Minimizing Cost of Provisioning in Fault-tolerant Distributed Data Centers with Durability Constraints

Rakesh Tripathi, S. Vignesh and Venkatesh Tamarapalli
Department of Computer Science and Engineering, IIT Guwahati, Assam, India-781039
{t.rakesh, s.vignesh, t.venkat}@iitg.ernet.in

***Abstract*—Many popular e-commerce applications run on geo-distributed data centers requiring high availability. Fault-tolerant distributed data centers are designed by provisioning spare compute capacity to support the load of failed data center, apart from ensuring data durability. The main challenge during the planning phase is how to provision spare capacity such that the total cost of ownership (TCO) is minimized. While the literature handled spare capacity provisioning by minimizing the number of servers, variation in electricity cost and PUE corroborate the need to minimize the operating cost for capacity provisioning. We develop an MILP model for spare capacity provisioning for geo-distributed data centers with durability requirements. We consider spare capacity provisioning problem with the objective of minimizing TCO. We model variation in the demand, fluctuation in electricity prices across locations, cost of state replication, carbon tax across different countries, and delay constraints while formulating the optimization model. Solving the model shows that TCO is reduced while leveraging the electricity price variation and demand multiplexing. The proposed model outperforms the CDN model by 50% and the minimum server model by 34%. Results also demonstrate the effect of power usage effectiveness (PUE), latency, number of data centers and demand on the TCO.**

I. Introduction

Many popular cloud services, web services, e-commerce applications and other large-scale applications are deployed over geo-distributed data centers. A geo-distributed data center is simply a collection of networked data centers interconnected via high capacity WAN links. See Fig. 1 for an illustration. The advantages of geo-distributed data centers are typically: increased data center availability and reliability, reduced access time for users across the globe and possibility of horizontal scale out against capacity constraints (electricity, physical space etc.). These advantages are driving several cloud service providers to move to geo-distributed data centers (*henceforth simply referred as distributed data centers*). For example Google is spread across 15 counties and has more than 30 data centers with an estimated 900,000 servers [1]. Akamai has nearly hundred thousand servers in over 1,900 networks across 71 countries.

However, business applications like e-commerce websites demand high availability as their downtime translates directly to lost revenue, lost productivity, and reduced customer satisfaction. A survey conducted by Gartner [2] on estimated financial cost of downtime suggests that a loss of revenue in the range of $250,000-$500,000 was incurred by 60% companies for an hour of downtime and one sixth of the organizations incurred a loss of $1 million or more. Data center unavailability has been reported by many cloud service providers like Amazon, Facebook and Google [3]–[5]. This could be due to reasons like building fire, power outage, human error, software bug, ISP router misconfiguration, and other man-made or natural disasters. Similarly, EMC survey 2013 [6], reported that 67% of IT managers and professionals give highest priority to storage technology for backup, recovery and archiving in their organizations. High availability and data durability to protect data destruction at the primary site, are ensured by state replication to a remote site, preferably to a nearest location. In this paper failure of single data center is the only kind of failure and state replication to nearest data center has been considered.

