Saha *et al. Journal of Cloud Computing: Advances, Systems and Applications* (2016) 5:1
DOI 10.1186/s13677-015-0050-8

Journal of Cloud Computing
a SpringerOpen Journal

RESEARCH Open Access

A novel revenue optimization model to address the operation and maintenance cost of a data center

Snehanshu Saha[1*], Jyotirmoy Sarkar[2], Avantika Dwivedi[1], Nandita Dwivedi[1], Anand M. Narasimhamurthy[3] and Ranjan Roy[4]

Abstract

Enterprises are enhancing investments in cloud services setting up data centers to meet growing demand. A typical investment is of the order of millions of dollars, infrastructure and recurring cost included. This paper proposes an algorithmic/analytical approach to address the issues of optimal utilization of the resources towards a feasible and profitable model. The economic sustainability of such a model is accomplished via Cobb-Douglas production function. The production model seeks to answer questions on maximal revenue given a set of budgetary constraints. The model suggests minimum investments needed to achieve target output.

Keywords: Cobb-Douglas production function, Data center, Concavity, Returns to scale, Profit function, Cost function

Motivation and background

IT operations are integral to most business organizations around the world. The business communities need to rely on the information systems to run their organizational operations. Therefore, a company may incur loss due to disruptions and unavailability of information systems. It is necessary to have an IT infrastructure, which houses all the information systems to minimize all kind of disruptions and obstacles related to information systems. This reliable IT infrastructure is called data center. The cost to run a data center is generally associated with power, cooling, networking and storage equipment. A data center houses thousand of information and computing systems deployed in computer racks. A rack is an Electronic Industries Association enclosure, which is 2 meters high, 0.61 meters wide and 0.76 meters deep. A standard rack accommodates 40-42 computing units and dense rack configuration servers (Blade rack) will accommodate 200 computing units. The heat dissipated by a standard rack is 10 KW and Blade rack will dissipate heat up to 30 KW. Hence, a data center containing 2000 racks may require 20 MW power [1]. Increasingly, PC-based computing and storage services are relocating to Internet services. While early Internet services were mostly informational, many recent Web applications offer services that previously resided in the client, including email, photo and video storage and office applications. The shift from PC-based computing services to server-side computing is driven primarily not only for the improvements in services, such as the ease of management (no configuration or backups needed) and ubiquity of access (a browser is all you need), but also by the advantages it offers to vendors. Now a days, Software as a service provides faster application development because it is easier for software vendors to make changes and improvements. Instead of updating millions of clients, vendors need to coordinate improvements and fixes inside their data centers and can restrict hardware deployment to a few well-tested configurations. Moreover, data center economics allows many application services to run at a low cost per user. For example, servers may be shared with thousands of active users (and many more inactive ones), resulting in better utilization. Similarly, the computation itself may become cheaper in a shared service (e.g., an email attachment received by multiple users can be stored once rather than many times). Finally, servers and storage in a data center can be easier to manage than the desktop or laptop

*Correspondence: snehanshusaha@pes.edu
[1] Center for Applied Mathematical Modeling and Simulation (CAMMS) & Department of Computer Science and Engineering, PESIT-BSC, Bangalore, 560100, India
Full list of author information is available at the end of the article

equivalent because they are under the control of a single, knowledgeable entity [2]. Though each data center is different based on the operations, facilities and the average cost per year to operate, a large data center costs between $10 million to $25 million. 42 % of costs are associated with hardware, software, uninterrupted power supplies, and networking. 58 % of the expenses is due to heating, air conditioning, property and sales tax. In a traditional data center, most of the cost is consumed by infrastructure for maintenance. A few surveys have estimated the maintenance cost up to 80 % of the total cost. As data centers have become important aspects in business organization, it is imperative to examine the cost-revenue dynamics and design an effective way to optimize it. Cisco in its Global Cloud Index (GCI) Reports-2012, forecasted data center traffic to shoot up to 554 exabytes (EB) per month by 2016, from 146 exabytes. There are a few techniques to optimize cost and profit. Cobb-Douglas is a widely used production model, but to the best of our knowledge, has never been used in the study of optimization issues arising in data centers.

Introduction & overview

The authors find it necessary to discuss different aspects of data centers and optimization, before proceeding to relevant scholarly work available in the public domain.

Data center key subsystems

A data center consists of many subsystems. Three main subsystems would be discussed here. These are often referred as 'Power, Ping, and Pong', required to run a data center [1].

1. Continuous power supply.
2. Air conditioning.
3. Network Connectivity.

Power is essential to provide uninterrupted services throughout the year. At large data centers, electricity is supplied either from a grid or from on-site generators. The electricity supplied to the computer racks is dissipated as heat. Therefore, a cooling system is required to mitigate the heat. Generally data centers have chillers to supply cold water, which is used for air conditioning. Network connectivity is necessary for data transmission within and outside of a data center. The power subsystem consists of a grid and backup generator. The network subsystems comprise of all the connectivity except rack switches, whereas cooling subsystem includes chiller and air conditioning system.

Traditional data center

Running a traditional data center is expensive in comparison to a cloud data center. Thousands of applications are running in traditional data centers along mixed hardware tool. Maintaining existing infrastructure consume the biggest chunk of the total cost and multiple management tools are required for operation and management. Lately, traditional data centers are being used by internet service providers for housing their own or third party servers. Traditionally data centers were either constructed for meeting the purpose of large organizations or Network-neutral data centers. These facilities establish interconnection of carriers and act as regional fiber hubs, providing services to local business in addition to hosting content servers.

Cloud data center

Cloud data center is a place, where 10,000 or more servers are hosted to provide services for applications. Consistent infrastructure components like racks, hardware, OS, networking etc. are generally used to build the cloud data center. One of the important features of cloud data center is that they are not remodeled traditional data centers. Salient features of cloud data centers are listed below:

- Constructed for serving different objectives.
- Built to a different scale.
- Created at a different time than the traditional data center.
- Unlike traditional data center, these are responsible for executing and managing different workload.
- Not constrained by limitations of traditional data centers.

The cost associated with cloud data centers are composed of three factors. Labor cost takes the smallest chunk of total operation cost, nearly 6 % of the total cost, whereas power and cooling cost and computing costs are 20 % and 48 % respectively. Other costs account for remaining 26 %. Cloud data centers add new cost, unlike traditional data centers [3].

Data center tiers

Data centers have been divided based on the destination available; each data center has been modeled for addressing specific business requirement and has operational problems and issues for various reasons:

- Data centers related to Corporate house.
- Data centers responsible for computer infrastructure as a service (IaaS) and hosting Web application.
- Data centers that provide services of TurnKey Solutions
- Data centers, where Web 2.0 has been implemented.

Data center optimization

Powerful operations, technology, and economic forces have converged to drive changes in enterprise data centers. From an operational standpoint, organizational

leadership thinks of the data center as a factory-like utility that collects and processes information. Top management understands the value of data that's available in real time to help inform and shape decision-making. They also expect the data center to be fast enough to adopt new, rapidly deployed, public facing and internal user applications. From a technology standpoint, data center of today must support mobility, provisioning on demand, scalability, virtualization and the flexibility to respond to fast-changing operational situations. From an economic standpoint, a few years of edgy fiscal conditions have imposed tight budgets on IT organizations in both the public and private sectors. That's the reason behind new thinking being introduced in data center modeling and management. Organizations expect maximum output for every dollar invested in IT. They also face pressure to reduce power usage as a component of overall organizational strategies for reducing their carbon footprint [4]. The challenges any data center faces in power consumption are listed below

- Lack of Access to Basic Data: Accurate data is required for achieving short-term and long-term capacity planning to ensure organizations don't overinvest or underinvest in power and cooling infrastructure. Specifically, IT industries need the ability to keep track of individual and aggregated server power usage and temperature data at any point of time and historical trends as well.
- Inefficient Use of Power: Server power is typically over allocated and racks are under populated in worst-case. This may create a situation where power infrastructure is inefficiently used and contributes to overinvestment in racks and power capacity. IT solutions need to figure out a mechanism to control power consumption below theoretical peak values so that kilowatt capacity of each rack may be fully utilized.
- Power and Cooling Excursions: The availability of services during business-critical time is a top priority though power supply disruption or cooling tower failures can affect business operations massively. IT must be able to survive such failures to avoid downtime and deliver reliably on service level agreements (SLAs).
- Higher density endowed computing environments are more efficient but can lead to localized hot spots during periods of peak utilization. Organizations need to identify and mitigate hot spots and optimize workload placement based on power and cooling availability and efficiency [5].

Rising energy costs and environmental responsibility have placed the data center industry under increasing pressure to improve its operational efficiency. According to Koomey, data center consumed 1.3 % of the global energy usage in 2010 [6]. At this scale, even relatively small efficiency improvements will result in significant cost savings and prevent millions of tons of carbon emissions. Google and other major Internet companies have made significant contributions towards improving the data center efficiency. The overall pace of PUE reduction has slowed down,given diminishing returns and the limitations of existing cooling technology [7]. Furthermore, best practice techniques such as hot air containment, water-de economization and extensive monitoring are now commonplace in largescale data centers [8]. Figure 1 shows Google's PUE performance track record from an annualized fleetwide PUE of 1.21 in 2008 to 1.12 in 2013, due to the implementation of best practices and natural progression down the learning curve [9]. Note the asymptotic decline of the trailing twelvemonth (TTM) PUE graph [10].

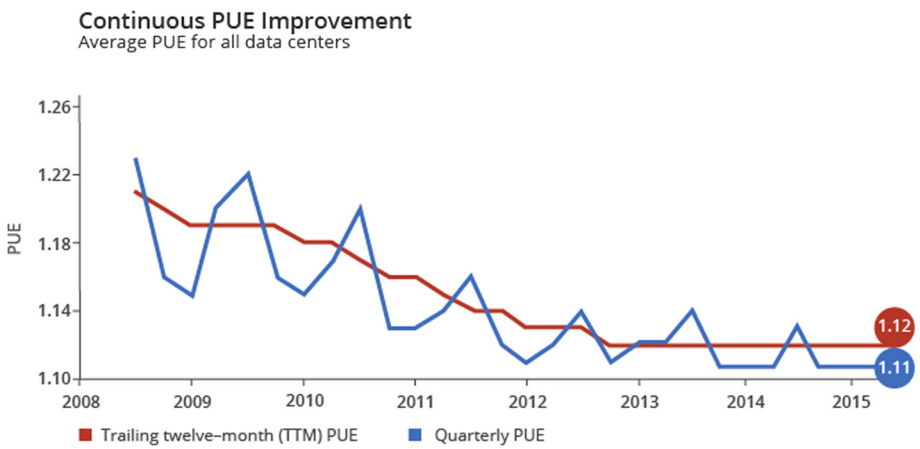

Fig. 1 PUE performance track record

In this paper, we propose a revenue model, which is best suited for the data centers in the current scenario. Mathematically, we have established the relevance of the model with the data centers. We find Cobb-Douglas model (CD) to be most suitable for optimizing the cost associated with the data centers. 3D graphs of cost data collected from various sources, have been generated for better understanding of the model. The remainder of the paper is organized as follows: Related work in the same field by other authors and Industry practices have been discussed in Section Introduction & overview. The analytical foundation of our proposed decision model has been elaborated in Section Related work. The decision model has been applied on data center's real time data set and results have been analyzed in Section Analytical foundations of the decision model. We have compared our model with other mathematical models used in data center and explored the ways to overcome the drawbacks of CD function in conclusion Section Results and discussion. Appendix contains the mathematical proofs of optimal revenue and cost, computation of elasticities and relevant Matlab codes.

Related work

Industry practices

Enterprises make investments in millions of dollars on setting up data centers. The elite list includes many Fortune 500 companies Google, Apple, Facebook Inc., Amazon.com, and Microsoft, to name a few. Each of the four new Google data center projects unveiled in 2007, cost an estimated $600 million, which includes capital investment for construction, infrastructure, and servers for two data center buildings. In its earnings reports, Google reported $1.9 billion spent on data centers in 2006 and $2.4 billion in 2007 [11]. Apple operates a data center in Newark, California, which it acquired in 2006 for approximately $45 million at a significant discount to its construction cost. The data center occupies 108,000 square feet of total space. Apple intends to invest more than $1 billion over the next 10 years on its 183-acre data center campus in Maiden, North Carolina [12]. Facebook has invested more than $1 billion in the infrastructure that powers its social network, which now serves more than 845 million users a month around the globe. The company spent $606 million on servers, storage, network gear and data centers in 2011, and spent another $500 million in 2012 [13]. In his research at Microsoft and Amazon Web Services, Hamilton has focused on cost models for operating hyper-scale data centers. His presentation at the Amazon open house reviewed cost assumptions for an 8 megawatt data center, which could include 46,000 servers. The cost was estimated at $88 million (about $11 million per megawatt), inclusive of monthly operating costs for a facility, which is dominated by the cost for servers (57 %), followed by power and cooling (18 %) and electric power (13 %) [14].

A similarity between cloud and traditional data center is that both can be used for data storage. Cloud is an example of off-premises computing, whereas data centers are being used on premise storing system. Nowadays, data centers are effectively being utilized in cloud computing. Cloud services are now being provided through data centers, which house cloud services and cloud-related resources. Cloud service providers also own data centers, which is located in different geographical location, for provisioning of uninterrupted services in case of outage and unpredictable situations. IaaS (Infrastructure as a service), which provides facilities like virtual machines, storage and load balancing maintains a large pool of resources in data centers. Data centers, which are largely being used for cloud computing are called cloud data center. Lately the demarcation of the terms has disappeared and all are referred as data centers. Existing data centers are often restructured with modern equipment so that it can take advantage of greater performance and energy efficient facilities of cloud computing. The entire process of modernization of data center is called data center transformation [15].

