_i is a data location having $Data(L_i)$ amount of data, $i = 1 \dots n$; m is the number of Cloud storage providers that can be accessed from any data location, each provider C_j, $j = 1 \dots m$, being able to store an large amount of data ($DataCapacity(C_j) \gg \sum_i Data(L_i)$); D is a processing datacenter that has the capability to process any amount of data, collected from Clouds; $c_{ij} = cost(L_i, C_j)$ represents the transfer cost of data from location L_i to a Cloud storage provider C_j (for example latency in *ms* or transfer price in EUR/GB); $c_{jD} = cost(C_j, D)$ represents the transfer cost of data stored in a Cloud location C_j to the datacenter where the data will be processed (similar costs like c_{ij}); $b_j = cost(C_j)$ represents the budget needed to store data for a specific Cloud storage provider C_j (represented by price in EUR/GB); A is the *assignment binary matrix* with the following mean: $A(i,j) = 1$ if all amount of data from location L_i, denoted by $Data(L_i)$ is stored on Cloud provider C_j. The request Req is a n elements binary array that specifies a data processing request; if $Req_j = 1$ then all data from Cloud provider C_j will be transfered to the datacenter D.

The quantity $Data(C_j)$ used in (1) represents the all amount of data gathered from one or many geographical locations by a Cloud storage provider C_j, after a feasible assignment is computed: $Data(C_j) = \sum_{i=1}^{n} A(i,j) \times Data(L_i)$.

Now, according with (2), we formulate a new optimization problem from (1), as follow:

$$\begin{aligned} \forall j = 1 \dots m \\ \sum_{i=1}^{n} A(i,j) \times Data(L_i) = Data(C_j); \\ \min_j \left\{ \sum_{i=1}^{n} A(i,j) \times c_{ij} \times Data(L_i) \right\}. \end{aligned} \tag{3}$$

and we need to solve n linear optimization problems to find all elements of assignment matrix.

A. Continuous Linear Programming Solver

The solution for problem (3) is to compute in n iterations the global allocation, by solving in each step an optimization problem. We have the following parameters: c is a matrix containing the objective function coefficients (each column contains the transfer cost from a specific data source to all Cloud Providers - see Table I);
a is an array containing the constraints coefficients (storage costs); B is a number containing the right-hand side value for each iteration (constant for all iterations); lb is an array containing the lower bound on each of the variables (the default lower bound is zero); ub is an array containing the