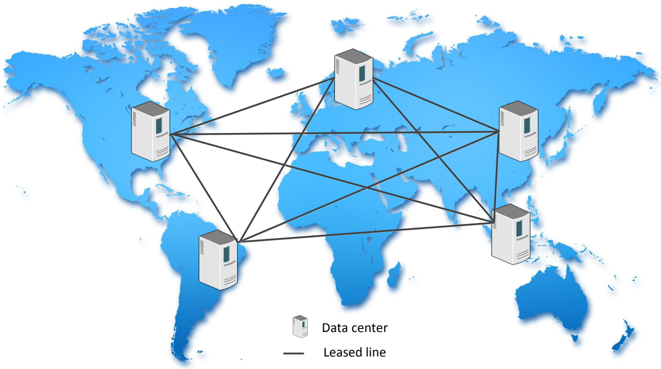

Fig. 1: Illustration of geo-distributed data center

Designing a fault-tolerant, durable, distributed data center involves spare capacity provisioning across data centers (allocation of additional servers and state replication to mask failures).There is an immediate question of where this spare capacity needs to be allocated. We call this problem as cost-aware capacity provisioning (CACP) problem, satisfying a set of constraints based on electricity prices, infrastructure cost, demand at each location, and latency constraints dictated by clients. A straight-forward approach for spare capacity provisioning could be to simply provision additional capacity uniformly across data centers lying within the latency bounds of the concerned client region. However, the main challenge in designing the fault tolerant distributed data center is in minimizing the cost of operating services over distributed data center reaping the benefit of variation in the carbon tax, electricity cost, and bandwidth cost across space and time. It may be noted that CACP problem is a variant of facility location problem proved to be NP hard [7].

Motivation: The work in this paper is motivated by the following observations from the literature:

- *High inter-data center communication cost:* Usually, cloud service providers connect their data centers with dedicated WAN links which are significantly expensive. For example, a cost of \$1 to transfer 2.7 GB of data over 100km was reported in [8]. Therefore, informed state replication must be carried out in order to minimize the operating cost involved.
- *Multiplexing of resources:* In all Internet-scale applications, client demand varies over time of the day. This variation can be exploited by cloud providers by multiplexing the demand on compute servers [9].
- *Electricity price variations:* In current multi-electricity market, electricity price varies across space and time. Electricity cost is also slowly exceeding the cost of data center equipment. For example, let us assume life of a server to be 4 years and the cost of a server to be \$2000 [10]. We calculate the energy to acquisition cost (EAC) defined to be the ratio of cost of running a server for 4 years to its acquisition cost, as given below in Eqs. (1) and (2).

$$\begin{aligned}\text{Power cost} &= 4 \text{ years} * (8760 \text{ hours/year}) \\ &\quad * (\text{electricity cost}) * \text{server power} * \text{PUE}\end{aligned} \tag{1}$$

$$\text{EAC} = \frac{\text{power cost}}{\text{server cost}} * 100 \tag{2}$$

We used the electricity prices given in Table I,obtained from [11], peak power used per server as 400W and an optimistic power usage effectiveness (PUE) value of 1.2, to calculate the EAC in different countries. The EAC values in Table I indicate that for most of the countries cost of power and cooling exceeds the cost of buying servers which suggests that a greater attention should be put on optimizing data center power consumption cost instead of only minimizing the servers while designing fault-tolerant data centers.

Country	Electricity price(\$/kWh)	Operating Cost(in \$)	EAC
Belgium	0.29	4877	243%
Brazil	0.16	2691	134%
Canada	0.11	1850	92%
France	0.19	3261	163%
Germany	0.32	5382	269%
Hong Kong	0.18	3027	151%
Ireland	0.28	4769	238%
Italy	0.28	4774	238%
Japan	0.22	3700	185%
Netherlands	0.29	4857	243%
Russia	0.08	1345	67%
Singapore	0.28	3920	196%
Switzerland	0.25	4204	210%
UK	0.25	3363	168%
USA	0.12	2018	101%

TABLE I: EAC for different countries

Therefore, we propose an optimization framework for cost-aware capacity provisioning in distributed data centers such that, apart from meeting the customer demand after a single data center fails, it is still cost-effective. The main contributions of this paper are as follows:

- We develop an MILP model for cost-aware capacity provisioning (CACP) with the objective of minimizing the TCO, subject to latency and availability constraints (i.e. to mask failure of single data center). Along with the server acquisition cost we also consider electricity cost, state replication cost and carbon tax.
- By solving the optimization model using real-world data, we demonstrate that CACP model has greater potential of reducing the TCO by exploiting the variation in electricity prices, bandwidth prices and client demand.
- We demonstrate that the CACP model has 50% improvement over the model which minimizes the average response time by routing to nearest data center (CDN model) and 35% over the model that minimizes total number of servers (MS model).
- We also demonstrate the impact of number of data centers, customer latency requirements,PUE, and variation in client demand on the TCO with the help of the proposed optimization model.