Cloud services are scalable, implying it will allocate resources based on your demands. We are considering the storage usage, which may range from terabytes to petabytes. Say, for example, if one organization need store 4000 GB (4 TB) of data in Amazon S3 (Simple Storage Service) then it would cost $118.50 per month for disk space (Considering Amazon charges $0.03 per GB for the first TB and $0.0295 per GB for the next 49 TB). Apart from storage charges, cloud service providers also charge the network usage as sometimes it requires to transfer data out of storage. Amazon S3 charges $0.090 per GB to transfer data up to 10 TB whereas google charges $0.11, adding $720 for AWS storage or it may incur $880 for google storage. API requests such as get, put, delete, copy, post etc. may also incur some cost. The charges may vary from $0.005 to $0.01 per 1000 or 10,000 requests based on the cloud service providers. There are escalating demands for data center space as services such as big data are migrating to cloud. Recent lease activities by the big data center players reveal the high demand for cloud services. Rackspace has leased 58,000 square feet at Digital Reality's 69 acre data center park in Dallas. The rising demand is attracting investment from non- traditional players into the data center space. Cousins Properties has transformed its 170,000 square feet of unused space out of a total of 1 million square foot American Cancer society center in Atlanta into a data center [16].

Since, energy consumption of cloud data centers is a key concern for owners owing to rising energy costs (fuel), CO2 emissions related to this consumption have become relevant [17]. Therefore, saving money in the energy budget of a cloud data center, without sacrificing Service Level Agreements (SLA) is an excellent incentive for cloud data center owners, and would at the same time be a great success for environmental sustainability. The ICT resources, servers, storage devices and network equipments consume maximum power. Processors [18] are the main contributors to the server's power consumption whereas other components [19] like multiple level caches, RAM, I/O activities also contribute to the total power consumption of the server. The storage devices range from a single hard disk to SAN (Storage Area Network) devices, which consume a significant amount of power. The other significant contributors to power consumption are network equipments which includes routers and switch fabrics.

Academic work

James Hamilton [20] has shown that, quite significantly, power is not the largest cost, if the amortization cost of power, cooling infrastructure for 15 years and new server amortization cost over 3 years are taken into consideration. He concluded that, cooling amortization and server amortization monthly payments have been computed using 5 % per annum cost and server hardware costs are the largest. But power infrastructure cost will rise and server hardware cost may fall, resulting in the domination of power cost over all other data center expenses in not so distant future. Generally, a typical data center comprises 100 fully loaded racks with the current generation 1U servers needing $1.2 million for power and an additional $1.2 million for cooling infrastructure per annum. Moreover, $1.8 million annual cost is incurred due to maintenance, amortization of power and cooling equipment. Thus, power is the most significant cost of the data center while server hardware contributes to the biggest chunk of the total operating cost. These two cost factors are primarily considered for calculation of output elasticities using 3D graphs for three phases of returns in the enterprise lifecycle.

Returns are generally used to measure the corresponding change in output subsequent to change in physical dose of an input. Every enterprise has an initial increasing returns to scale, followed by constant returns and finally decreasing returns. The reasons for the occurrence of these phases are described later in the section titled "Analytical foundations of the decision model".

Cobb-Douglas function has been widely used in economics and various sectors. Askhan Hassani has used this production function in construction management, in construction schedule crashing and project risk analysis related to duration of construction projects [21]. Moyazzem, Ajit and Tapati have used Cobb-Douglas function to decide the most suitable functional form of production process for major manufacturing sectors of a country. They have applied Cobb-Douglas model with additive error and multiplicative error term [22]. De-Min Wu [23], have shown the exact distribution of the indirect least squares estimator of the coefficients of the Cobb-Douglas production function within the context of a stochastic production model of Marschak-Andrews type. Efstratios Rappos, Stephan and Rudlof have proposed integer programming optimization model of data center for determining the optimal allocation of data components among a network of Cloud data servers in such a way that it minimizes the total costs of additional storage, estimated data retrieval costs and network delay penalties [24]. Geo-optimization technique considers the geographical location of the servers and customers while trying to optimize cost of cloud services [25]. Combinatorial optimization method has been developed to determine the best allocation process of virtual servers to target servers or virtual resource to actual resources [26]. Efforts have been made to reduce the cost of electricity in data center under multiple electricity market environment without compromising quality of services. The model proposed in Distributed Internet Data Centers in a Multi-Electricity-Market Environment is an example of constrained mixed integer programming [27]. Budget constraints force organizations to explore strategies that yield optimal revenues.The proposed production model using Cobb-Douglas production function [28–30] is very relevant by paving an optimal way to attain the maximum revenue. In this paper, four major segments of the cost associated with data centers such as server, infrastructure, power, and network [31] are considered for optimization. Power is the fastest growing cost among all other costs. Several initiatives have been contemplated to curtail the cost associated with power. Dynamic smart cooling techniques, equipped with temperature- aware cooling algorithm has been adopted to reduce the cost. Using scale processor and system power enables data centers save energy and reduce cost [32]. Here in this paper, an attempt is made to achieve the dual goal of profit maximization and cost minimization within certain constraints. It is proved mathematically that cost minimization can be achieved at increasing return to scale, whereas profit maximization can be attained at decreasing return to scale. The Cobb-Douglas production function, which has been used rigorously in this revenue model, is endowed with a flexible functional form and less restriction over output elasticity.

Next, we define key terminologies used in the scientific investigation & deployment of our model in the revenue optimization problem.

Key terminologies & techniques
Mathematical optimization

Optimization is a technique to select the best element from a set of available alternatives in the field of mathematics, computer science, economics or management science [33]. An optimization problem can be represented in various ways.

Given: a function $f : A \rightarrow R$ from set A to the real numbers An element x_0 in A such that $f(x_0) \leq f(x)$ for all x in A (minimization). $f(x_0) \geq f(x)$ for all x in A (maximization). The optimization technique is useful for modeling many real world problems In the above formulation, the domain A is called search space of function f and elements of A are called candidate solutions or feasible solutions. The function is suitably termed as cost function, revenue function or utility function based on the area of interest. A feasible solution that minimizes (or maximizes, if that is the goal) the objective function is called an optimal solution.

Computational optimization techniques
To solve a particular problem, researchers may use an algorithm that will find the solution in finite steps. Iterative method will converge to the optimal solution and heuristic will give an approximate solution to a problem used in optimization scenario, is combinatorial algorithms.

Iterative methods:
Iterative methods are usually applied to solve problems of non-linear programming. The iterative methods differ according to the use of Hessians, Gradients or function values. Evaluating Hessians and Gradient helps improve the rate of convergence of the functions, but such methods introduce computation complexity to each iteration. In some cases, the computational complexity may be very high. One major criterion for optimizers is the number function evaluations required as this often may require large computational efforts. The derivatives some time give detailed information for such optimizers but are hard to calculate.

Methods that evaluate Hessian:

- Newton's method.

Methods that evaluate Gradient:

- Quasi-Newton method.
- Conjugate gradient method.
- Interior point method.
- Gradient descent.
- Sub gradient method.
- Ellipsoid method.
- Reduced gradient method.

Gradient descent method has been used in the proposed work to compute optimal costs.

- *Increasing returns to scale*: In the initial phase, the output may increase in a higher proportion [34]. This phase is called the phase of increasing returns. This change occurs as:
  1. Greater application of the variable factor ensures better utilization of the fixed factor. Actually, this enables the utilization of idle capacity (potential) of the fixed factor.
  2. It facilitates better division of the variable factor.
  3. It improves co-ordination between the factors. This paper establishes that cost minimization of an enterprise that invests on servers, infrastructure, network, power, etc. is achieved at this phase. The 3D plots obtained are neither concave nor convex.
  4. *Constant returns to scale*: An increase in one input may yield an increase in corresponding output in the same proportion. However, this phase rarely happens and even if it occurs, it would be for a very negligible period. Actually, it is only a passing phase between increasing and diminishing returns.
- *Decreasing returns to scale*: Ultimately, the phase of decreasing or diminishing returns will set in, whereby the deployment of an additional input will result into increase in output but at a diminishing rate or lower ratio [34].

This happens because:

1 As more and more units of a variable factor are combined with the fixed factor, the latter gets over-utilized. Hence, the rate of corresponding growth of output goes on diminishing.

2 Factors of production are imperfect substitutes of each other. The divisibility of their units is not comparable.

3 The coordination between factors get distorted so that marginal product of the variable factor declines.

Our work proves that profit maximization of an enterprise is achieved in this phase [Fig. 7]

The marginal product is the change in total output owing to a unit change in the input of a variable factor. It is also shown that marginal product increases for the initial phase, i.e. increasing returns to scale, subsequently stabilizes for constant returns and finally decreases for last phase, i.e., decreasing returns to scale.

Analytical foundations of the decision model

We propose necessary results that will be used to model production, cost and profit of the cloud data center.

Theorem 1: production maximization
Consider an enterprise that has to choose its consumption bundle (S, I, P, N) where S, I, P and N are number of servers, investment in infrastructure, cost of power and networking cost respectively of a cloud data center. The enterprise wants to maximize its production, subjected to the constraint that the total cost of the bundle does not exceed a particular amount. The company has to keep the budget constraint in mind and keep total spending within this amount.

The production maximization is done using Lagrangian Multiplier. The Cobb-Douglas function is:

$$f(S, I, N, P) = kS^{\alpha} I^{\beta} P^{\gamma} N^{\delta} \tag{1}$$

Let m be the cost of the inputs that should not be exceeded.

$$w_1 S + w_2 I + w_3 P + w_4 N = m$$

w_1: Unit cost of servers
w_2: Unit cost of infrastructure
w_3: Unit cost of power
w_4: Unit cost of network

Optimization problem for production maximization is:

$$max\ f(S, I, P, N)\ subject\ to\ m$$

The following values of S, I, P and N thus obtained are the values for which the data center has maximum

production of satisfying the given constraints on the total investment.

$$S = \frac{m\alpha}{w_1}(1+\beta+\gamma+\delta) \tag{2}$$

$$I = \frac{m\beta}{w_2}(1+\alpha+\gamma+\delta) \tag{3}$$

$$P = \frac{m\gamma}{w_3}(1+\alpha+\beta+\delta) \tag{4}$$

$$N = \frac{m\delta}{w_4}(1+\alpha+\beta+\gamma) \tag{5}$$

The above results are proved in Appendix 1.

A quick heuristic for CRS: Revenue optimization

If we consider again the case of constant return to scale, where all the elasticities of different cost components are equal. $y = \prod_{i=1}^{n} x_i^{\alpha_i}$, where all α_i are equal and $\sum \alpha_i = 1$. In such scenario, the response variable or output turns out to be the geometric mean of all inputs.

Theorem 2: cost minimization

Consider an enterprise that has a target level of output to achieve by investing a minimum amount. The Cobb-Douglas function is of the form:

$$y_{tar} = f(S,I,N,P) = kS^{\alpha}I^{\beta}P^{\gamma}N^{\delta} \tag{6}$$

y_{tar} is the target output of the firm that needs to be achieved and w_1, w_2, w_3 and w_4 are unit prices of servers, infrastructure, power and network respectively. Cost minimization problem is formulated as follows:

$$\min_{S,\,I\,,\,P,\,N} w_1S + w_2I + w_3P + w_4N \textit{ subject to } y_{tar} \tag{7}$$

The cost for producing y_{tar} units in cheapest way is c, where

$$c = w_1S + w_2I + w_3P + w_4N \tag{8}$$

c can be written as;

$$c = Q\left[w_1^{\alpha}w_2^{\beta}w_3^{\gamma}w_4^{\delta}\right]^{\frac{1}{\alpha+\beta+\gamma+\delta}} y_{tar}^{\frac{1}{\alpha+\beta+\gamma+\delta}} \tag{9}$$

where,

$$Q = k^{\frac{-1}{\alpha+\beta+\gamma+\delta}}\left[\frac{\alpha^{\beta+\gamma+\delta}}{\beta^{\beta}+\gamma^{\gamma}+\delta^{\delta}} + \frac{\beta^{\alpha+\gamma+\delta}}{\alpha^{\alpha}+\gamma^{\gamma}+\delta^{\delta}} + \frac{\gamma^{\alpha+\beta+\delta}}{\alpha^{\alpha}+\beta^{\beta}+\delta^{\delta}} + \frac{\delta^{\alpha+\beta+\gamma}}{\alpha^{\alpha}+\beta^{\beta}+\gamma^{\gamma}}\right]^{\frac{1}{\alpha+\beta+\gamma+\delta}}$$

The above results are proved in Appendix 2.

$$C_{avg} = \frac{C}{y_{tar}} = Q\left[w_1^{\alpha}w_2^{\beta}w_3^{\gamma}w_4^{\delta}\right]^{\frac{1}{\alpha+\beta+\gamma+\delta}} y_{tar}^{\frac{1}{\alpha+\beta+\gamma+\delta}-1}$$

Average cost $\left(C_{avg}\right)$ should decrease in order to achieve minimum cost with output y_{tar},

$$\frac{1-\alpha-\beta-\gamma-\delta}{\alpha+\beta+\gamma+\delta} < 0$$

$$\alpha+\beta+\gamma+\delta > 1$$

Therefore, the enterprise will have cost minimization at the phase of increasing returns to scale.

Global Minima for cost minimization: a heuristic approach

Apart from the above calculation, Gradient Descent method has been used to retrieve the values of elasticities where cost minimization is ensured. For simplification of equations, let us consider two cost segments X and Y. w_1 and w_2 are unit prices of X and Y. Rewriting the cost function using the newly elected variables, we obtain

$$c = w_1X + w_2Y \tag{10}$$

The newly formed CD function is,

$$y_{tar} = X^{\alpha}Y^{\beta}$$

$$X^{\alpha} = \frac{y_{tar}}{Y^{\beta}}$$

$$X = \left(\frac{y_{tar}}{Y^{\beta}}\right)^{\frac{1}{\alpha}}$$

Putting the value of X in cost function 10, we obtain

$$c = W_1\left(\frac{y_{tar}}{Y^{\beta}}\right)^{\frac{1}{\alpha}} + W_2Y$$

$$\frac{\partial c}{\partial \alpha} = -\left(w_1 y_{tar}^{\frac{1}{\alpha}} Y^{-\frac{\beta}{\alpha}} ln\left(\frac{y_{tar}}{Y^{\beta}}\right)\right)\frac{1}{\alpha^2}$$

$$\frac{\partial c}{\partial \beta} = -\left(w_1 y_{tar}^{\frac{1}{\alpha}} Y^{-\frac{\beta}{\alpha}} ln y_{tar} ln\left(\frac{y_{tar}}{Y^{\beta}}\right)\right)\frac{1}{\alpha^3}$$

The above partial derivatives are used in gradient descent method for cost minimization.