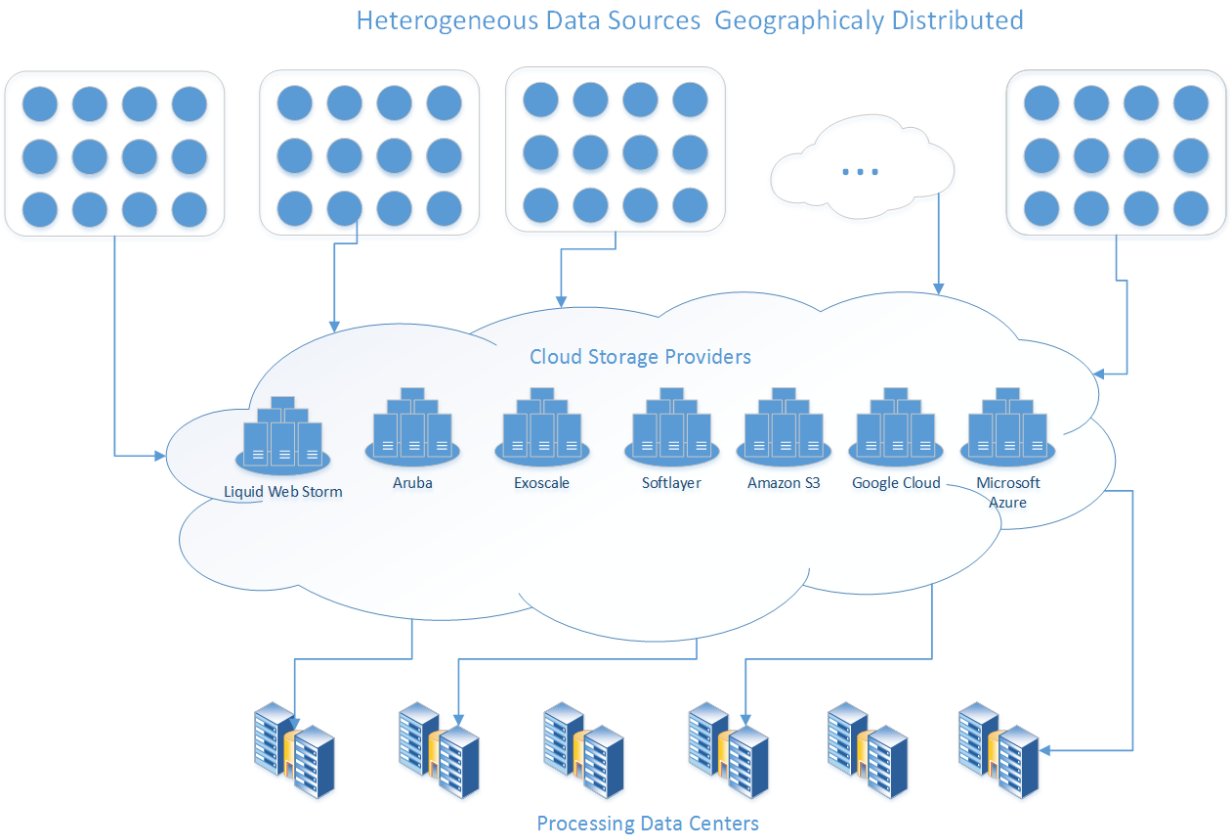

Fig. 1. The model used for data processing: data sources (on the top), seven public Cloud Storage providers and several datacenters.

TABLE I
TRANSFER COST MATRIX (c).

Cloud Provider	ds1	ds2	ds3	ds4
cp1	0.0000	0.2292	0.0500	0.0791
cp2	0.1097	0.0000	0.2292	0.0500
cp3	0.0182	0.1097	0.0000	0.2292
cp4	0.0000	0.0182	0.1097	0.0000
cp5	0.0791	0.0000	0.0182	0.1097
cp6	0.0500	0.0791	0.0000	0.0182
cp7	0.2292	0.0500	0.0791	0.0000

upper bound on each of the variables (we consider the values from Table II);

TABLE II
CLOUD PROVIDERS CAPACITY ($Data(C_j)$ - ub).

Cloud Provider	Capacity (GB)
cp1	5.000
cp2	5.000
cp3	5.000
cp4	3.000
cp5	2.000
cp6	1.000
cp7	1.000

$CTYPE$ represents the sense of constraint ("U" means an inequality constraint with an upper bound); $VARTYPE$ encodes a continuous variable; $itlim$ is the simplex iterations limit (it is decreased by one each time when one simplex iteration has been performed, and reaching zero value signals the solver to stop the search); $msglev$ specify that error and warning messages can be displayed during the solver run. Finally, If $SENSE$ is 1, the problem is a minimization and if $SENSE$ is -1, the problem is a maximization (we want to maximize the amount of stored data having a specific budget).

