Cloud Computing Operations Research

Ilyas Iyoob[1], Emrah Zarifoglu[2], A. B. Dieker[3]

Abstract

This paper argues that the cloud computing industry faces many decision problems where operations research could add tremendous value. To this end, we provide an OR perspective on cloud computing in three ways. First, we compare the cloud computing process with the traditional pull supply chain and introduce the *Cloud IT Supply Chain* as the system of moving information from suppliers to consumers through the cloud network. Second, based on this analogy, we organize the cloud computing decision space by identifying the problems that need to be solved by each player in the supply chain, namely (1) cloud providers, (2) cloud consumers, and (3) cloud brokers. We list the OR problems of interest from each player's perspective and discuss the tools which may need to be developed to solve them. Third, we survey past and current research in this space and discuss future research opportunities in Cloud Computing Operations Research.

Keywords: Cloud Computing, Operations Research, Cloud IT, Green IT, Supply Chain

1. Introduction

Many of today's Information Technology (IT) applications rely on access to state-of-the-art computing facilities. For instance, as business decisions are increasingly driven by (data) analytics, the practice of operations research and business analytics becomes inherently intertwined with the management of IT resources. In response to the resulting demand for flexible computing resources, cloud computing has taken the IT industry by storm over the past few years. According to the National Institute of Standards and Technology (NIST), cloud computing is "... a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (for example, networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service-provider interaction" (Mell and Grance 2011). Cloud computing is a service where computing is provided as a commodity, much akin to electricity or cable television. Thus, cloud computing is not about a specific technology; rather it is a step in the commoditization of IT enabled by technological advances.

The paradigm shift from *IT as a product* to *IT as a service* and the accompanying flexibility gives rise to a vast array of resource management decisions, and this paper discusses the decision spaces of the cloud computing stakeholders. It is important to optimize cloud computing for everyone in the business of cloud, both from a cost perspective and a green (sustainability) perspective. It is our objective to argue that the stakeholders could benefit from Operations Research due to the nature of the problems they face, and that similarly the OR community could benefit from an emerging field which has the potential to drive new research questions. Even though the process of IT commoditization is not yet complete, operations research can already be applied to cloud computing as it stands now. In fact, operations research can be used every step of the way.

There are several reasons why, from an OR point of view, cloud computing is fundamentally different from traditional data centers and server farms. The aforementioned flexibility of cloud computing has

[1] Gravitant and University of Texas at Austin
[2] IBM
[3] Georgia Institute of Technology

resulted in a large degree of specialization among service providers, which has decoupled and significantly widened the decision space. Moreover, cloud computing enables decisions on an unprecedented level of granularity, as a result of advances in technology that allow for great flexibility in fulfilling resource requests.

We draw connections between OR and cloud computing in three ways. First, we compare the cloud computing process to a familiar operations research concept in network modeling – the traditional pull supply chain model. In both cloud computing as well as the traditional pull supply chain models, the end consumer requests services from a service provider who aggregates services by collaborating with other providers through a network. Finally, the service provider responds to the end consumer through another network. Due to these similarities, we define the Cloud IT Supply Chain as the flow of IT services from providers to consumers through the cloud. We also discuss the differences between the pull supply chain model and the Cloud IT Supply Chain, and describe the challenges specific to the cloud.

A second way in which we provide an OR perspective on cloud computing is by identifying the optimization problems that need to be solved by the different players in the Cloud IT Supply Chain: providers, consumers, and brokers. By categorizing these problems according to their role in the supply chain, this allows us to organize the cloud computing decision space and to discuss which OR tools seem appropriate for each decision problem.

Third, we advocate an interplay between cloud computing and OR by arguing that cloud computing can stimulate and drive new research in OR. To this end, we use our classification of decision problems for cloud stakeholders to survey the status of current and past research on a problem-by-problem basis. We identify areas in which quite a bit of research is being done and those in which not much work has been done so far. Based on this analysis, we discuss the prospects of Cloud Computing Operations Research.