Gradient descent: Algorithm

1. **procedure** GRADIENTDESCENT()
2. $\frac{\partial c}{\partial \alpha} \leftarrow -\left(w_1 y_{tar}^{\frac{1}{\alpha}} Y^{-\frac{\beta}{\alpha}} ln\left(\frac{y_{tar}}{Y^{\beta}}\right)\right)\frac{1}{\alpha^2}$
3. $\frac{\partial c}{\partial \beta} \leftarrow -\left(w_1 y_{tar}^{\frac{1}{\alpha}} Y^{-\frac{\beta}{\alpha}} ln y_{tar} ln\left(\frac{y_{tar}}{Y^{\beta}}\right)\right)\frac{1}{\alpha^3}$
4. **repeat**
5. $\alpha_{n+1} \leftarrow \alpha_n - \delta\frac{\partial c}{\partial \alpha}$
6. $\beta_{n+1} \leftarrow \beta_n - \delta\frac{\partial c}{\partial \beta}$
7. $\alpha_n \leftarrow \alpha_{n+1}$
8. $\beta_n \leftarrow \beta_{n+1}$
9. **until** $(\alpha_{n+1} > 0)\,||\,(\beta_{n+1} > 0)\,||\,(\alpha_{n+1}+\beta_{n+1} > 1)$
10. **end procedure**

Using the above algorithm, the optimal values of α, β and cost have been computed (cf. Results and discussion).

Theorem 3: profit maximization

Consider an enterprise that needs to maximize its profit. The Profit function is:

$$\pi = pf(S,I,N,P) - w_1S - w_2I - w_3P - w_4N$$

Profit maximization is achieved when:

$$(1)\ p\frac{\partial f}{\partial S} = w_1\ (2)\ p\frac{\partial f}{\partial I} = w_2\ (3)\ p\frac{\partial f}{\partial P} = w_3\ (4)\ p\frac{\partial f}{\partial N} = w_4$$

The calculations yield the following values of S, I, P and N as obtained:

$$S = \left(pk\alpha^{1-(\beta+\gamma+\delta)}\beta^{\beta}\gamma^{\gamma}\delta^{\delta}w_1^{\beta+\gamma+\delta-1}w_2^{-\beta}w_3^{-\gamma}w_4^{-\delta}\right)^{\frac{1}{1-(\alpha+\beta+\gamma+\delta)}} \tag{11}$$

$$I = \left(pk\alpha^{\alpha}\beta^{1-(\alpha+\gamma+\delta)}\gamma^{\gamma}\delta^{\delta}w_1^{-\alpha}w_2^{\alpha+\gamma+\delta-1}w_3^{-\gamma}w_4^{-\delta}\right)^{\frac{1}{1-(\alpha+\beta+\gamma+\delta)}} \tag{12}$$

$$P = \left(pk\alpha^{\alpha}\beta^{\beta}\gamma^{1-(\alpha+\beta+\delta)}\delta^{\delta}w_1^{-\alpha}w_2^{-\beta}w_3^{\alpha+\beta+\delta-1}w_4^{-\delta}\right)^{\frac{1}{1-(\alpha+\beta+\gamma+\delta)}} \tag{13}$$

$$N = \left(pk\alpha^{\alpha}\beta^{\beta}\gamma^{\gamma}\delta^{1-(\alpha+\beta+\gamma)}w_1^{-\alpha}w_2^{-\beta}w_3^{-\gamma}w_4^{\alpha+\beta+\gamma-1}\right)^{\frac{1}{1-(\alpha+\beta+\gamma+\delta)}} \tag{14}$$

The above results are proved in Appendix 3. These values of S, I, P and N are the "profit maximizing data center's" demand for inputs, as a function of the prices of all the inputs, and of the price of output. Substituting values of S, I, P and N into Eq. (1); we get

$$y = \left(kp^{\alpha+\beta+\gamma+\delta}\alpha^{\alpha}\beta^{\beta}\gamma^{\gamma}\delta^{\delta}w_1^{-\alpha}w_2^{-\beta}w_3^{-\gamma}w_4^{-\delta}\right)^{\frac{1}{1-(\alpha+\beta+\gamma+\delta)}} \tag{15}$$

y increases in price of its output and decreases in price of its inputs iff:

$$1-(\alpha+\beta+\gamma+\delta) > 0 \qquad \alpha+\beta+\gamma+\delta < 1$$

Therefore, the enterprise will have profit maximization at the phase of decreasing returns to scale. It is later shown in Appendix 4, that profit maximization is scalable provided for an arbitrary, n, number of input variables (constant), the result stands as long as $\sum_{i=1}^{n}\alpha_i < 1$; where α_i is the i^{th} elasticity of the input variable x_i.

Consider again the CD function for maximization.

$$y = A^{\alpha}B^{\beta} \tag{16}$$

where A and B are constants. Let $[\alpha_{min}, \alpha_{max}]$ be the range of permissible values for α, similarly $[\beta_{min}, \beta_{max}]$ be the range for β, where $\alpha_{min}, \alpha_{max}, \beta_{min}, \beta_{max} > 0$. To maximize y, if $A > 1$ then $\alpha = \alpha_{max}$ (α should be as large as possible and α_{max} is the largest permitted value). Similarly, if $A < 1$, then $\alpha = \alpha_{min}$. Since the terms involving α are independent of those involving β, the same logic can be applied independently to the term B^{β}. An easy way to see the above is by taking log of both sides of (15), we get

$$log(y) = \alpha log(A) + \beta log(B) \tag{17}$$

To maximize $log(y)$, if $log(A)$ is negative, α needs to be as small as possible (since $\alpha > 0$) else α must be as large as possible. Same applied to β (Table 1).

Consider the case where we have a set of data points i.e. instead of constants A and B we have,

$$y_i = u_i^{\alpha}v_i^{\beta} \tag{18}$$

where i = 1 to N Our criterion is to choose α and βso as to maximize $y = \prod_{i=1}^{n} y_i$ i.e. maximize

$$\prod_{i=1}^{N} y_i = \left(\prod_{i=1}^{N} u_i\right)^{\alpha}\left(\prod_{i=1}^{N} v_i\right)^{\beta} \tag{19}$$

The RHS of (18) is similar in form to (16) and hence same rule can be applied i.e. If $\prod_{i=1}^{N} u_i < 1$ then $\alpha = \alpha_{min}$ else $\alpha = \alpha_{max}$. The term involving β can be minimized similarly and independently. The only remaining step is to determine the permissible ranges. Let ϵ be the smallest value that α and β can take. Suppose in the above example, $\prod_{i=1}^{N} u_i < 1$ and $\prod_{i=1}^{N} v_i > 1$. We know that α should be minimized and β should be maximized. Since $\alpha + \beta < 1$, let $\alpha + \beta = 1 - \delta$, where δ is a small non-negative number. We then have $\alpha_{min} = \epsilon$ and $\beta_{max} = 1 - \delta - \epsilon$.

Results and discussion

As mentioned earlier, server and power/cooling costs form the biggest chunk of the total cost. These two inputs are considered for computing the values of the elasticities using 3D plots. However, the results obtained hold good for any number of inputs. It is also possible to aggregate the inputs into two broad categories-operational expenditure and capital expenditure and use these as the two inputs in the proposed cost model. Operational expenditures include the recurring costs like power/cooling, server management etc; whereas capital expenditure includes initial investment such as new server cost, infrastructure costs etc.

Table 1 Maximization of CD function for fixed A and B

A, B	α, β
$A < 1, B < 1$	$[\alpha_{min}, \beta_{min}]$
$A < 1, B > 1$	$[\alpha_{min}, \beta_{max}]$
$A > 1, B < 1$	$[\alpha_{max}, \beta_{min}]$
$A > 1, B > 1$	$[\alpha_{max}, \beta_{max}]$

The data associated with data center costs from various sources have been accumulated and Cobb-Douglas function is applied on varying elasticities to find the optimal solution for revenue of data center, and finally revenue maximization is demonstrated graphically. All simulation results have been generated by a computer system using Matlab.

The approximate data from the Fig. 2 for two types of costs, namely server management/administrative cost and power/cooling cost are captured. The optimal elasticity of each input and maximum revenue for each year using Matlab code [Appendix 5] are obtained. The experiment has been conducted for the following three cases:

1) Increasing Returns to Scale
2) Constant Returns to Scale
3) Decreasing Returns to Scale

Case 1: increasing returns to scale

Applying the constraints:
$\alpha + \beta > 1$
$\alpha > 0$
$\beta > 0$
to the function:
$f = kx^{\alpha} y^{\beta}$

Using fmincon function of matlab [Appendix 5], the values of elasticities for which revenue is maximized for each year are obtained.

In Table 2, all units are in \$B. The optimal revenue for all the years is obtained at $\alpha = 1.8$ and $\beta = 0.1$ Using these results, 3D-simulations are created and the corresponding graphs are obtained.

In Figs. 3 and 4, X axis represents output elasticity α of new server expenditure, Y axis represents output elasticity β of power/cooling cost and Z axis represents revenue. The graphs obtained depict the effects of Cobb-Douglas production function over worldwide IT spending in data center.It is observed that the graphs obtained are not concave graphs, which is shown mathematically as well [Appendix 6]. It is prominent from the graphs that revenues in the range of α around 1.8 and β around 0.1 give the optimal revenue for each year.

It is seen from the data set that if new server cost is increased by 1 unit from year 2007(\$56B) to 2008(\$57B), the revenue changes by 63 units and when it is increased by 1 unit from year 2008(\$57B) to 2009(\$58B), the revenue changes by 64.68 units. This proves that marginal product of input (new server cost) increases in increasing returns to scale.

Case 2: constant returns to scale

Applying the constraints:
$\alpha + \beta = 1$
$\alpha > 0$
$\beta > 0$
to the function: $f = kX^{\alpha} Y^{\beta}$
and using fmincon function of matlab [Appendix 5], the values of elacticities are obtained which maximize the revenue for each year, as shown in Table 3.

In Table 3, all units are in \$B. The optimal revenue for all years are obtained at $\alpha = 0.9$ and $\beta = 0.1$ Using these results, 3D-simulations are created and the graphs of Constant Return to scale are obtained.

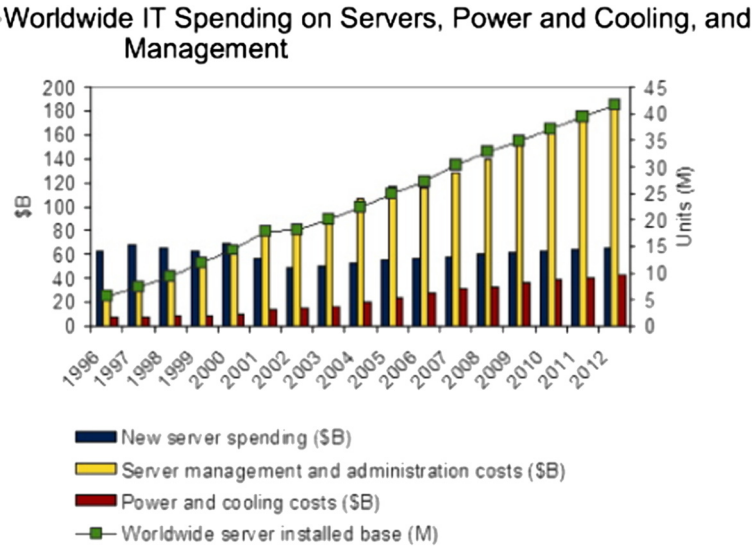

Fig. 2 World wide IT spending on servers, power and cooling, and management

Table 2 Simulation output for IRS

Year	New server	Power and cooling	Elasticity (α)	Elasticity (β)	Max. revenue
1996	62	5	1.8000	0.1000	1977.88
1997	65	5	1.8000	0.1000	2153.48
1998	62	10	1.8000	0.1000	2119.84
1999	60	10	1.8000	0.1000	1998.35
2000	65	10	1.8000	0.1000	2308.04
2001	55	15	1.8000	0.1000	1779.35
2002	45	15	1.8000	0.1000	1239.91
2003	47	15	1.8000	0.1000	1340.86
2004	50	20	1.8000	0.1000	1542.58
2005	52	20	1.8000	0.1000	1655.42
2006	55	20	1.8000	0.1000	1831.28
2007	56	30	1.8000	0.1000	1969.92
2008	57	30	1.8000	0.1000	2033.69
2009	58	30	1.8000	0.1000	2098.37
2010	59	40	1.8000	0.1000	2227.09
2011	60	40	1.8000	0.1000	2295.50
2012	60	40	1.8000	0.1000	2295.50