The rest of the paper is organized as follows. Section II discusses the work related to capacity provisioning in distributed data centers. Section III discusses the proposed CACP model and list the other models used in comparison. The advantages of the proposed CACP model over the existing models in minimizing the TCO are showed in Section IV. Section V concludes the paper.

II. Related Work

Data center placement and capacity provisioning has been addressed previously in both industry white papers and research papers. A few studies addressed data center placement, expansion and capacity provisioning based on MILP optimization which considers capital cost and operating cost. Their objective functions span across minimizing total cost, minimizing total carbon footprint, and minimizing average service latency abiding by the QoS constraints [8], [10], [12].

The work in [13] presented an approach for solving placement, capacity provisioning and request routing jointly. The objective was to minimize the total cost of building a new or expanding existing data center, subject to latency constraint on every routing path bounded by a maximum value. The authors of [14] proposed general guidelines to design a disaster resilient data center which described quality of resilience metrics like service availability and time to recovery and discussed how these affected the decision of selecting data center recovery mechanism, site placement and topology design, VM placement and backup strategies.

Although the problem of data center placement and capacity provisioning has been addressed in most articles, failure of data centers which affects the revenue of cloud service providers running business critical applications has been ignored. The closest work to our study of capacity provisioning in fault tolerant distributed data center is that in [9], where the authors designed a simple optimization model to minimize the total number of servers required across all the data centers such that latency constraints and availability constraints (masking failure of single data center) are met. However, it has been observed that server operating cost is comparable to or in some countries even dominate the server acquisition cost. Therefore, we use minimization of the TCO as an objective to minimize total cost (operating cost and server acquisition) of spare capacity provisioning. Our model reaps the benefit of electricity and bandwidth price variation and demand multiplexing (for diurnal applications) to minimize the TCO.

III. Optimization Model

In this section, we formulate the problem of cost aware spare capacity provisioning (CACP) as an MILP model. We first state the assumptions made in building the model, followed by system model used, and the present the optimization model. Table II summarizes the notation used in the model with the definitions.

Variable	Meaning
Input Parameters	
s	Data center location
u	Client region
P_s^{fh}	Power consumed at data center s during hour h with failed data center f
$P_s^{h\ max}$	Maximum power available at data center s during hour h
γ_s^{fh}	Average server utilization at data center s during hour h and failed data center f
γ^{max}	Maximum value of γ to avoid waiting
L_u^h	Total number of requests generated from user location u during hour h
D_{su}	Propagation delay between client region u and data center s
D_{max}	The maximum tolerable latency
θ_s^h	Electricity price per kWh at data center s at hour h
ρ_s	Transmission loss of electricity at data center s
α	Server acquisition cost
δ_s	Carbon tax at data center s
M^{min}	Minimum number of servers at any data center
M^{max}	Maximum number of servers at any data center
ν_s	Bandwidth cost for state replication of data center s, to the nearest data center (this cost is a known constant value, for a given data center)
ξ	Number of bytes required for state replication of single request
Decision Variables	
m_s	Number of servers in data center s
λ_{su}^{fh}	Number of requests from user location u, served by data center s during hour h and failed data center f
y_{su}	Binary variable that denotes whether client location u lies within the latency bound of data center s
Cost Components	
F	Total cost of ownership, including server acquisition cost, operating cost and state replication cost
ψ	Operating cost and server acquisition cost
η	Cost of state replication to nearest data center for durability

TABLE II: Summary of notation used in the paper

A. Assumptions

The following assumptions were made in the model.