The proposed solver is described in the following listing. We used GNU Linear Programming Kit package [12].

```
% n - number of data locations
% m - number of Cloud Storage Providers
[m n] = size(c);

% seting the parameters of the
% optimization problem
ctype = "U";
vartype = "CCCCCCC";
s = -1;
param.msglev = 1;
param.itlim = 1000;

for i = 1 : n
[xopt(:,j), fopt(j), status, extra] =
glpk (c(:,j), a, b, lb, ub,
```

```
CTYPE, VARTYPE, SENSE,
param);
end
```

In this listing, $XOPT$ is the optimizer (the value of the decision variables at the optimum) and $FOPT$ is the optimum value of the objective function. We used GNU Linear Programming Kit solver software

B. Data Gathering Model

In Fig. 1 we represent the vision of our model: we have multiple heterogeneous data sources geographically distributed that are connected with a set of Cloud storage providers. Furthermore, the Cloud providers have link with different processing data center and periodically send data to be processed. Additionally we know the cost of transfer from the data source to the cloud provider. So we aim to select the best Cloud providers to sent data such that the cost to be minimum and to respect he budget.

TABLE III
STORAGE COSTS FOR CLOUD PROVIDERS.

Cloud Provider	Name of Cloud Provider	Cost[1] (EUR/GB/month)
cp1	Microsoft Azure Object Storage us-central	0.0220
cp2	Google Cloud Storage eu	0.0237
cp3	Amazon S3 ap-northeast-1	0.0274
cp4	SoftLayer Object Storage AMS	0.0366
cp5	Exoscale Object Storage CH-GV2	0.0517
cp6	Aruba Cloud Storage R1-CZ	0.0790
cp7	Liquid Web Storm Object Storage us-central	0.0914

C. API-based Cost Model

Table III presents the storage costs for different Cloud Providers. Also, Table I contains the transfer costs. All of these numerical values can be obtained using the following API-based cost model. Every Cloud storage provider offers publicly the price for available services. In order to obtain prices we can implement a web service that use the APIs offered by the Cloud Providers. A method for prices comparison is offered by Cloudorado which find the best solution for your specific requirements (https://www.cloudorado.com).

IV. RESULTS

We run the optimizer solver with n iterations considering two value for the available budget. I the first case ($B = 100$), only the first 3 Cloud Providers were selected, all of them having the lowest price (see Figure 2). The second case ($B = 500$) selects all Cloud Providers, even the $cp4, \ldots cp7$ are more expensive. These are selected because $cp1$, $cp2$ and $cp3$ reached their maximum capacity (see Table II). As a conclusion, the optimum value of the objective function increases two times even the total available budget was increased by five times.

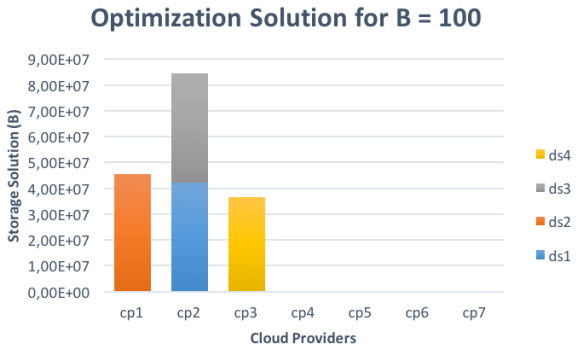

Fig. 2. $XOPT$ - Cloud Provider Allocation for a $B = 100$. Only chipper Cloud Providers are used; $sum(FOPT) = 3308, 3$.

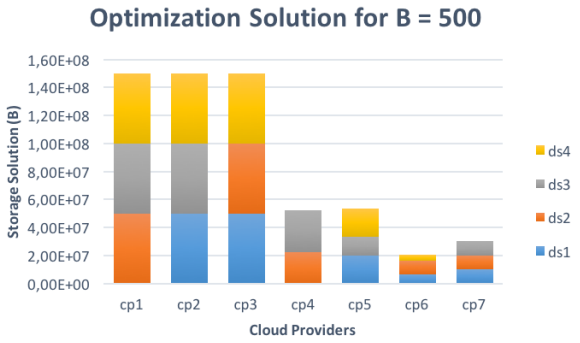

Fig. 3. $XOPT$ - Cloud Provider Allocation for a $B = 500$. All Cloud Providers are used; $sum(FOPT) = 6770, 0$.