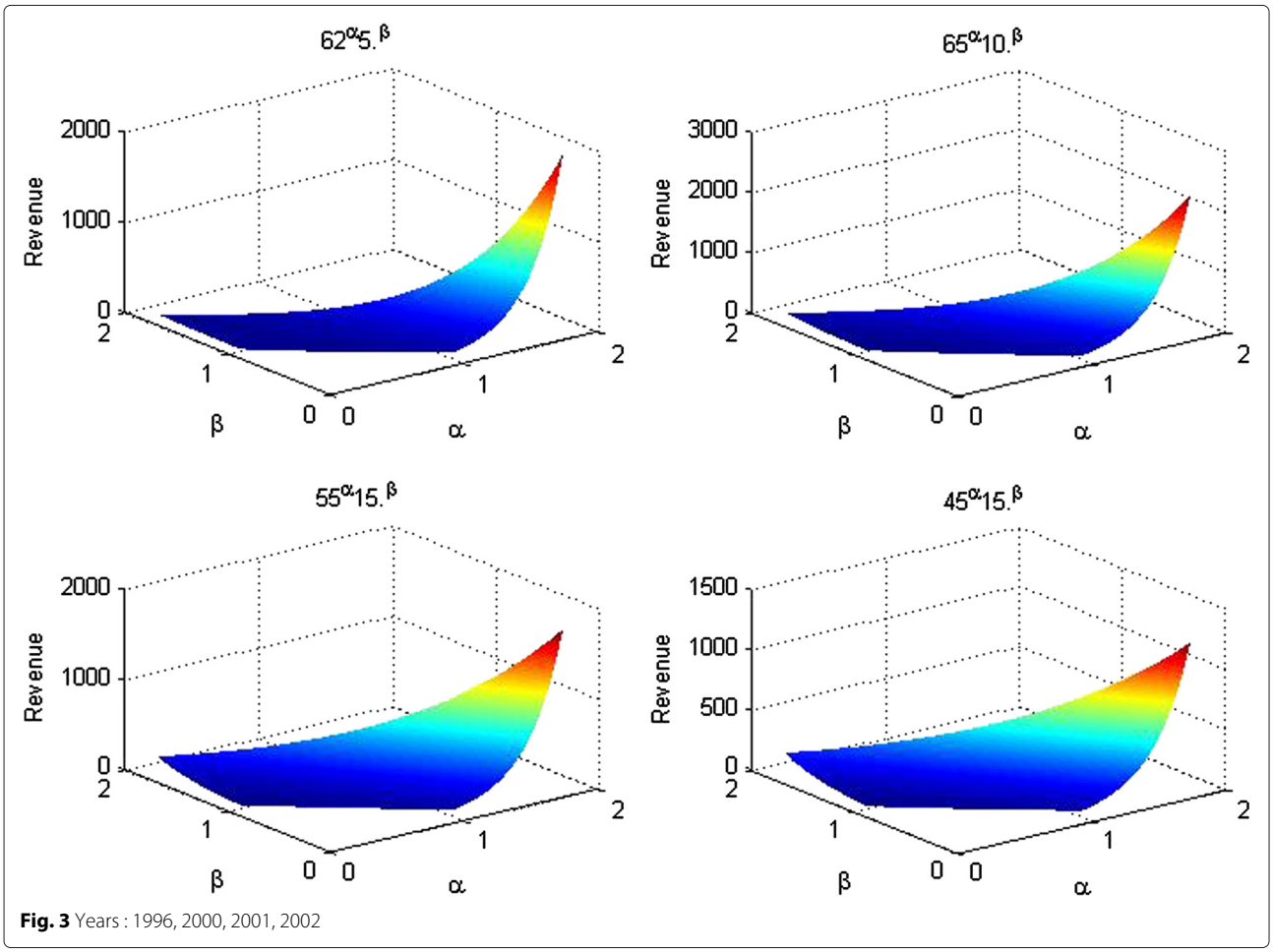

Fig. 3 Years : 1996, 2000, 2001, 2002

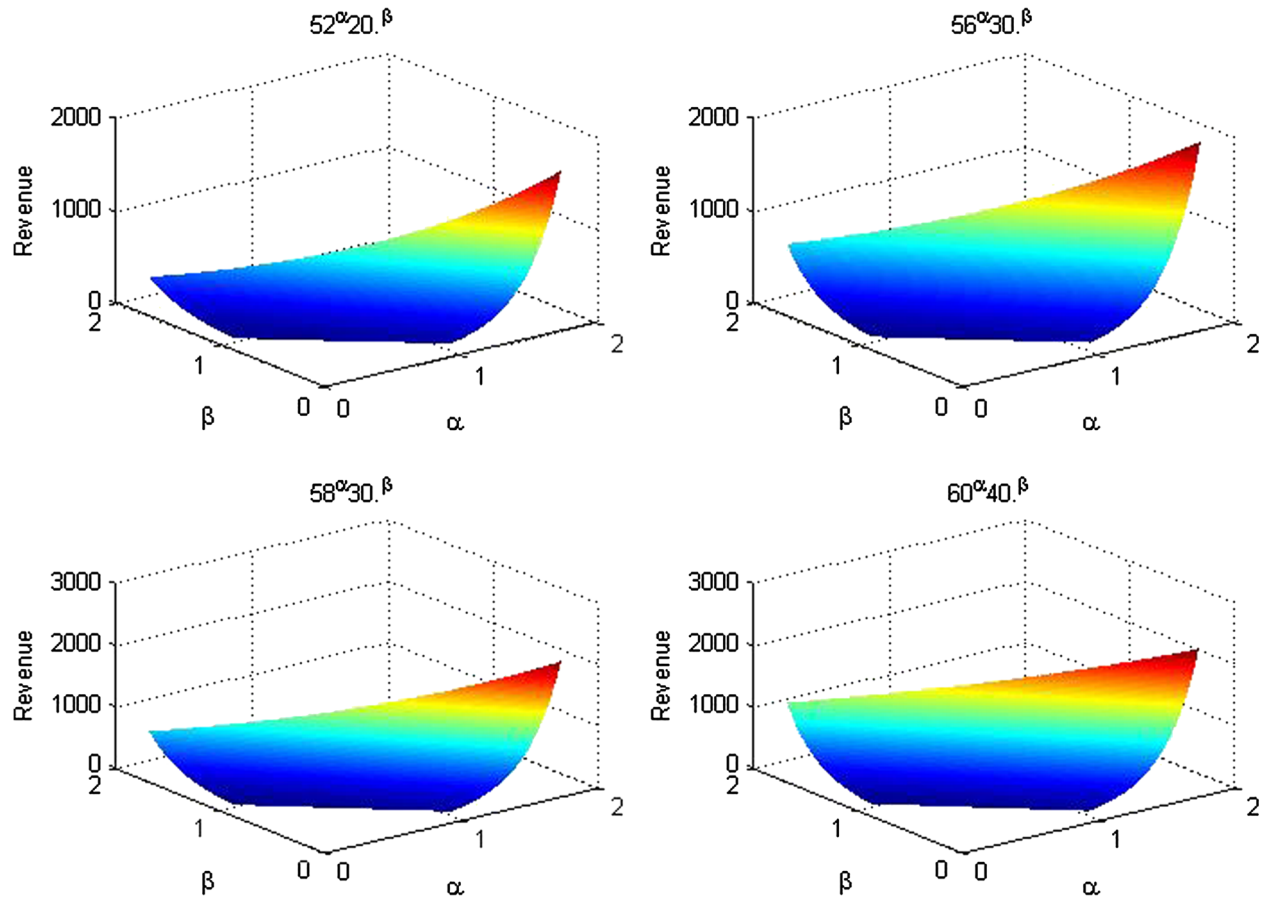

Fig. 4 Years : 2005, 2007, 2009, 2011

Table 3 Simulation output for CRS

Year	New server	Power and cooling	Elasticity (α)	Elasticity (β)	Max. revenue
1996	62	5	0.9000	0.1000	48.2003
1997	65	5	0.9000	0.1000	50.2943
1998	62	10	0.9000	0.1000	51.6598
1999	60	10	0.9000	0.1000	50.1575
2000	65	10	0.9000	0.1000	53.9041
2001	55	15	0.9000	0.1000	48.2987
2002	45	15	0.9000	0.1000	40.3181
2003	47	15	0.9000	0.1000	41.9273
2004	50	20	0.9000	0.1000	45.6222
2005	52	20	0.9000	0.1000	47.2613
2006	55	20	0.9000	0.1000	49.7084
2007	56	30	0.9000	0.1000	52.6116
2008	57	30	0.9000	0.1000	53.4564
2009	58	30	0.9000	0.1000	54.2997
2010	59	40	0.9000	0.1000	56.7509
2011	60	40	0.9000	0.1000	57.6159
2012	60	40	0.9000	0.1000	57.6159

In Figs. 5 and 6, X axis represents output elasticity α of new server expenditure, Y axis represents output elasticity β of power/cooling costs and Z axis represents revenue. The graphs obtained demonstrate the effects of Cobb-Douglas production function over worldwide IT spending in data center. It is observed that the graphs obtained are concave graphs, otherwise proved mathematically [Appendix 6]. The graphs reveal that revenues in the range of α close to 0.9 and β close to 0.1 are optimal.

It is evident from the data set that if new server cost is increased by 1 unit from year 2007($56B) to 2008($57B), the revenue changes by 0.84 units. And when it is increased by 1 unit from year 2008($57B) to 2009($58B), the revenue changes by 0.84 units. This proves that marginal product of input (new server cost) is constant in constant returns to scale.

Case 3: decreasing returns to scale

Applying the constraints:

$\alpha + \beta < 1$

$\alpha > 0$

$\beta > 0$

to the function:

$f = kX^{\alpha}Y^{\beta}$

and using fmincon function of matlab [Appendix 5], the values of elacticities are obtained for which revenue is maximized for each year.

In Table 4, all units are in $B. The optimal revenue for all years are obtained at $\alpha = 0.8$ and $\beta = 0.1$.

In Figs. 7 and 8, X axis represents output elasticity α of new server spending, Y axis represents output elasticity β of power/cooling costs and Z axis represents revenue. The graphs obtained reflect the effects of Cobb-Douglas production function over worldwide IT spending in data center. We observe that the graphs obtained are concave graphs which has been proved mathematically also [Appendix 6]. It is prominent from the graphs that revenues in the range α near 0.8 and β near 0.1 give the optimal revenue for each year.

It can be seen from the data set that if new server cost is increased by 1 unit from year 2007($56B) to 2008($57B), the revenue changes by 0.50 units. And when it is increased by 1 unit from year 2008($57B) to 2009($58B), the revenue changes by 0.49 units. This