- Failure of only single data center (a site) is considered. Failure of more than one data center at the same time is assumed to be unlikely because, two sites do not share a common resource group by choice of locations.
- Though the failure of a data center is inevitable, its probability is often observed to be low. Hence, no failure case ($f = 0$) has a probability of 0.95 and any one of the data centers $f \in \{1, |S|\}$ can fail uniformly at random with probability 0.05, where S is the set of data center locations [15].
- Mechanism for failure detection and request re-routing is already in place. We can use any state-of-art approach similar to the one in [16].
- Each pair of data center sites are connected by dedicated link and it is charged based on actual usage over a billing cycle.
- Client demand at a location is proportional to its population. Propagation delay within the client region is assumed to be negligible.

B. System Model

Cost: Let S and U denote the set of data centers and client locations, respectively. The cost of server (acquisition cost) be α. Since the electricity cost usually varies across time and space, θ_s^h denotes the electricity cost at data center location $s \in S$ during the hour h of the day. Let δ_s be the carbon tax levied at data center location $s \in S$ and ν_s be the bandwidth cost for state replication from data center s, to its nearest data center (this cost is a assumed to be constant for a pair).

Demand: Let λ_{su}^{fh} denote the number of requests from client region u served by data center at location s, during hour h when data center indexed $f \in F = \{0, |S|\}$ has failed. Let L_u^h be the total demand for user location u at hour h.

Delay: Let D_{max} be the maximum latency allowed for a client based on the service level agreements with the cloud provider. Let D_{su} be the propagation delay between user location u and data center location s. The data center must be designed such that even after the failure of a site, the latency continues to be lower than D_{max}.

State replication: For every request served by a data center s, let ξ be the size of data that needs to be replicated to its nearest data center and ν_s be the bandwidth cost to the nearest data center of s. The cost of state replication denoted by η is modeled as

$$\eta = \sum_{f,u,s,h} \left(\lambda_{su}^{fh}\ \xi\ \nu_s \right) \tag{3}$$

Server Provisioning: Let m_s denote the number of servers required in a data center at s. We define M^{min} and M^{max} to be the minimum and maximum number of servers that can be provisioned at any data center due to limitations associated with space and power.

Power Consumption: Let P_{idle} be the average power drawn in idle condition and P_{peak} be the power consumed when server is running at peak utilization. Then total power consumed at a data center location $s \in S$, at hour $h \in H$ is given by [8]:

$$\begin{aligned} P_s^{fh} \ = \ & m_s(P_{idle} \ + \ (E_s \ - \ 1)P_{peak}) \\ & + \ m_s(P_{peak} \ - \ P_{idle})\gamma_s^{fh} \ + \ \epsilon \end{aligned} \tag{4}$$

where E_s is the PUE of a data center at s, ϵ is an empirical constant, and γ_s^{fh} is the average server utilization defined by

$$\gamma_s^{fh} \ = \ \frac{\sum_{u \in U} \lambda_{su}^{fh}}{m_s \mu} \tag{5}$$

where μ is the service rate of the server.

C. CACP Model

For notational convenience, we define the following variables:
$\mathbf{m} \triangleq [m_s, \forall s \in S]$,
$\boldsymbol{\lambda} \triangleq [\lambda_{su}^{fu}, \forall s \in S,\ \forall u \in U,\ \forall h \in H, \forall f \in F]$ and
$\boldsymbol{y} \triangleq [y_{su}, \forall s \in S,\ \forall u \in U]$

We define the TCO, denoted by $\mathcal{F}$, to be the sum of state replication cost η (defined in Eq. 3), and the sum of operating cost and server acquisition cost, denoted by Ψ, defined as

$$\Psi = \sum_{s \in S} \Big(\sum_{h \in H} \Big(\theta_s^h P_s^{fh} \ + \delta_s(\rho_s \ + \ 1) P_s^{fh} \Big) + m_s \alpha \Big) \tag{6}$$

With these cost factors the CACP problem is expressed by the following optimization model.