We believe that OR techniques can make a difference in the practice of cloud computing, as confirmed by the facts that (1) the cloud computing market has grown to several billion U.S. dollars a year and is still rapidly growing (this is a conservative estimate; estimates vary greatly by source), so there is a significant potential for savings, (2) usage data is continuously and often automatically collected, so that informed decisions can be made, and (3) migration to the cloud is fast, so in a variety of settings it can be relatively easy to implement decisions or to provide a test bed.

This paper is organized as follows. Section 2 gives some background on cloud computing and gives an introduction to the decision space. We present the Cloud IT Supply Chain in Section 3 in order to organize this space. Sections 4-6 describe the decisions from the perspective of each of the three aforementioned players in the Cloud IT Supply Chain. As part of our treatment, we summarize past research and we identify areas where more work is needed. We also discuss areas where cloud computing could potentially drive new theoretical developments in operations research. Section 7 discusses the prospects for this emerging field.

2. Cloud Computing and decision making

This section gives a brief summary of the benefits and challenges of cloud computing, with a focus on decision making. It does not serve as a comprehensive introduction to cloud computing; for background on different aspects of cloud computing, we refer to (Armbrust et al. 2010, Menascé and Ngo 2009, Foster et al 2008) instead. We discuss various major decisions involved in ‘going cloud’. In practice, most consumers make these decisions by choice as opposed to using analytical tools. Still, this section allows us to clarify what cloud computing is, and to clear up any potential confusion surrounding our use of this popular terminology.

As it stands today, there are several reasons why any organization or individual would want to "go cloud". Cloud computing brings two unique features for enterprises: elasticity and flexibility. High level of elasticity in the cloud allows enterprises to scale their IT resources up and down within very short amounts of time. Flexibility presents a large set of options for an enterprise to configure its IT resources, such as operating systems, software, memory, CPU, etc. These features are enabled by virtualization, which allows the control of computing resources to be physically separated from the resources themselves.

The benefits of cloud computing has led increasingly many enterprises to outsource their IT resources to the cloud, despite some challenges. This outsourcing frees limited resources of enterprises to be concentrated on their core functions. However, losing sole control on issues like resource availability, security, privacy, confidentiality and management of IT resources to outer partners is of concern to some businesses. Another challenge is that part of IT resource management and access is typically shared by regular work force in the enterprise, which makes management related considerations in the enterprise more complex. Also, transition to the cloud employs new privacy protection paradigms and methods that will allow third party users to process data without the need of accessing it (Gentry 2010). In spite of these challenges and concerns, consumer demand for cloud services is growing by the day. Apparently groups of consumers are confident that providers are well-equipped to handle some of these issues if they arise, and this process can be closely governed by stricter Service Level Agreements (SLAs) between the consumer and the provider.

A number of options are available to consumers as they embark on their journey to the cloud, and the rest of this section summarizes the key decisions they face. The first major decision for a consumer is the level of IT outsourcing (Figure 1). There are three major service delivery options;

1. outsource only computing infrastructure (also known as **Infrastructure-as-a-Service** or IaaS) from Amazon, GoGrid, Rackspace etc., or
2. outsource the development platform in addition to the infrastructure (known as **Platform-as-a-Service** or PaaS) from Google App Engine, Microsoft Azure, Force.com, or
3. outsource the entire software including the platform and the infrastructure (known as **Software-as-a-Service** or SaaS) from Google apps, salesforce.com etc.