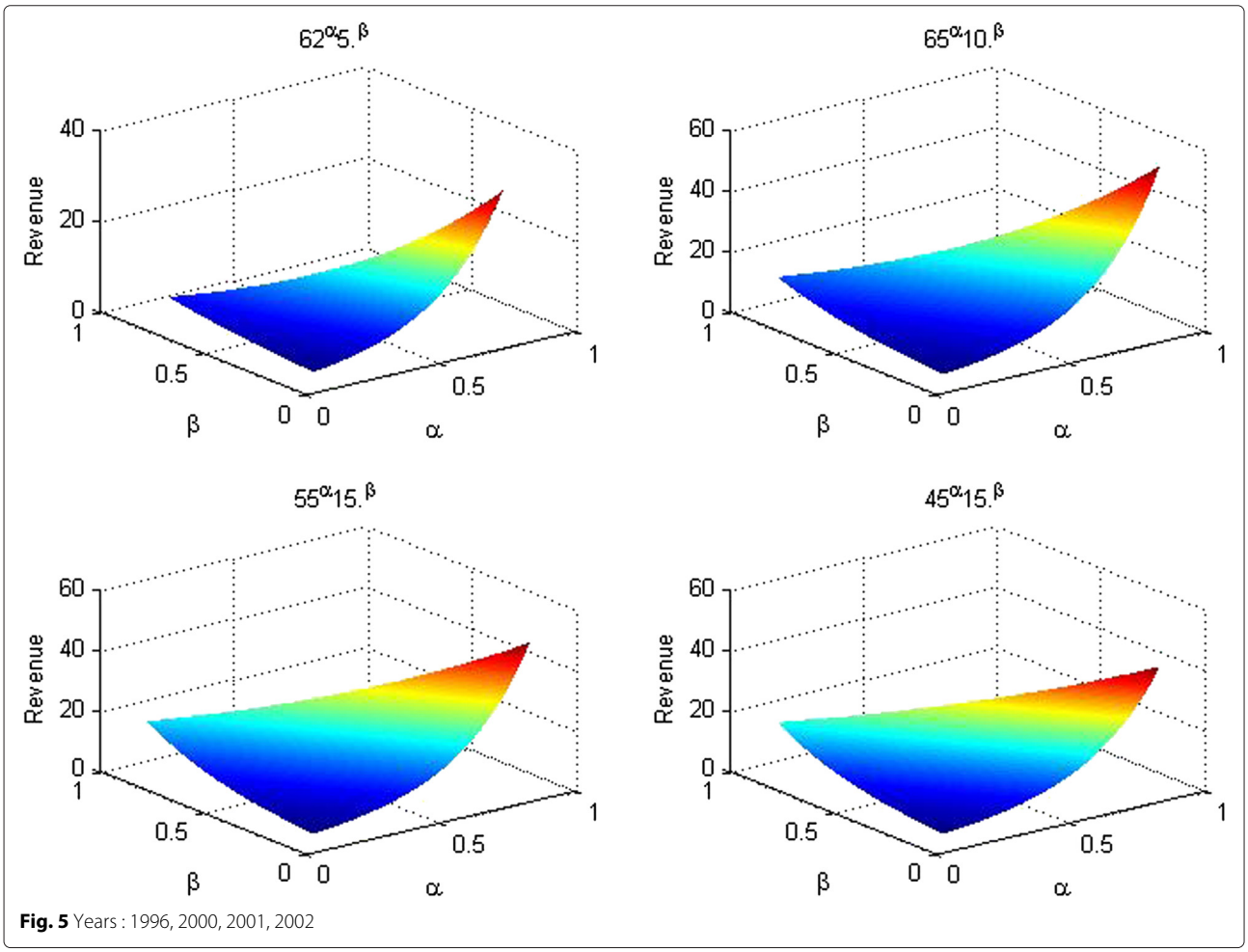

Fig. 5 Years : 1996, 2000, 2001, 2002

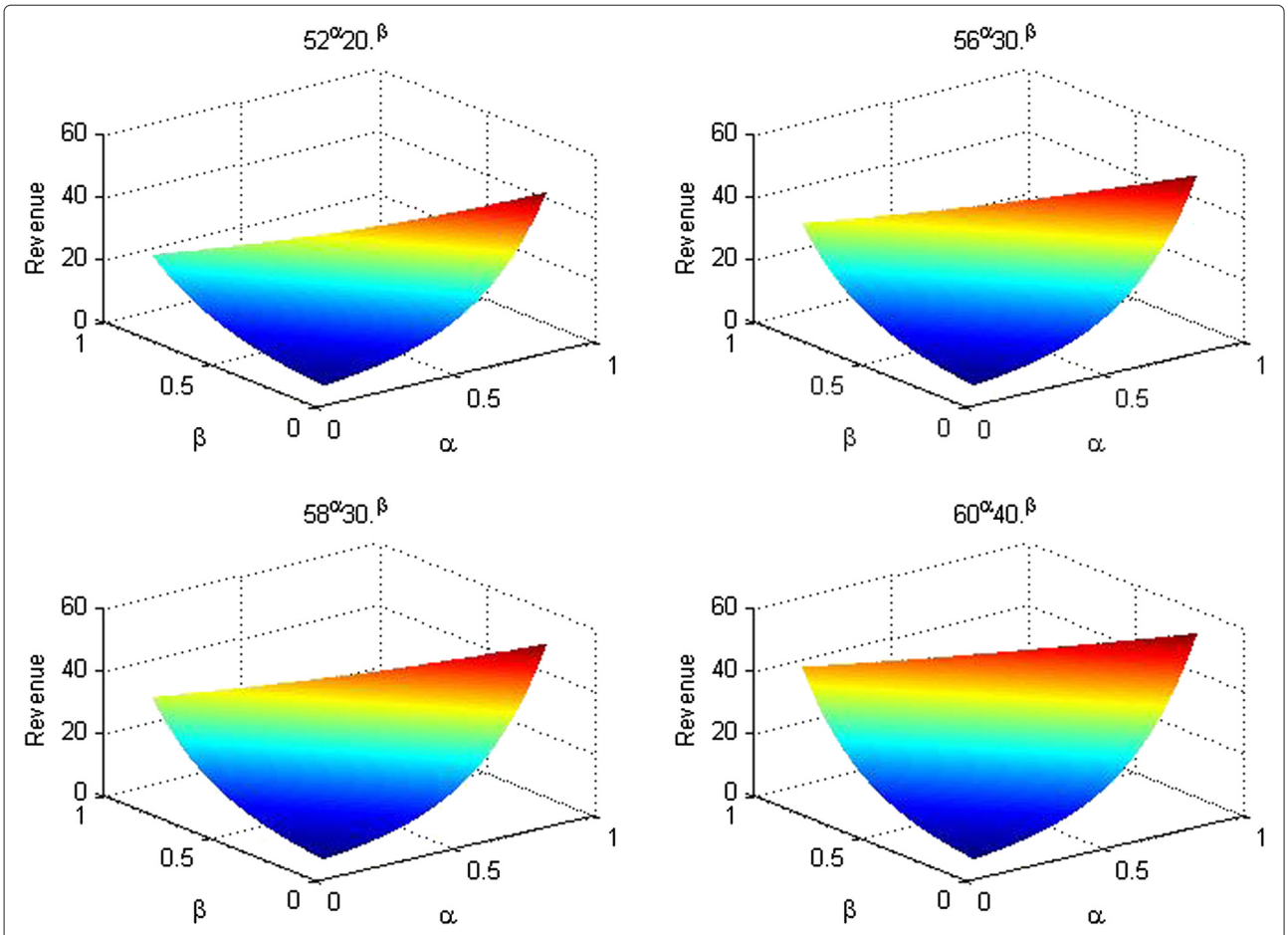

Fig. 6 Years : 2005, 2007, 2009, 2011

Table 4 Simulation output for DRS

Year	New server	Power and cooling	Elasticity (α)	Elasticity (β)	Max. revenue
1996	62	5	0.8000	0.1000	31.9014
1997	65	5	0.8000	0.1000	33.1305
1998	62	10	0.8000	0.1000	34.1911
1999	60	10	0.8000	0.1000	33.3059
2000	65	10	0.8000	0.1000	35.5084
2001	55	15	0.8000	0.1000	32.3519
2002	45	15	0.8000	0.1000	27.5536
2003	47	15	0.8000	0.1000	28.5291
2004	50	20	0.8000	0.1000	30.8517
2005	52	20	0.8000	0.1000	31.8351
2006	55	20	0.8000	0.1000	33.2961
2007	56	30	0.8000	0.1000	35.1773
2008	57	30	0.8000	0.1000	35.6789
2009	58	30	0.8000	0.1000	36.1788
2010	59	40	0.8000	0.1000	37.7474
2011	60	40	0.8000	0.1000	38.2584
2012	60	40	0.8000	0.1000	38.2584

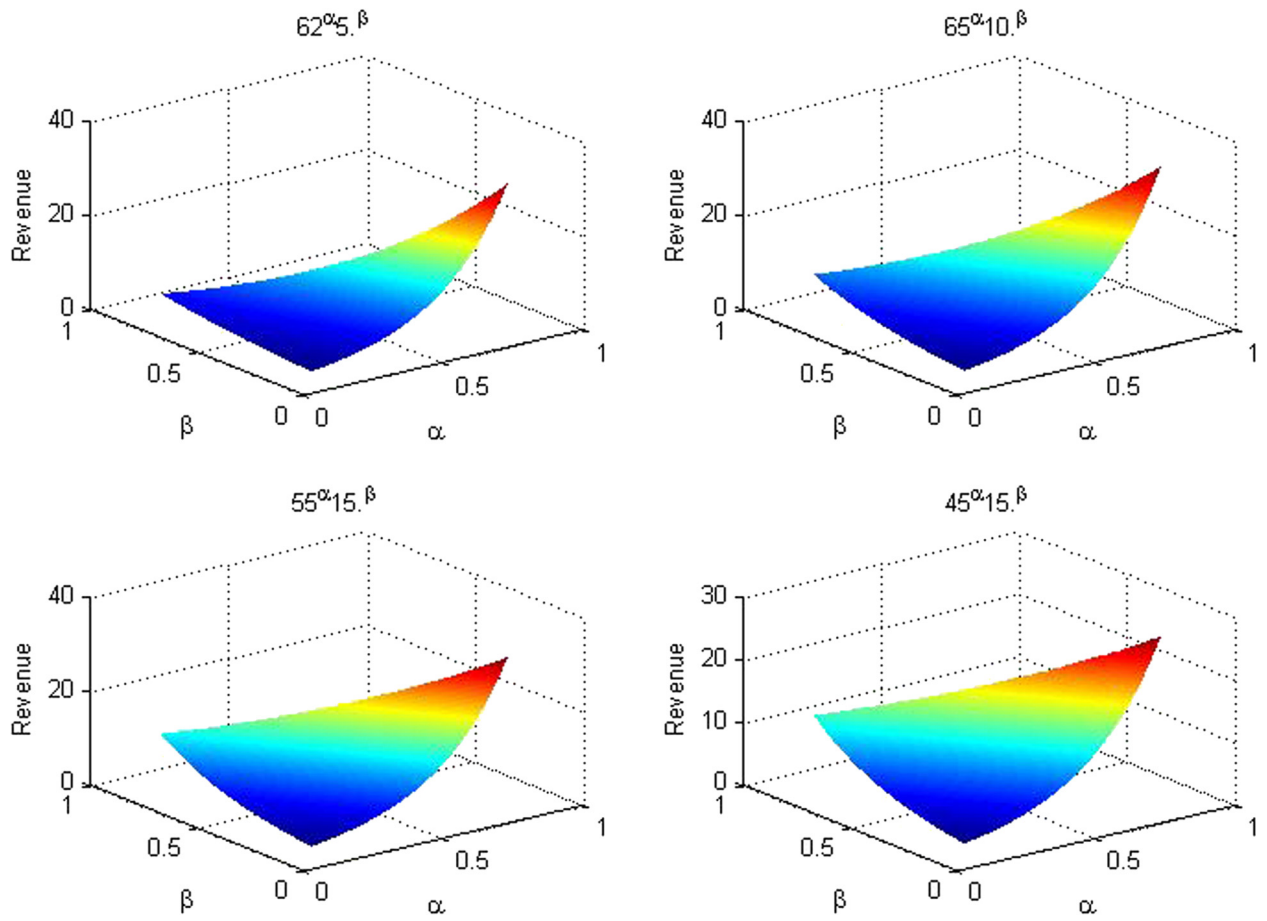

Fig. 7 Years : 1996, 2000, 2001, 2002

proves that marginal product of input(new server cost) decreases in decreasing returns to scale.

Gradient descent method has been applied on the same world wide IT spending dataset to find out optimal elasticities for cost minimization. As we have already proved that cost minimization can be achieved in increasing return to scale, the sum of the elasticities should be grater than 1. The initial values of the elasticities have been assumed as 1.2 and 0.7 whereas step size for each iteration has been considered as 0.001 (Table 5).

In the gradient descent calculation, we assumed the target revenue as \$120B and unit cost of new server installation as 0.6.

Conclusion

In this paper, the proposed production model using Cobb-Douglas production function is used to quantify boundaries of the inputs (cost of servers, networking, infrastructure and power), for which the maximum revenue subject to the constraint that the total cost does not exceed a particular amount, is attained. These values for the inputs are mentioned in Eqs. 2 through 5, and can be used to obtain the maximum revenue by substituting in the production model. Similarly, the production model is also used to obtain the minimum total cost subject to a certain amount of production (output) that has to be achieved. This value of total minimum cost can be computed using Eq. 9. Further, the inputs contributing to the maximum profit are deduced, as evident from Eqs. 11 through 14. Computation of revenue function using the production model, a subsequent operation, becomes straight forward enough.

Hence, the total cost required to achieve the maximum revenue, the minimum cost required to achieve a pre-defined revenue as well as the total cost required to achieve the maximum profit are successfully calculated. These strategies can be used for deploying existing resources optimally so that a responsible and beneficial balance can be achieved over the longer term. Economic sustainability implies utilizing the assorted assets of the company efficiently to deliver functioning profitability over a sustained period.

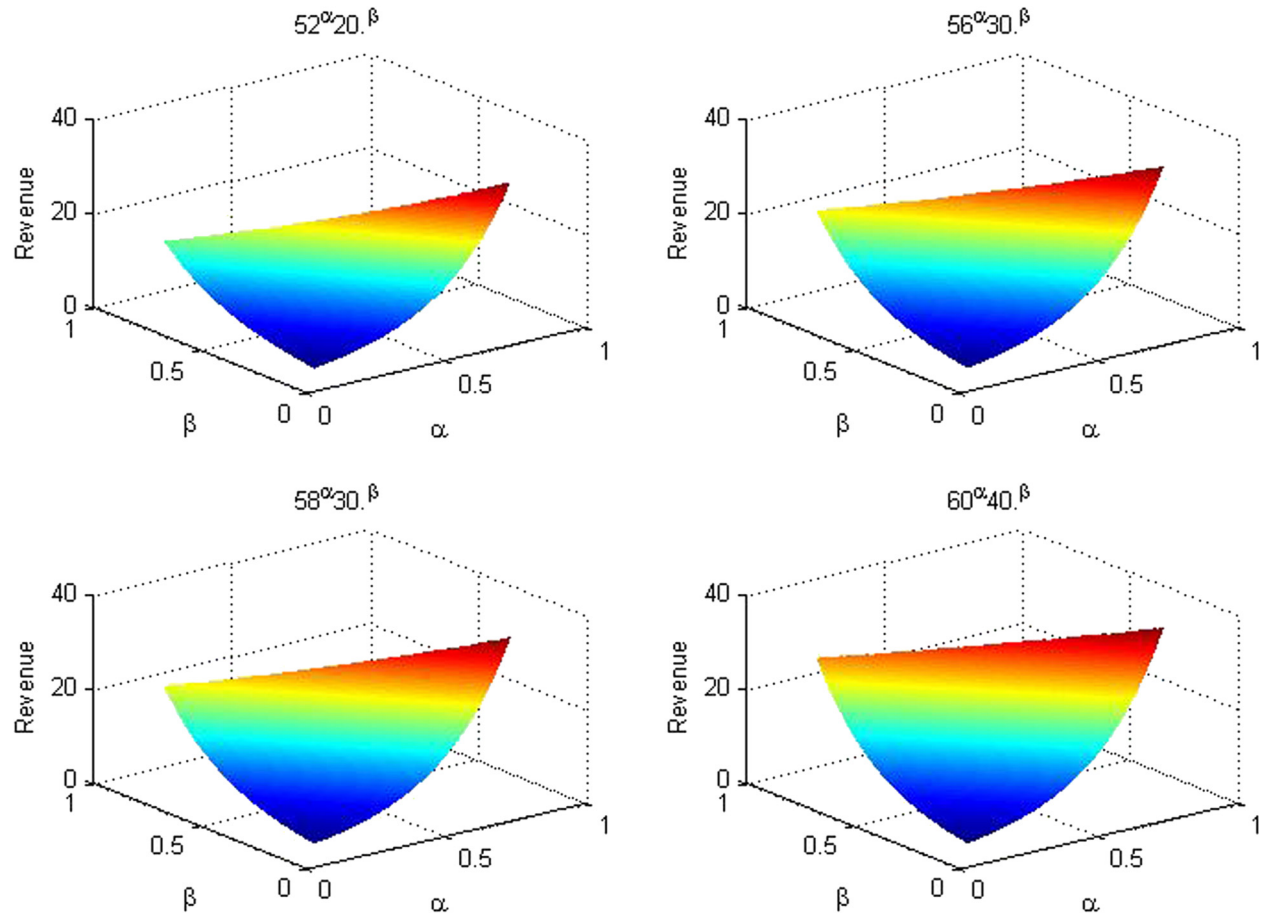

Fig. 8 Years : 2005, 2007, 2009, 2011

Table 5 Gradient descent output for costminimization

Year	New server	Power and cooling	Elasticity (α)	Elasticity (β)	Min. Cost
1996	62	5	1.3732	0.0091	56.2716
1997	65	5	1.3729	0.0102	58.0163
1998	62	10	1.3732	0.0091	56.2716
1999	60	10	1.3747	0.0031	55.3465
2000	65	10	1.3729	0.0102	58.0163
2001	55	15	1.3724	0.0123	51.9474
2002	45	15	1.3712	0.0170	45.7922
2003	47	15	1.3745	0.0039	47.5236
2004	50	20	1.3742	0.0049	49.2799
2005	52	20	1.3726	0.0116	50.1856
2006	55	20	1.3724	0.0123	51.9474
2007	56	30	1.3737	0.0069	52.7832
2008	57	30	1.3750	0.0017	53.6839
2009	58	30	1.3722	0.0131	53.7031
2010	59	40	1.3735	0.0080	54.5258
2011	60	40	1.3747	0.0031	55.3465
2012	60	40	1.3747	0.0031	55.3465

Further, it is also established that the cost minimization with production constraint will be achieved at the phase of increasing returns to scale (IRS) of the enterprise.

$$\alpha + \beta + \gamma + \delta > 1$$

and profit maximization will take place at the phase of decreasing returns to scale (DRS).

$$\alpha + \beta + \gamma + \delta < 1$$

Finally, this paper shows simulations for a given data set, and it is seen that for every phase (IRS, CRS or DRS), there exists a common optimal value of elasticities (α and β) for annual data, which maximizes the revenue. It is also observed from the 3D graphs that the production function is concave for constant and decreasing returns to scale this signifies the enterprise will definitely reach a point where its profit will be maximum for a particular investment. Whereas, it is neither concave or convex for the increasing return to scale, which signifies that the enterprise will reach a point where it is able to minimize the investments made and achieve a target output. Therefore, it can be seen from these 3D graphs and optimal output elasticities that the production model proposed agrees with the current practices in industry.