$$\underset{\boldsymbol{m},\boldsymbol{\lambda},\boldsymbol{y}}{\textbf{minimize}} \ \ \mathcal{F} = \Psi + \eta \tag{7}$$

subject to,

$$\sum_{s \in S} \lambda_{su}^{fh} \ = \ L_u^h \qquad \forall u, h, f \tag{8}$$

$$0 \ \leq \ \lambda_{su}^{fh} \ \leq \ y_{su} L_u^h \qquad \forall s, u, h, f \tag{9}$$

$$M^{min} \ \leq m_s \ \leq M^{max} \qquad \forall s \tag{10}$$

$$P_s^{fh} \leq P_s^{h \ max} \qquad \forall s, h \tag{11}$$

$$2 D_{su} \, y_{su} \leq D_{max} \tag{12}$$

$$\gamma_s^{fh} \ \leq \ \gamma^{max} \qquad \forall s, h \tag{13}$$

$$y_{su} \in \{0, 1\} \qquad \forall s, u \tag{14}$$

Among the constraints, Eq. (8) ensures that demand of all client regions in every hour is met. Eq. (9) ensures that all the client requests are served by data centers within the latency limit. Eq. (10) ensures that capacity limit of a data center (in terms of number of servers) is not exceeded. There is also a constraint on the total power available at a data center which is taken care by Eq. (11). Eq. (12) ensures that the delay experienced by a client lies within the maximum delay bound. Since our focus is only on network latency [9], in order to ensure that the service time is independent of the data center used, we capped the server utilization by Eq. (13).

D. Existing Models

We present two models from the literature with different objectives that are used in data center planning. We compare the proposed model with these in the next section.

- **CDN Model:** In principle, CDN's objective is to balance load such that average response time is minimized. By defining the average response time as

$$\Delta = \frac{\sum_{f,u,s,h} \lambda_{su}^{fh} D_{su}}{\sum_{f,u,h} L_u^h} \tag{15}$$

  The CDN model for fault-tolerant operation is

$$\begin{aligned} &\underset{\boldsymbol{m},\boldsymbol{\lambda},\boldsymbol{y}}{\textbf{minimize}} \ \ \Delta \\ &\textbf{subject to eqs.} \ \ (8) \text{ - } (14) \end{aligned} \tag{16}$$

  This model is designed to route the requests after the failure to the nearest data center with the spare capacity. It does not consider any other operating cost factors to optimize the TCO.
- **MS Model:** This model has been defined in [9]. The main objective of this model is to minimize the total number of servers deployed across all the data centers with availability and latency constraints. This model also does not consider any cost factors associated with the operation of the data centers.

IV. Numerical Results

In this section, we evaluate the proposed CACP model and the two existing models (MS and CDN) under different scenarios. The proposed optimization model is an MILP, solved using CPLEX with Matlab on a server with Intel Xeon processor and 64 GB of RAM, running Ubuntu 14.04 (64 bit) OS. To understand the advantage of considering the operating cost in spare capacity provisioning, we compared the cost of solutions obtained with the other models. We not only demonstrate that CACP model outperforms MS and CDN models in the reduction of the TCO but also, show the sensitivity of the models to the number of data centers, user demand, maximum latency bound and PUE.

A. Experimental Setup

Data center locations: Availability of (cheaper) power dictates the selection of locations for data centers on global scale. We consider data center locations around the world from the USA, Europe and Asia viz. Oregon, Pennsylvania, South Carolina, Chicago, U.K., Germany, France, Belgium, Japan, China, Hong Kong and Singapore. The minimum and maximum number of servers at each location are set to 1000 and 100,000, respectively.

Client locations: From the data in [17], we considered top 15 countries in terms of Internet usage as our client locations. These are Brazil, China, Egypt, France, Germany, India, Indonesia, Japan, Mexico, Nigeria, Russia, South Korea, UK, USA, and Vietnam. The propagation delay between data center and client location is considered to be linearly proportional to the distance between them, and it increases by 10 ms for every 1000 km [8].