V. CONCLUSION

Service selection problem in multi-Cloud environments is a difficult and current research problem. The cost reduction should be the main benefit. In this paper we presented an aggregated linear programming technique in order to find a cost efficient solution for the service selection.

The results show that the allocation of cloud providers depends on transfer costs and on the available budget, can be selected only few or all Cloud providers. The proposed method satisfy the data demand.

The work presented in this paper can be applied in CyberWater research project (http://cw.hpc.pub.ro). The project aims to build a platform using advanced computational and communication technology for management of water resources. The main activity is to gather diverse data from various heterogeneous sources (e.g. sensors, WWW, third party institutions, etc.) in a common digital platform in order to provide assistance in critical situations (e.g. accidental pollution) [13]. The proposed solution will allocate the appropriate Cloud Providers to store the maximum amount of data gathered from hydrological sensors for a given budget.

ACKNOWLEDGMENT

The research presented in this paper is supported by projects: *CyberWater* grant of the Romanian National Authority for Scientific Research, CNDI-UEFISCDI, project number 47/2012; *clueFarm*: Information system based on cloud services accessible through mobile devices, to increase product quality and business development farms - PN-II-PT-PCCA-2013-4-0870; *DataWay*: Real-time Data Processing Platform for Smart Cities: Making sense of Big Data - PN-II-RU-TE-2014-4-2731.

REFERENCES

[1] C. Lynch, "Big data: How do your data grow?" *Nature*, vol. 455, no. 7209, pp. 28–29, 2008.

[2] A. SZALA, "Science in an exponential world," *Nature*, vol. 440, p. 2020, 2006.

[3] T. Economist, "The data deluge," February 2010. [Online]. Available: http://www.economist.com/node/15579717

[4] C. Negru and V. Cristea, "Cost models–pillars for efficient cloud computing: position paper," *International Journal of Intelligent Systems Technologies and Applications*, vol. 12, no. 1, pp. 28–38, 2013.

[5] B. Palanisamy, A. Singh, L. Liu, and B. Jain, "Purlieus: locality-aware resource allocation for mapreduce in a cloud," in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, 2011, p. 58.

[6] C. Negru, F. Pop, O. C. Marcu, M. Mocanu, and V. Cristea, "Budget constrained selection of cloud storage services for advanced processing in datacenters," in *RoEduNet International Conference-Networking in Education and Research (RoEduNet NER), 2015 14th*. IEEE, 2015, pp. 158–162.

[7] C. Esposito, M. Ficco, F. Palmieri, and A. Castiglione, "Smart cloud storage service selection based on fuzzy logic, theory of evidence and game theory," 2015.

[8] W. Yao and L. Lu, "A selection algorithm of service providers for optimized data placement in multi-cloud storage environment," in *Intelligent Computation in Big Data Era*. Springer, 2015, pp. 81–92.

[9] J. M. Ferris, "Migrating data among cloud-based storage networks via a data distribution service," Mar. 17 2015, uS Patent 8,984,269.

[10] M. Unuvar, S. Tosi, Y. N. Doganata, M. G. Steinder, and A. N. Tantawi, "Selecting optimum cloud availability zones by learning user satisfaction levels," *Services Computing, IEEE Transactions on*, vol. 8, no. 2, pp. 199–211, 2015.

[11] M. D. De Assuncao, M. A. S. Netto, L. Renganarayana, and C. C. Young, "System, method and program product for cost-aware selection of stored virtual machine images for subsequent use," May 19 2015, uS Patent 9,038,085.

[12] A. Makhorin, "Glpk (gnu linear programming kit), version 4.42," *URL http://www. gnu. org/software/glpk*, 2004.

[13] C. Negru, M. Mocanu, C. Chiru, A. Draghia, and R. Drobot, "Cost efficient cloud-based service oriented architecture for water pollution prediction," in *Intelligent Computer Communication and Processing (ICCP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 417–423.