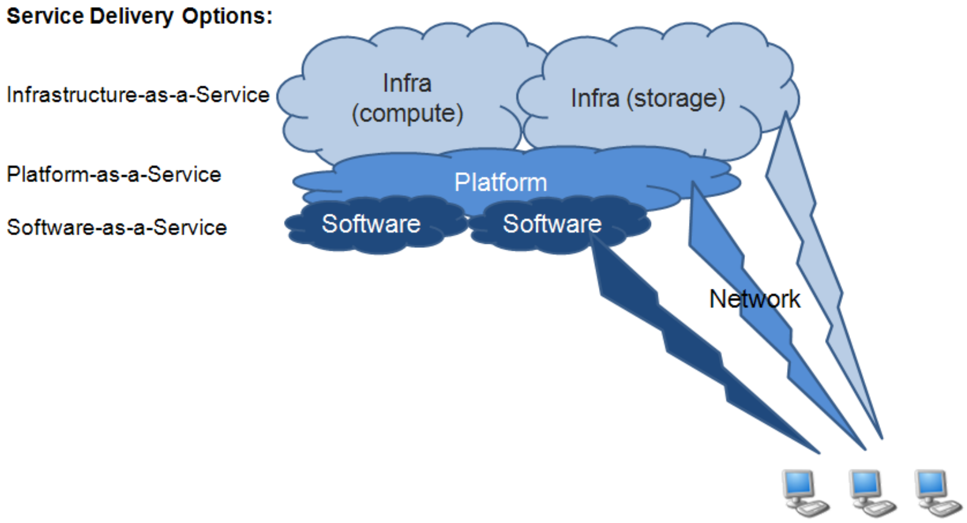

Figure 1. Cloud Service Delivery Options

Based on the service delivery option selected, the consumer has a number of other options for deployment, functionality, infrastructure location, and data location (Figure 2). These options have functional and security impacts on the consumer as well as a major financial impact on the consumer. There may be a role for optimization models to identify the best combination of options that works best for a consumer while satisfying budget and customer satisfaction constraints, but often only a subset of the options are available to cloud consumers due to the nature of their business.

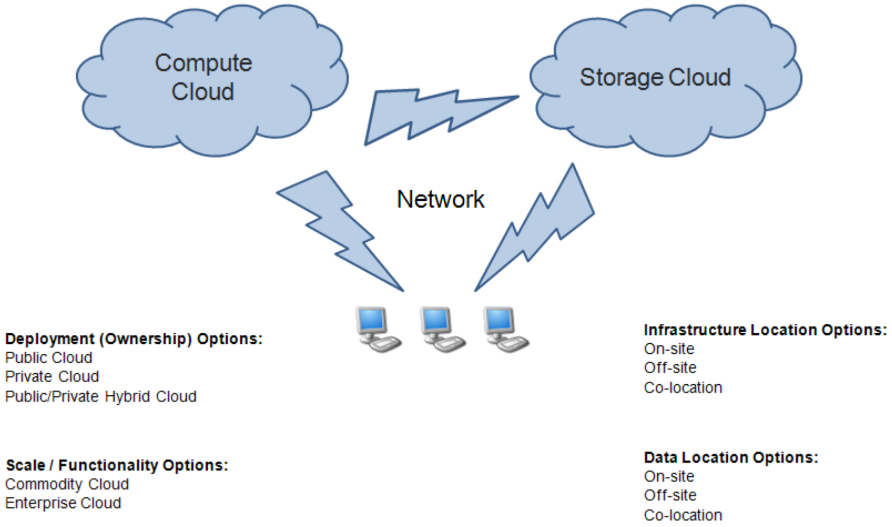

Figure 2. Other Major Cloud Decisions

Deployment (Ownership) Options

In terms of ownership, consumers have the option of going with;

- Public Cloud – which does not involve any ownership and consumers share a public cloud with other consumers. This option is for companies or individuals who want elasticity and flexibility at a low operational cost.
- Private Cloud – which is owned and operated by the consumer. This option is for companies large enough to share resources between departments within their organization. Here the consumer purchases hardware with virtualization technology as a capital expenditure that is amortized over time.
- Public/Private Hybrid – which is a combination of the first two options. This is a commonly chosen option where the consumer purchases a small piece of hardware with virtualization technology for some of its applications and deploys the rest of its applications on a public cloud. This also requires some capital expenditure.