Input parameters: For the electricity price θ_s^h, we considered the data reported in [18]. We used carbon tax of 0.49 cents/kWh [8] for the power being consumed. P_{idle} and P_{peak} are taken to be 400W and 200W, respectively [19]. Average PUE is set to 1.8 [20]. The empirical constant ϵ is assumed to be 0 [13]. P^{max} is taken to be 100MW/hr for all the locations [12]. The default value for maximum latency is taken to be 300ms. We considered the data size for replication per request to be 1KB [21] and inter-data center bandwidth cost (for state replication) as \$1 for 2.7 GB over 100km [8].

Traffic model: We assumed that the demand from a client region is proportional to the number of Internet users from that region [17]. Aggregated demand is a uniform random variable lying between 100,000 and 300,000 requests/second [22]. Table III shows the distribution of requests across various client regions. A day is divided into 12 off-peak, 6 mid-peak and 6 on-peak hours, to accommodate the behavior of diurnal applications.

B. Results

We present the results obtained from solving the models under different scenarios constructed by varying the number

Country	Brazil	China	Egypt	France	Germany	India	Indonesia	Japan	Mexico	Nigeria	Russia	S. Korea	UK	USA	Vietnam
% Demand	5.33%	31.76%	2.17%	2.78%	3.50%	15.43%	2.04%	5.64%	2.65%	3.37%	4.50%	2.13%	2.93%	13.69%	2.09%
Request Range (x10^3)	10-16	63-95	4-6	6-8	7-10	30-46	4-6	11-17	5-8	7-10	9-13	4-7	6-9	27-46	4-6

TABLE III: Demand variation across space and time

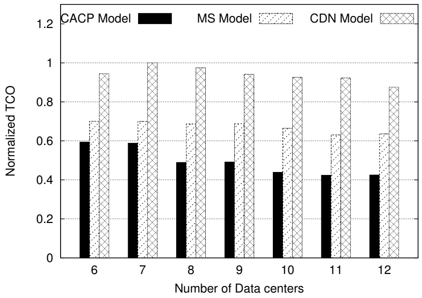

Fig. 2: Normalized TCO with varying number of data centers

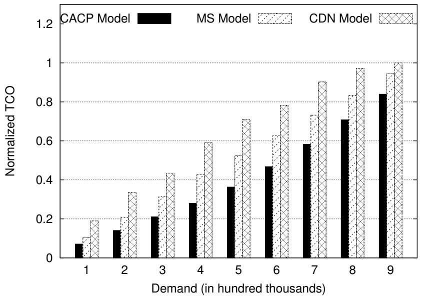

Fig. 4: Normalized TCO with varying demand

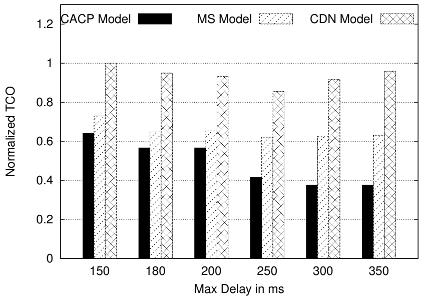

Fig. 3: Normalized TCO with varying maximum latency bound

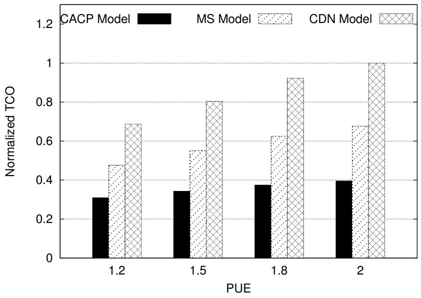

Fig. 5: Normalized TCO with varying PUE values

of data centers, the demand, latency bound, and PUE value. The metric used to compare the three optimizations models is the total cost of ownership as defined in Eq. 6. In each graph, we plotted the normalized values of TCO wherein, the normalization is done with respect to the maximum TCO across all the experiments.