It has also been established that the behavior of the marginal product of an input depends on the phase of the enterprise. It will rise during increasing returns to scale phase, stay stable for constant returns phase and fall for the last phase of decreasing returns to scale.

An enterprise investing in the data center may be willing to suffer a decrease in production by 10 % or less, at the cost of a major decrement in total investment made on the inputs. This compromise with the yearly production can increase the overall profit of the enterprise drastically. This result needs confirmation by analyzing the different values of inputs for a deviation of revenue ranging from 0–10 %. For example, consider the investments made in the year 2009, (Table 1). The maximum revenue is $36.1788B. If the enterprise decides to decrease investment on servers to $55B and on power & cooling to $28B, the revenue decreases to $34.4354B. Therefore, the output decreases by $1.7434B (that is 4.8 %), but at the same time, the cost of input decreases by a total of $2B. Thus, overall profit increases. Tables 2 and 3 fortify our observation further. Table 2 shows that an increase in the cost of server while keeping power and cooling costs did not impact the revenue. In fact, the revenue increased. Table 3 demonstrates the fact that a decrease in server cost coupled with a twofold increase in power plus cooling cost did not affect revenue in a negative way. Gradient descent ensures that cost is achieved to be minimal by manipulating the elasticities and the increase in revenue is observed accordingly.

The above results are obtained taking into consideration two major cost factors, servers, and power. These results can be extended to any cost model with more number of inputs. In spite of its application in various fields, Cobb-Douglas production function has the following limitations-

- It does not allow identification of the nature of technological progress [35].
- It may suffer from curvature violation and sometimes it requires estimation of many parameters [36].

In the context of the our proposed model, nature of technological progress is not very important as we have not instrumented this in our model. Curvature violation is a major issue, in case of flexible functional form. We expect the global curvature conditions to be consistent with economic theory when estimations of cost, profit, revenue are required from a functional form. Along with that, the task of maintaining the flexibility of functional form is also necessary. Translog function is a generalized form of Cobb-Douglas function, a flexible functional form providing second order approximation. Both the Cobb-Douglas and Translog functions are linear in parameters and can be estimated using least squares methods. It is possible to impose restrictions on the parameters (homogeneity conditions). C-D functions are simplistic, assumes all firms have same production elasticities and that substitution elasticities equal 1. Translog function, which is a generalization of Cobb-Douglas, suffers from curvature violation. But we have not implemented Translog function in our model. Multiple test cases run on several different values don't indicate any curvature violation (cf. Fig. numbers 1–5).

General Notes on Curvature violations (Change of sign):

- Translog function is very commonly used.
- It is a generalization of the Cobb-Douglas function.
- It is a flexible functional form providing a second order approximation.
- Cobb-Douglas and Translog functions are linear in parameters and can be estimated using least squares methods.

$$lnq_i = b_0 + b_1 lnx_{1i} + b_2 lnx_{2i} + v_i + u_i$$

Translog:

$$lnq_i = b_0 + b_1 lnx_{1i} + b_2 lnx_{2i} + 0.5 b_{11} (lnx_{1i})^2 + 0.5 b_{22} (lnx_{2i})^2 + b_{12} lnx_{1i} lnx_{2i} + v_i + u_i$$

The disadvantages of Translog function are listed below:

- It is more difficult to interpret.
- Translog function requires many parameters for estimation.
- It can suffer from curvature violation.

The curvature violation is not evident for Cobb-Douglas in the case of CRS (Constant return to scale) and DRS (Decreasing return to scale). It may arise when we consider IRS (Increasing return to scale). This IRS curvature violation issue can be tackled using Monte Carlo simulation. In theory, it's a technical possibility but the authors have not encountered in this study. In fact, Translog production functions requires estimation of many parameters: K+3+K(K+1)/2 and may suffer from curvature violations. On the other hand, Cobb-Douglas, having linear parameters (in logs) estimated or simulated at par with this assumption, and empirically speaking, we do not see too many curvature violations (on higher orders) for such production/cost functions. If there are, it sometimes arises due to added local (meaning industry-specific, product specific, as opposed to global meaning universalized) restrictions, which we haven't instrumented in our structure and need not be concerned about. In cases of stochastic frontier analysis estimates with Translog (and therefore quadratic forms) the curvature violation is much more feasible than a Cobb Douglas production function. Considering all these points, Cobb-Douglas is a more suitable option in comparison to Translog as a foundation of our mathematical model.

This production model obtained from Cobb-Douglas function might also prove to be helpful in forecasting the revenue of a cloud data center. An enterprise planning to set-up a data center may want to know the approximate revenue for a particular capital expenditure on the inputs. The future revenue estimates may be determined by using this production function [Appendix 7].

Chandrakant et al in their paper Cost Model for Planning, Development and Operation of a Data Center have considered four key cost components (space, power, cooling, operation) in total cost calculation. Individual cost component has been discussed in details based on their dependency on different parameters such as amortization cost, maintenance cost and the influence on the total cost. These apart, other cost factors (Licensing cost, Personnel cost, Operation cost) have also been incorporated in the cost model. In contrast, our proposed model is not dependent on the number of cost components. The Cobb-Douglas model can be expanded as it can accommodate any number of cost factors. Our model not only highlighted cost optimization but also had shown how to achieve profit maximization, revenue maximization. Using real world data set we have established the efficiency of our suggested mathematical model. Jim Gao, in his Machine Learning Applications for Data Center Optimization paper, rigorously used machine learning application to model data center performance and energy efficiency. Various challenges related to data center have been discussed and neural network has been implemented to build the mathematical framework. Energy efficiency is the prime objective of the optimization model and the performance of the model is limited by the quality and quantity of data inputs like any other machine learning applications. In contrast, our model is independent on the training set and methods of training the machine.

Paul J.J. Welfens in his paper 'A Quasi-Cobb Douglas Production Function with Sectoral Progress: Theory and Application to the New Economy' proposed a new mathematical model based on Cobb-Douglas function, which can shed light on process innovation dynamics – this includes a distinction between Harrod neutrality and Solow neutrality of technological progress. The model is an example of endogenous growth approach in which one can study different types of technological progress. It is assumed that Solow type technological progress was determined by ICT capital only. The model has been developed based on two sector production function, where the sectoral Solow progress depends on the hybrid sectoral capital intensity.

$$Y = [B(K'/L)K']^{\beta}[K'']^{\beta''}[AL]^{1-\beta'-\beta''} \tag{20}$$

where $K' =$ ICT -capital $K'' =$ non-ICT capital. K denotes capital and L denotes labor. According to the assumption, sectoral Solow progress B in the sector using ICT capital is associated with K'/L (hybrid sectoral capital intensity). The parameter B(K'/L) implies the ICT sector, characterized by capital-saving technological progress which can be found in various fields such as computer chip (Moore's law which says that the power of computing chip will double within 3 or more recently 2 years) and fiber optical cables. The information and communication technology (ICT) and other few factors are the key reasons behind the technological progress.

Additional remarks & future work

Technological progress

AWS and other data center providers are constantly improving the technology and define the cost of servers as the principle component in the revenue model. For example, AWS [31] spends approximately 57 % of their budget towards servers and constantly improvise in the procurement pattern of three major types of servers. The model used by the authors need to look at such situation i.e include three different input variables for the costs of three different types of servers. In essence, a whole body of future work may include incorporating a lot of input parameters and subject the Cobb Douglas model to multiple levels of the same parameter, namely power cost. This leads to a possibility of interesting analytical exercise, coupled with a full factorial design of an experiment.

Experimental design

As discussed above, input parameters may have multiple levels. It is relevant to study the effects of such input

variables on the revenue in terms of percentage contribution of each variable. An efficient, discrete factorial design could be implemented to study the effects and changes in all relevant parameters regarding revenue. Revenue may be modeled as y, dependent on a constant (market force) and a bunch of input variables, quantitative or categorical in nature. The road-map to design a proper set of experiments for simulation involves the following:

- Develop a model best suited for the data obtained.
- Isolate measurement errors and gauge confidence intervals for model parameters.
- Check the adequacy of the model.

Response variable is the outcome, e.g. revenue output due to factors such as cost, man-hours and the levels of those factors. The primary and secondary factors as well as replication patterns need to be ascertained such that the impact of variation among the entities is minimized. Interaction among the factors need not be ignored. A full factorial design with the number of experiments equal to $\sum_{i=1}^{k} n_i$ would capture all interactions and explain variations due to technological progress, the authors believe. Here, n denotes the number of factors and k stands for different levels each factor may have [37].

Appendix 1

The Lagrangian function for the optimization problem is:

$$\mathcal{L} = y - \lambda(w_1 S + w_2 I + w_3 P + w_4 N - m)$$
$$\mathcal{L} = kS^{\alpha} I^{\beta} P^{\gamma} N^{\delta} - \lambda(w_1 S + w_2 I + w_3 P + w_4 N - m)$$

The first order conditions are:

$$\frac{\partial \mathcal{L}}{\partial S} = k\alpha S^{\alpha-1} I^{\beta} P^{\gamma} N^{\delta} - w_1\lambda = 0 \tag{21}$$
$$\frac{\partial \mathcal{L}}{\partial I} = k\beta S^{\alpha} I^{\beta-1} P^{\gamma} N^{\delta} - w_2\lambda = 0 \tag{22}$$
$$\frac{\partial \mathcal{L}}{\partial P} = k\gamma S^{\alpha} I^{\beta} P^{\gamma-1} N^{\delta} - w_3\lambda = 0 \tag{23}$$
$$\frac{\partial \mathcal{L}}{\partial N} = k\delta S^{\alpha} I^{\beta} P^{\gamma} N^{\delta-1} - w_4\lambda = 0 \tag{24}$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = -(w_1 S + w_2 I + w_3 P + w_4 N - m) = 0 \tag{25}$$

Dividing (22), (23), (24) by (21),

$$I = \frac{\beta}{\alpha}\frac{w_1}{w_2} S$$
$$P = \frac{\gamma}{\alpha}\frac{w_1}{w_3} S$$
$$N = \frac{\delta}{\alpha}\frac{w_1}{w_4} S$$

Substituting these values in (25),

$$S = \frac{m\alpha}{w_1}(1 + \beta + \gamma + \delta) \tag{26}$$

Similarly,

$$I = \frac{m\beta}{w_2}(1 + \alpha + \gamma + \delta) \tag{27}$$
$$P = \frac{m\gamma}{w_3}(1 + \alpha + \beta + \delta) \tag{28}$$
$$N = \frac{m\delta}{w_4}(1 + \alpha + \beta + \gamma) \tag{29}$$

Appendix 2

The Lagrangian function for the optimization problem is:

$$\mathcal{L} = w_1 S + w_2 I + w_3 P + w_4 N - \lambda(f(S, I, P, N) - y_{tar}) \tag{30}$$

The First order conditions are;

$$\frac{\partial \mathcal{L}}{\partial S} = w_1 - \lambda k\alpha S^{\alpha-1} I^{\beta} P^{\gamma} N^{\delta} = 0 \tag{31}$$
$$\frac{\partial \mathcal{L}}{\partial I} = w_2 - \lambda k\beta S^{\alpha} I^{\beta-1} P^{\gamma} N^{\delta} = 0 \tag{32}$$
$$\frac{\partial \mathcal{L}}{\partial P} = w_3 - \lambda k\gamma S^{\alpha} I^{\beta} P^{\gamma-1} N^{\delta} = 0 \tag{33}$$
$$\frac{\partial \mathcal{L}}{\partial N} = w_4 - \lambda k\delta S^{\alpha} I^{\beta} P^{\gamma} N^{\delta-1} = 0 \tag{34}$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = kS^{\alpha} I^{\beta} P^{\gamma} N^{\delta} - y_{tar} = 0 \tag{35}$$

Dividing Eqs. (32), (33) and (34) by (31); we obtain

$$I = \frac{\beta}{\alpha}\frac{w_1}{w_2} S$$
$$P = \frac{\gamma}{\alpha}\frac{w_1}{w_3} S$$
$$N = \frac{\delta}{\alpha}\frac{w_1}{w_4} S$$