Scale / Functionality Options

Based on scale and functionality of the applications, consumers can choose either a

- Commodity Cloud – where they are billed by the hour and the public can access their servers, or
- Enterprise Cloud – where they are billed on a monthly basis with private access to servers that may or may not be accessible to the public. This option is selected by consumers that require higher service levels from the cloud providers.

Infrastructure Location Options

Regarding the location of the infrastructure, consumers can opt for

- On-site – where the hardware is stored in the consumer's data center,
- Off-site – where the hardware is stored at the provider's data center, or

- Co-location – where the hardware is stored at a secure neutral location.

Data Location Options
Similar to infrastructure, consumers can specify data to be stored on-site, off-site, or at a co-located facility.

Note that these decisions are highly interdependent. For example, if a consumer chooses a private cloud deployment, then they will not have any scale / functionality options because these are only relevant to public clouds. Moreover, their infrastructure and data can only be located on-site or at a co-located facility. On the other hand, if a consumer chooses a public cloud deployment, then they can choose either a commodity or enterprise cloud, but infrastructure and data may be restricted to off-site locations only.

3. Cloud IT Supply Chain

This section introduces the Cloud IT Supply Chain, which provides a direct connection between OR and cloud computing. The supply chain analogy helps in identifying the possible OR problems in the cloud decision space. Furthermore, the existing approaches to solving each of these supply chain-like problems are discussed in detail in each of the following sections, thus drawing a distinction between those problems that can be solved by existing supply chain methods and those that require extensive research and new methodology.

The first step in applying operations research to the cloud is to understand how the cloud works. Let us take an example of a retail company that employs IaaS cloud services. Suppose the manager wants to run a query on the customer database to identify all the customers that have a birthday on a specific day so that special coupons could be sent to these customers. First, the manager accesses a virtual machine (VM) on the compute cloud (example GoGrid) to create a request for the query to be run. The VM initiates the query and accesses the data from the disks in the storage cloud (example Amazon). Then the VM executes the query code on the dataset within the compute cloud and aggregates the information in a suitable format. Now that the report is ready, the VM sends a report of all the birthday customers through the network to the manager's computer. The VM also updates the storage cloud with query results. Note that in the case of PaaS or SaaS, the user goes through a similar process with a different interface (e.g., using a virtual host or a URL).

This process can be generalized as follows (Figure 3):

1. Client *generates request* through network to compute cloud
2. VM on compute cloud *gets data* from storage cloud
3. VM on compute cloud *aggregates information* for client
4. VM on compute cloud *responds with information* to client through network.

As we go through this example, we start noticing similarities with more traditional supply chain models. Consider the **pull supply chain model** made famous by Dell's direct business model. When customers place an order (on the phone or online) with Dell.com, Dell collects the parts from different suppliers. Then, Dell assembles the parts to build the product for the customer. Finally, Dell ships the product through physical networks (by land or air) to the customer. Dell also updates the expected inventory levels on the suppliers end for future forecasts. This process is as follows (Figure 4):

1. Customer *generates order* through phone, web, etc. with Dell
2. Dell Plant *gets parts* from suppliers
3. Dell Plant *assembles product* for customer
4. Dell Plant *ships product* to customer through ground/air transportation.