- **TCO comparison** : In this experiment, we compared the TCO from the CACP model with that from the existing models. We increased the number of data centers between 6 and 12, with the total demand fixed at 300,000 requests/sec and maximum latency at 300 ms. For each case(fixed number of data centers), we solved the optimization models and computed the normalized TCO. Fig. 2, shows the normalized TCO as the number of data centers increases with the state being replicated only at the nearest data center.
  We can see that the TCO for CACP and MS model reduces with the number of data centers since, they take advantage of diurnal pattern in the demand resulting in better multiplexing. The reduction in the TCO with the CACP model is larger compared to that with the MS model. The CACP model also leverages the electricity price variation across space and time reducing the TCO further. We can observe that the CACP model achieved a TCO reduction of upto 34% compared to the MS model, and upto 50% compared to the CDN model.
- **Impact of Latency**: In this experiment we studied the impact of maximum latency bound on the TCO. We ran the models with 12 data centers, 15 client regions, an aggregate demand of 300,000 requests/second. The maximum latency was chosen from [150,180,200,250,300,350] sec. Fig. 3 shows the normalized TCO for the three models. We can see that CACP improved the TCO upto 55% and 40% compared to the CDN and MS models, respectively. In the CACP model, we get greater number of data centers capable of serving the requests from a particular client region, leading to better multiplexing of resource, which in turn reduces the TCO. Along with this, CACP model also selects the data centers in regions with lower electricity prices while meeting the latency bound. Although the CDN model gives minimum latency, request routing is oblivious to the variation in the electricity price and bandwidth price. Therefore, the TCO is higher with CDN model particularly when the latency requirements are not very stringent.
- **Impact of Demand:** We solved the optimization models on 12 data centers while varying the total demand with the maximum latency at 300ms. Results showed in 4 indicate that as the demand increases, the advantage of CACP model over the existing models decreases. Since, there is an upper bound on the capacity of a data center, increasing demand leads to saturation of all the data centers in the regions with cheaper electricity. This reduces the choice available for CACP model and leading to the selection of locations with higher electricity prices. We can conclude that the proposed model is advantageous only when the data center does not operate at peak utilization. Under heavy load, the CACP model can help the provider to determine an optimal data

center upgrade plan while minimizing the TCO.

- **Impact of PUE:** Finally, we studied the sensitivity of the models to data center PUE values. We considered 12 data centers serving 15 client regions, peak demand of 300,000 requests per second and maximum latency of 300ms. With an increasing PUE the power efficiency of the data center reduces, leading to greater power consumption and higher operating cost. Fig. 5 shows that CACP model significantly reduces the TCO for every PUE value considered. For a PUE value of 2 the reduction in the cost is 60% with respect to CDN model and is 40% with respect to the MS model. This suggests that cloud service providers should use CACP model when the PUE is large.

V. Conclusion

In this paper we consider the problem of spare capacity provisioning while planning fault tolerant and durable geo-distributed data centers, capable of masking single data center failures. In the CACP problem formulation, we minimized the TCO reaping the benefit of electricity cost and demand variation across the data centers. We used real world data, to show that the CACP model outperforms the existing models reported in the literature in minimizing the TCO. The CACP model achieves a cost reduction of upto 34% and 50% when compared to MS and CDN models, respectively. We also demonstrated that the CACP lowers the TCO when the variation in the electricity price and demand across the data centers is higher. We showed that the CACP model is promising for cloud providers when the load is high and the PUE values are high.

References

[1] P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav, "It's not easy being green," in *Proceedings of the ACM SIGCOMM 2012 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication*, ser. SIGCOMM '12, 2012, pp. 211–222.