Substituting these values in Eq. (35); we obtain

$$\Rightarrow y_{tar} = kS^{\alpha}\left(\frac{\beta}{\alpha}\frac{w_1}{w_2} S\right)^{\beta}\left(\frac{\gamma}{\alpha}\frac{w_1}{w_3} S\right)^{\gamma}\left(\frac{\delta}{\alpha}\frac{w_1}{w_4} S\right)^{\delta}$$
$$\Rightarrow y_{tar} = kS^{\alpha+\beta+\gamma+\delta}\alpha^{-\beta-\gamma-\delta}\beta^{\beta}\gamma^{\gamma}\delta^{\delta} w_1^{\beta+\gamma+\delta} w_2^{-\beta} w_3^{-\gamma} w_4^{-\delta}$$
$$\Rightarrow S^{\alpha+\beta+\gamma+\delta} = k^{-1}\alpha^{\beta+\gamma+\delta}\beta^{-\beta}\gamma^{-\gamma}\delta^{-\delta} w_1^{-\beta-\gamma-\delta} w_2^{\beta} w_3^{\gamma} w_4^{\delta} y_{tar}$$
$$\Rightarrow S = \left(k^{-1}\alpha^{\beta+\gamma+\delta}\beta^{-\beta}\gamma^{-\gamma}\delta^{-\delta} w_1^{-\beta-\gamma-\delta} w_2^{\beta} w_3^{\gamma} w_4^{\delta} y_{tar}\right)^{\frac{1}{\alpha+\beta+\gamma+\delta}}$$
$$\Rightarrow w_1 S = \left(k^{-1}\alpha^{\beta+\gamma+\delta}\beta^{-\beta}\gamma^{-\gamma}\delta^{-\delta} w_1^{\alpha} w_2^{\beta} w_3^{\gamma} w_4^{\delta} y_{tar}\right)^{\frac{1}{\alpha+\beta+\gamma+\delta}} \tag{36}$$

Similarly,

$$w_2 I = \left(k^{-1}\alpha^{-\alpha}\beta^{\beta+\gamma+\delta}\gamma^{-\gamma}\delta^{-\delta}w_1^{\alpha}w_2^{\beta}w_3^{\gamma}w_4^{\delta}y_{tar}\right)^{\frac{1}{\alpha+\beta+\gamma+\delta}} \tag{37}$$

$$w_3 P = \left(k^{-1}\alpha^{-\alpha}\beta^{-\beta}\gamma^{\beta+\gamma+\delta}\delta^{-\delta}w_1^{\alpha}w_2^{\beta}w_3^{\gamma}w_4^{\delta}y_{tar}\right)^{\frac{1}{\alpha+\beta+\gamma+\delta}} \tag{38}$$

$$w_4 N = \left(k^{-1}\alpha^{-\alpha}\beta^{-\beta}\gamma^{-\gamma}\delta^{\beta+\gamma+\delta}w_1^{\alpha}w_2^{\beta}w_3^{\gamma}w_4^{\delta}y_{tar}\right)^{\frac{1}{\alpha+\beta+\gamma+\delta}} \tag{39}$$

The cost for producing y_{tar} units in cheapest way is c, where

$$c = w_1 S + w_2 I + w_3 P + w_4 N \tag{40}$$

From (31), (32), (33) and (34), Eq. (35) can be written as;

$$c = Q\left[w_1^{\alpha}w_2^{\beta}w_3^{\gamma}w_4^{\delta}\right]^{\frac{1}{\alpha+\beta+\gamma+\delta}} y_{tar}^{\frac{1}{\alpha+\beta+\gamma+\delta}} \tag{41}$$

where,

$$Q = k^{\frac{-1}{\alpha+\beta+\gamma+\delta}}\left[\frac{\alpha^{\beta+\gamma+\delta}}{\beta^{\beta}+\gamma^{\gamma}+\delta^{\delta}} + \frac{\beta^{\alpha+\gamma+\delta}}{\alpha^{\alpha}+\gamma^{\gamma}+\delta^{\delta}} + \frac{\gamma^{\alpha+\beta+\delta}}{\alpha^{\alpha}+\beta^{\beta}+\delta^{\delta}} + \frac{\delta^{\alpha+\beta+\gamma}}{\alpha^{\alpha}+\beta^{\beta}+\gamma^{\gamma}}\right]^{\frac{1}{\alpha+\beta+\gamma+\delta}}$$

$$C_{avg} = \frac{C}{y_{tar}} = Q\left[w_1^{\alpha}w_2^{\beta}w_3^{\gamma}w_4^{\delta}\right]^{\frac{1}{\alpha+\beta+\gamma+\delta}} y_{tar}^{\frac{1}{\alpha+\beta+\gamma+\delta}-1}$$

Appendix 3

The conditions for optimization:

$$p\alpha k S^{\alpha-1} I^{\beta} P^{\gamma} N^{\delta} = w_1 \tag{42}$$
$$p\beta k S^{\alpha} I^{\beta-1} P^{\gamma} N^{\delta} = w_2 \tag{43}$$
$$p\gamma k S^{\alpha} I^{\beta} P^{\gamma-1} N^{\delta} = w_3 \tag{44}$$
$$p\delta k S^{\alpha} I^{\beta} P^{\gamma} N^{\delta-1} = w_4 \tag{45}$$

Multiplying these equations with S, I, P and N, respectively-

$$p\alpha k S^{\alpha} I^{\beta} P^{\gamma} N^{\delta} = w_1 S \Rightarrow p\alpha y = w_1 S \tag{46}$$
$$p\beta k S^{\alpha} I^{\beta} P^{\gamma} N^{\delta} = w_2 I \Rightarrow p\beta y = w_2 I \tag{47}$$
$$p\gamma k S^{\alpha} I^{\beta} P^{\gamma} N^{\delta} = w_3 P \Rightarrow p\gamma y = w_3 P \tag{48}$$
$$p\delta k S^{\alpha} I^{\beta} P^{\gamma} N^{\delta} = w_4 N \Rightarrow p\delta y = w_4 N \tag{49}$$

Dividing Eqs. (47), (48) and (49) by (46) following equations are obtained:

$$I = \frac{\beta}{\alpha}\frac{w_1}{w_2}S \tag{50}$$
$$P = \frac{\gamma}{\alpha}\frac{w_1}{w_3}S \tag{51}$$
$$N = \frac{\delta}{\alpha}\frac{w_1}{w_4}S \tag{52}$$

Substituting these values of I, P and N in (42), we get

$$\begin{aligned} & p\alpha k S^{\alpha-1} I^{\beta} P^{\gamma} N^{\delta} = w_1 \\ \Rightarrow & p\alpha k S^{\alpha-1}\left(\frac{\beta}{\alpha}\frac{w_1}{w_2}S\right)^{\beta}\left(\frac{\gamma}{\alpha}\frac{w_1}{w_3}S\right)^{\gamma}\left(\frac{\delta}{\alpha}\frac{w_1}{w_4}S\right)^{\delta} = w_1 \\ \Rightarrow & pkS^{\alpha+\beta+\gamma+\delta-1}\beta^{\beta}\gamma^{\gamma}\delta^{\delta}w_1^{\beta+\gamma+\delta-1}w_2^{-\beta}w_3^{-\gamma}w_4^{-\delta} = 1 \\ \Rightarrow & S = \left(pk\alpha^{1-(\beta+\gamma+\delta)}\beta^{\beta}\gamma^{\gamma}\delta^{\delta}w_1^{\beta+\gamma+\delta-1}w_2^{-\beta}w_3^{-\gamma}w_4^{-\delta}\right)^{\frac{1}{1-(\alpha+\beta+\gamma+\delta)}} \end{aligned} \tag{53}$$

Performing similar calculations the following values of I, P and N are obtained:

$$I = \left(pk\alpha^{\alpha}\beta^{1-(\alpha+\gamma+\delta)}\gamma^{\gamma}\delta^{\delta}w_1^{-\alpha}w_2^{\alpha+\gamma+\delta-1}w_3^{-\gamma}w_4^{-\delta}\right)^{\frac{1}{1-(\alpha+\beta+\gamma+\delta)}} \tag{54}$$

$$P = \left(pk\alpha^{\alpha}\beta^{\beta}\gamma^{1-(\alpha+\beta+\delta)}\delta^{\delta}w_1^{-\alpha}w_2^{-\beta}w_3^{\alpha+\beta+\delta-1}w_4^{-\delta}\right)^{\frac{1}{1-(\alpha+\beta+\gamma+\delta)}} \tag{55}$$

$$N = \left(pk\alpha^{\alpha}\beta^{\beta}\gamma^{\gamma}\delta^{1-(\alpha+\beta+\gamma)}w_1^{-\alpha}w_2^{-\beta}w_3^{-\gamma}w_4^{\alpha+\beta+\gamma-1}\right)^{\frac{1}{1-(\alpha+\beta+\gamma+\delta)}} \tag{56}$$

These values of S, I, P and N are the profit maximizing data center's demand for inputs, as a function of the prices of all the inputs, and of the price of output. Substituting values of S, I, P and N into Eq. (1), we get

$$y = \left(kp^{\alpha+\beta+\gamma+\delta}\alpha^{\alpha}\beta^{\beta}\gamma^{\gamma}\delta^{\delta}w_1^{-\alpha}w_2^{-\beta}w_3^{-\gamma}w_4^{-\delta}\right)^{\frac{1}{1-(\alpha+\beta+\gamma+\delta)}} \tag{57}$$

Appendix 4

Consider the following production function:

$$y = \prod_{i=1}^{n} k x_i^{\alpha_i}$$

To prove:

$$\sum_{i=1}^{n}\alpha_i < 1$$

Consider the profit function:

$$\pi_n = \prod_{i=1}^{n} k x_i^{\alpha_i} - \sum_{i=1}^{n} w_i x_i$$

w_i: Unit cost of inputs

Profit maximization is achieved when: $p\frac{\partial f}{\partial x_i} = w_i$. Deriving the condition for optimization:

$$pk\frac{\alpha_1}{x_1}\prod_{i=1}^{n} x_i^{\alpha_i} = w_1 \quad (58)$$

$$pk\frac{\alpha_2}{x_2}\prod_{i=1}^{n} x_i^{\alpha_i} = w_2 \quad (59)$$

$$\vdots$$

$$pk\frac{\alpha_n}{x_n}\prod_{i=1}^{n} x_i^{\alpha_i} = w_n \quad (60)$$

Multiplying these equations with x_i, respectively-

$$p\alpha_1\prod_{i=1}^{n} k x_i^{\alpha_i} = w_1 x_1 \Rightarrow p\alpha_1 y = w_1 x_1 \quad (61)$$

$$p\alpha_2\prod_{i=1}^{n} k x_i^{\alpha_i} = w_2 x_2 \Rightarrow p\alpha_2 y = w_2 x_2 \quad (62)$$

$$\vdots$$

$$p\alpha_n\prod_{i=1}^{n} k x_i^{\alpha_i} = w_n x_n \Rightarrow p\alpha_n y = w_n x_n \quad (63)$$

Dividing Eqs. (62) to (63) by (61), following equations are obtained:

$$x_2 = \frac{\alpha_2}{\alpha_1}\frac{w_1}{w_2}x_1$$
$$x_3 = \frac{\alpha_3}{\alpha_1}\frac{w_1}{w_3}x_1$$
$$\vdots$$
$$x_{n-1} = \frac{\alpha_{n-1}}{\alpha_1}\frac{w_1}{w_{n-1}}x_1$$
$$x_n = \frac{\alpha_n}{\alpha_1}\frac{w_1}{w_n}x_1$$

Substituting these values of x_i in Eq. (58),

$$pk\frac{\alpha_1}{x_1}\prod_{i=1}^{n} x_i^{\alpha_i} = w_1$$
$$\Rightarrow pk\alpha_1 x_1^{\alpha_1-1}\left(\frac{\alpha_2}{\alpha_1}\frac{w_1}{w_2}x_1\right)^{\alpha_2}\left(\frac{\alpha_3}{\alpha_1}\frac{w_1}{w_3}x_1\right)^{\alpha_3}\ldots\left(\frac{\alpha_{n-1}}{\alpha_1}\frac{w_1}{w_{n-1}}x_1\right)^{\alpha_{n-1}} \times\left(\frac{\alpha_n}{\alpha_1}\frac{w_1}{w_n}x_1\right)^{\alpha_n} = w_1$$
$$\Rightarrow pk x_1^{(\alpha_1+\alpha_2+\ldots+\alpha_n)-1}\alpha_1^{1-(\alpha_2+\alpha_3+\ldots+\alpha_n)}\alpha_2^{\alpha_2}\ldots \alpha_n^{\alpha_n} w_1^{-1+(\alpha_2+\alpha_3+\ldots+\alpha_n)} w_2^{-\alpha_2}\ldots w_n^{-\alpha_n} = 1$$
$$\Rightarrow x_1 = \left(pk\alpha_1^{1-(\alpha_2+\alpha_3+\ldots+\alpha_n)}\alpha_2^{\alpha_2}\ldots\alpha_n^{\alpha_n} w_1^{-1+(\alpha_2+\alpha_3+\ldots+\alpha_n)} w_2^{-\alpha_2}\ldots w_n^{-\alpha_n}\right)^{\frac{1}{1-(\alpha_1+\alpha_2+\ldots+\alpha_n)}}$$

Performing similar calculations following values of x_i, $(i >= 2)$ are obtained,

$$x_2 = \left(pk\alpha_2^{1-(\alpha_1+\alpha_3+\ldots+\alpha_n)}\alpha_1^{\alpha_1}\ldots\alpha_n^{\alpha_n} w_2^{-1+(\alpha_1+\alpha_3+\ldots+\alpha_n)} w_1^{-\alpha_1}\ldots w_n^{-\alpha_n}\right)^{\frac{1}{1-(\alpha_1+\alpha_2+\ldots+\alpha_n)}}$$

$$\vdots$$

$$x_n = \left(pk\alpha_n^{1-(\alpha_1+\alpha_2+\ldots+\alpha_{n-1})}\alpha_1^{\alpha_1}\ldots\alpha_{n-1}^{\alpha_{n-1}} w_n^{-1+(\alpha_1+\alpha_2+\ldots+\alpha_n)} w_2^{-\alpha_2}\ldots w_{n-1}^{-\alpha_{n-1}}\right)^{\frac{1}{1-(\alpha_1+\alpha_2+\ldots+\alpha_n)}}$$

Substituting values of x_i in production function,

$$y = \left(kp^{(\alpha_1+\alpha_2+\ldots+\alpha_n)}\alpha_1^{\alpha_1}\alpha_2^{\alpha_2}\ldots\alpha_n^{\alpha_n} w_1^{-\alpha_1} w_2^{-\alpha_2}\ldots w_n^{-\alpha_n}\right)^{\frac{1}{1-(\alpha_1+\alpha_2+\ldots+\alpha_n)}}$$

y increases in price of its output and decreases in price of its inputs iff:

$$1 - \sum_{i=1}^{n}\alpha_i > 0$$
$$\sum_{i=1}^{n}\alpha_i < 1$$

Therefore decreasing returns to scale, is validated

Appendix 5

Matlab Code for Increasing return to scale:

A = [11; −1 − 1; −10; 0 − 1];
b = [1.9; −1.1; −0.1; −0.1];
x0 = [0.4; 0.1];
[*x*, *fval*] = *fmincon*(@*cobbfun*, *x*0, *A*, *b*);
function f = cobbfun(x)
%Cobb-Douglas function with k = 1
%f is a representation of Cobb-Douglas function.
% x(1), x(2) is representing elasticity constant of new server spending and power/cooling cost respectively.
$f = -62^{x(1)} * 5^{x(2)}$;
end

Matlab Code for Constant return to scale:

```
A = [−10; 0 − 1];
b = [ −0.1; −0.1];
Aeq = [11];
beq = [1]
x0 =[0.4; 0.1];
[x, fval] = fmincon(@cobbfun, x0, A, b, Aeq, beq);
function f = cobbfun(x)
%Cobb-Douglas function with k = 1
% f is a representation of Cobb-Douglas function.
%x(1), x(2) is representing elasticity constant of new server spending and power/cooling cost respectively.
f = −62^x(1) ∗ 5^x(2);
end
```

Matlab decreasing returns to scale:

```
A = [11; −10; 0 − 1];
b = [0.9; −0.1; −0.1];
x0 = [0.4; 0.1];
[x, fval] = fmincon(@cobbfun, x0, A, b);
function f = cobbfun(x)
%Cobb-Douglas function with k = 1
%f is a representation of Cobb-Douglas function.
%x(1), x(2) is representing elasticity constant of new server spending and power/cooling cost respectively.
f = −62^x(1) ∗ 5^x(2);
end
```

Appendix 6

A c^2 function $f : U \subset R^n \rightarrow R$ defined on a convex open set U is concave if and only if the Hessian matrix $D^2 f(x)$ is negative semi-definite for all $x \in U$. A matrix H is negative semi-definite if and only if it's $2^n - 1$ principal minors alternate in sign so that odd order minors are less than equal to 0 and even order minors are greater than equal to 0. Cobb-Douglas function for 2 inputs is:

$$f(x, y) = cx^a y^b$$

Its Hessian is

$$\begin{bmatrix} a(a-1)cx^{a-2}y^b & abcx^{a-1}y^{b-1} \\ abcx^{a-1}y^{b-1} & b(b-1)cx^a y^{b-2} \end{bmatrix}$$

$$\Delta_1 = a(a-1)cx^{a-2}y^b$$
$$\Delta_1 = b(b-1)cx^a y^{b-2}$$
$$\Delta_2 = abc^2 x^{2a-2} y^{2b-2}(1-(a+b))$$

Condition for a function to be concave,

$$\Delta_1 \leq 0$$
$$\Delta_2 \geq 0$$

For decreasing and constant returns to scale: $a + b \leq 1$ Therefore,

$$a \leq 1,\ b < 1$$
$$\Rightarrow (a-1) \leq 0$$
$$\Rightarrow \Delta_1 \leq 0$$

$$(1-(a+b)) \geq 0$$
$$\Rightarrow \Delta_2 \geq 0$$

Both conditions for concave function are satisfied by decreasing and constant returns to scale. Therefore, the graph obtained for decreasing and constant returns to scale is concave, while for increasing returns the graph is neither concave nor convex.

Appendix 7

Forecasting the revenue of the data center

Cobb-Douglas production function [28] that relates the 4 inputs to the output of the IaaS data center is:

$$y = kS^\alpha I^\beta P^\gamma N^\delta \tag{64}$$

y: total production of an IaaS data center
S: total number of servers
I: total cost of infrastructure
P: total unit watt of power drawn
N: total mbps data
k: total factor productivity
α, β, γ and δ are the output elasticities of servers, infrastructure, power drawn and network respectively.
In order to find the values of the constants and k, the Method of Least Squares [38] is applied. For this the equation is linearized, by taking the natural log of both sides. Therefore equation becomes:

$$logY = logk + \alpha logS + \beta logI + \gamma logP + \delta logN$$

Replacing the above values with-

$$Y' = logY;\ k' = logk;\ S' = logS;\ I' = logI;\ N' = logN$$

$$Y' = k' + \alpha S' + \beta I' + \gamma P' + \delta N'$$

Let Y_i' be the value of Y' corresponding to the value S_i', I_i', P_i' *and* N_i' of S', I', P' *and* N' respectively.

$$Y_i' = k' + \alpha S_i' + \beta I_i' + \gamma P_i' + \delta N_i'$$

The value of Y_i' is the estimated value of given y_i corresponding to S_i, I_i, P_i *and* N_i. Let,

$$S = \sum (y_i - Y_i)^2$$

$k', \alpha, \beta, \gamma, \delta$ are so determined that S is minimum. The necessary conditions for this are:

$$\frac{\partial S}{\partial k'} = 0;\ \frac{\partial S}{\partial \alpha} = 0;\ \frac{\partial S}{\partial \beta} = 0;\ \frac{\partial S}{\partial \gamma} = 0;\ \frac{\partial S}{\partial \delta} = 0$$

These 5 equations are used for determining the values of $\alpha, \beta, \gamma, \delta$ *and* k. Substituting these values of $\alpha, \beta, \gamma, \delta$ *and* k

in Eq. (64), the equation of curve of best fit for the economic data of the established data centers is obtained.

Now, if an enterprise planning to set-up a data center can predict its approximate revenue for a particular set of investments on the 4 inputs.

Appendix 8

3D Plot Code for Decreasing returns to scale:

```
syms xm ym;
dy = 0.001;
dx = 0.001;
f = 7000000.^x m. * 4700000.^y m;
[xm, ym] = meshgrid(.1 : dx : .9, .1 : dy : .9);
f(xm + ym > 0.9) = NaN;
surf(xm, ym, f, 'EdgeColor', 'none')
```

3D Plot Code for Constant returns to scale:

```
syms xm ym;
dx = 0.001;
dy = 0.001;
f = 1600.^x m. * 270.^y m;
[xm, ym] = meshgrid(.1 : dx : .9, .1 : dy : .9);
f(xm + ym > 1) = NaN;
surf(xm, ym, f, 'EdgeColor', 'none')
```

3D Plot Code for Increasing returns to scale:

```
syms xm ym;
dx = 0.001;
dy = 0.001;
f = 65.^x m. * 5.^y m;
[xm, ym] = meshgrid(.1 : dx : 1.9, .1 : dy : 1.9);
f(xm + ym < 1.1) = NaN;
f(xm + ym > 1.9) = NaN;
surf(xm, ym, f, 'EdgeColor', 'none')
```

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Snehanshu has been instrumental in theorizing and abstracting the models including the Lagrangian constained optimization. Avantika and Nandiata Dwivedi went through the models rigorously and worked out the validations theoretically as per Snehanshu's instructions. Jyotirmoy was responsible for simulation and data validation. Ranjan Roy and Anand Narasimhamurthy brought in the economics in to the modeling in terms of interpretation, especially increasing and decreasing returns to scale. All authors read and approved the final manuscript.

Acknowledgements
The authors wish to thank Professor B.P. Vani, Institute for Social and Economic Change, for carefully reading the manuscript and offering insightful suggestions. The authors are grateful to Dr. Saibal Kar, Centre for Studies in Social Sciences, Calcutta and IZA, Bonn for reading the paper and giving useful feedback.

Author details
[1]Center for Applied Mathematical Modeling and Simulation (CAMMS) & Department of Computer Science and Engineering, PESIT-BSC, Bangalore, 560100, India. [2]GE Healthcare, Bangalore, 560066, India. [3]BITS Pilani, Hyderabad, India. [4]State Resource Centre, Chattisgarh, India.

Received: 15 June 2015 Accepted: 18 December 2015
Published online: 19 January 2016

References

1. Patel CD, et al (2005) Cost Model for Planning, Development and Operation of a 1209 Data Center:4–17
2. André Barroso LA, Hölzle U The Datacenter as a Computer An Introduction to the Design of Warehouse-Scale Machines
3. Hurwitz J, Bloor R, Kaufman M, Halper F (2010) Cloud computing for dummies. Wiley Publishing, Inc, Indianapolis, Indiana
4. Schaapman P (2012) Data Center Optimization Bringing efficiency and improved services to the infrastructure
5. Better Power Management for Data Centers and Private Clouds. ftp://download.intel.com/newsroom/kits/xeon/e5/pdfs/Intel_Node_Manager-SolutionBrief.pdf. Accessed on 12/12/2014
6. Koomey J (2011) Growth in Data center electricity use 2005 to 2010. Analytics Press, Oakland, CA
7. Uptime Institute (2013) Data Center Industry Survey. https://uptimeinstitute.com/research-publications/asset/18. Accessed on 12/3/2015
8. Google's Green DataCenters (2011) Network POP Case Study. https://static.googleusercontent.com/media/www.google.com/en//corporate/datacenter/dc-best-practices-google.pdf. Accessed on 13/4/2015
9. Retrieved from. https://www.google.com/about/datacenters/efficiency/internal/. Accessed on 23/5/2015
10. Gao J (2014) Machine Learning Applications for Data Center Optimization Jim Gao, Google
11. How much do Google data centers cost [WWW document]. http://www.datacenterknowledge.com/google-data-center-faq-part-2/. Accessed on 12/3/2015
12. How Big is Apple's North Carolina Data Center? [WWW document]. http://www.datacenterknowledge.com/the-apple-data-center-faq/. Accessed on 8/6/2015
13. How much Does Facebook Spend on Its Data Centers? [WWW document]. http://www.datacenterknowledge.com/the-facebook-data-center-faq-page-three/. Accessed on 4/5/2015
14. A Look Inside Amazon's Data Centers [WWW document]. http://www.datacenterknowledge.com/archives/2011/06/09/a-look-insideamazons-data-centers. Accessed on 4/4/2015
15. Nicholas M (2011) HP Updates Data Center Transformation Solutions
16. Drummer R As 'Big Data' Moves Into The Cloud, Demand for Data Center 1207 Space Soars. http://www.bigdatacow.com/content/%E2%80%98big-data%E2%80%99-moves-cloud-demand-data-center-space-soars. Accessed on 2/4/2015
17. Basmadjian R, De Meer H, Lent R, Giuliani G (2012) Cloud computing and its interest in saving energy: the use case of a private cloud. J Cloud Comput Adva Syst Appl 1:5. doi:10.1186/2192-113X-1-5
18. Fan X, Weber WD, Barroso LA (2007) Power provisioning for a warehouse-sized computer. ACM SIGARCH Comput Archit News 35(2):13–23
19. Saravana M, Govidan S, Lefurgy C, Dholakia A (2009) Using on-line power modeling for server power capping. In: Workshop on Energy-Efficient Design 2009, University of Texas and IBM
20. Hamilton J (2008) Cost of Power in Large-Scale Data Centers [WWW 1190 document]. http://perspectives.mvdirona.com/2008/11/cost-of-power-1191in-large-scale-data-centers/. Accessed on 9/3/2015
21. Hassani A (2012) Applications of Cobb-Douglas Production Function in Construction Time-Cost Analysis
22. Hossain Md. M, Majumder AK, Basak T (2012) An Application of Non-Linear Cobb-Douglas Production Function to Selected Manufacturing Industries in Bangladesh. Open Journal of Statistics 2(4)
23. Wu D-M (1975) Estimation of the Cobb-Douglas Production Econometrica 43(4). doi:10.2307/1913082

24. Rappos E, Robert S, Riedi RH (2013) A Cloud Data Center Optimization Approach Using Dynamic Data Interchanges. IEEE 2nd International Conference on Cloud Networking (CloudNet)
25. Speitkamp B, Bichler M (2010) A mathematical programming approach 178 2013 IEEE 2nd International Conference on Cloud Networking (CloudNet): Short Paper for server consolidation problems in virtualized data centers. In: IEEE Transactions on Services Computing Vol. 3, no. 4. pp 266–278
26. Jiao L, Li J, Xu T, Fu X (2012) Cost optimization for online social networks on geo-distributed clouds. In: Network Protocols (ICNP), 20th IEEE International Conference on. pp 1–10
27. Rao L, Liu X, Xie L, Liu W (2010) Distributed Internet Data Centers in a Multi-Electricity-Market Environment. INFOCOM,Proceedings IEEE
28. Cobb CW, Douglas PH (1928) A Theory of Production. Am Econ Rev 18(Supplement):139–165
29. Tan BH (2008) Cobb-Douglas Production Function [Online Database]. http://docentes.fe.unl.pt/jamador/Macro/cobb-douglas.pdf. Accessed on 8/3/2015
30. Huang K-W, Wang M (2009) Firm-Level Productivity Analysis for Software as a Service Companies. In: Proceedings of the 30th International Conference on Information Systems, Phoenix, USA, December 15-18. paper 21
31. Greenberg A, Hamilton J, Maltz DA, Patel P (2009) The cost of a cloud: research problems in data center networks. Comput Commun Rev 39(1):68–73
32. Patel C, Shah A (2005) Cost Model for Planning, Development and Operation of a Data Center. Internet Systems and Storage Laboratory, HP Laboratories Technical Report, HPL-2005-107R1, Palo Alto, CA
33. Fletcher R (2000) Practical methods of optimization. Wiley. ISBN 978-0-471-49463-8
34. Preston McAfee R, Stanley Johnson J (2005) Professor of Business, Economics & Management, California Institute of Technology. Introduction to Economic Analysis
35. Coelli TJ, Rao DSP, O'Donnell CJ, Battese GE (2005) An Introduction to Efficiency and Productivity Analysis. Springer
36. Welfens PJJ (2005) A Quasi-CobbDouglas production function with sectoral progress: Theory and application to the new economy. EIIW Discussion Paper, No. 132, University of Wuppertal
37. Raj Jain The Art of Computer Systems Performance Analysis, Techniques for Experimental Design, Measurement, Simulation, and Modeling. Wiley publishers, Delhi
38. Miller SJ (2006) The Method of Least Squares Mathematics Department Brown University, Providence: Brown University:1–7