[2] L. G. Christy Pettey, "Gartner Reveals Top Predictions for IT Organizations and Users for 2011 and Beyond," http://www.gartner.com/newsroom/id/1480514, 2010.

[3] A. AWS, "Summary of the Amazon EC2 and Amazon RDS Service Disruption in the US East Region," http://aws.amazon.com/message/65648/, 2011.

[4] R. Johnson, "More Details on Today's Outage," https://www.facebook.com/notes/facebookengineering/more-details-on-todays-outage/431441338919, 2010.

[5] R. M. D. W. T. Anderson, "Understanding BGP Misconfiguration," https://djw.cs.washington.edu/papers/sigcomm2002-misconfigs.pdf.

[6] "Managing storage trends, challenges and options,2013-2014," https://education.emc.com/content/_common/docs/articles/Managing_Storage_Trends_Challenges_and_Options_2013_2014.pdf.

[7] V. Verter, "Uncapacitated and Capacitated Facility Location Problems," in *Foundations of Location Analysis*, ser. International Series in Operations Research & Management Science, H. A. Eiselt and V. Marianov, Eds. Springer US, 2011, vol. 155, p. 2537. [Online]. Available: http://dx.doi.org/10.1007/978-1-4419-7572-0_2

[8] A.-H. Mohsenian-Rad and A. Leon-Garcia, "Energy information transmission tradeoff in green cloud computing," *Carbon*, vol. 100, 2010.

[9] I. Narayanan, A. Kansal, A. Sivasubramaniam, B. Urgaonkar, and S. Govindan, "Towards a leaner geo-distributed cloud infrastructure," in *6th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 14)*, Jun 2014.

[10] I. Goiri, K. Le, J. Guitart, J. Torres, and R. Bianchini, "Intelligent placement of datacenters for internet services," in *Distributed Computing Systems (ICDCS), 2011 31st International Conference on*, June 2011, pp. 131–142.

[11] "Electricity pricing," https://en.wikipedia.org/wiki/Electricity_pricing.

[12] M. Wardat, M. Al-Ayyoub, Y. Jararweh, and A. Khreishah, "To build or not to build? addressing the expansion strategies of cloud providers," in *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, Aug 2014, pp. 477–482.

[13] S. Chen, Y. Wang, and M. Pedram, "Concurrent placement, capacity provisioning, and request flow control for a distributed cloud infrastructure," in *Design, Automation and Test in Europe Conference and Exhibition (DATE), 2014*, March 2014, pp. 1–6.

[14] R. Souza Couto, S. Secci, M. Mitre Campista, and L. Kosmalski Costa, "Network design requirements for disaster resilience in iaas clouds," *Communications Magazine, IEEE*, vol. 52, no. 10, pp. 52–58, October 2014.

[15] A. Winokur, "Industry perspective: Remote data center sites for disaster recovery," *The Data Center Journal*.

[16] H. Xu and B. Li, "Joint request mapping and response routing for geo-distributed cloud services," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 854–862.

[17] "Internet world stats," http://www.internetworldstats.com/unitedstates.htm.

[18] "Electricity price in usa," http://www.eia.gov/.

[19] Y. Li, H. Wang, J. Dong, J. Li, and S. Cheng, "Operating cost reduction for distributed internet data centers," in *Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on*, May 2013, pp. 589–596.

[20] http://www.vertatique.com/no-one-can-agree-typical-pue.

[21] Z. Wu, M. Butkiewicz, D. Perkins, E. Katz-Bassett, and H. V. Madhyastha, "Spanstore: Cost-effective geo-replicated storage spanning multiple cloud services," in *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*. ACM, 2013, pp. 292–308.

[22] L. Rao, X. Liu, L. Xie, and W. Liu, "Minimizing electricity cost: optimization of distributed internet data centers in a multi-electricity-market environment," in *INFOCOM, 2010 Proceedings IEEE*, 2010, pp. 1–9.