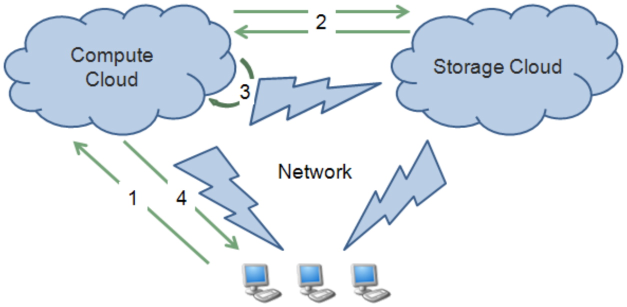

Figure 3. IaaS example

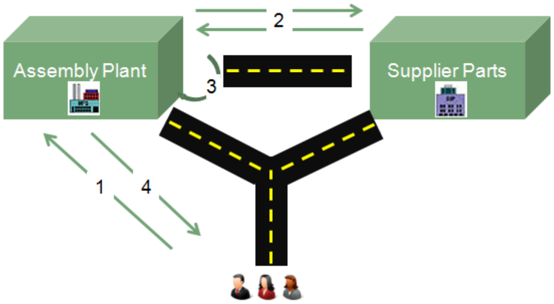

Figure 4. Pull Supply Chain example

Since the cloud execution process closely resembles the pull supply chain model, we refer to the cloud execution process as the **Cloud IT Supply Chain**, which we define as the system of organizations, people, technology, activities, information and resources involved in moving IT services from supplier to consumer through the Cloud.

Similarities and Differences

A number of similarities exist between the Cloud IT Supply Chain and the Traditional Pull Supply Chain (Table 1) especially in terms of specialization of labor, infrastructure ownership, customer experience and dependency on the network.

Specialization of Labor: In both cases, each player in the supply chain can focus on their business value instead of trying to become a jack of all trades but master of none. Dell focuses on supplier relationships and logistics instead of hardware manufacturing. Similarly, compute cloud providers can focus on energy conservation through better load balancing while storage cloud providers can focus on better utilization through storage partitioning. Moreover, cloud consumers can focus on value added services instead of having to manage large data centers.

Ownership: Since each player is focusing on their own skill sets, they only need to own their portion of the overall infrastructure. Dell does not need to own all the parts that make up their product. They also need not purchase any inventory, which significantly cuts costs. In the same way, compute cloud providers need not own storage space while storage cloud providers need not own much computing capacity. Cloud consumers also do not need to commit large portions of their annual budget to data center assets.

Customer Experience: Customers enjoy staying out of the logistics and still have access to resources they need. Dell customers can order a customized product and have it shipped to their homes within 7-14 days. Similarly, cloud consumers can remote login to a virtual machine on a virtual data center and instantly have access to effectively unlimited compute power and storage.

Network Risk: In both the Cloud IT Supply Chain as well as the Traditional Pull Supply Chain, end customers are highly dependent on the network. Dell customers depend on the transportation network to receive their products, and cloud consumers depend heavily on the internet and intranet to fulfill their computing needs.

Table 1. Cloud IT Supply Chain vs Traditional Pull Supply Chain

Similarities	Differences
Specialization of Labor: Each player in the supply chain can focus on their own skill set	***Cost of Risk:*** Different players take on the risk of failure
Ownership: Each player in the supply chain needs to own only their portion of the assets	***Lead Time:*** The time from order placement to receipt of goods and services varies significantly
Customer Experience: End customers can enjoy the product but can stay out of the logistics	
Network Risk: End customers are highly dependent on the network to receive their goods and services	

However, there are also some key differences.

Cost of Risk: Even though both Dell customers and cloud consumers are dependent on the network, there is a big difference in who is responsible if something goes wrong. Dell has to absorb the cost of not having enough capacity on the transportation network or for having to pay high prices to shipping companies. On the other hand, in the Cloud IT Supply Chain it is the consumers who have to absorb the cost of lost business due to network outages. This is also because of the fact that in the Cloud IT Supply Chain, a portion of the network is owned and managed by the consumer, which is their local area network. This is usually not the case in the Traditional Pull Supply Chain.

Lead Time: Dell customers have to wait approximately 7-14 days to receive a product after placing an order whereas cloud consumers do not need to wait more than 3-4 seconds to get a response after executing a command on the virtual machine. The only exception to this is when cloud consumers run batch jobs on the cloud, but even then the processing time is usually no more than a day.

Perspectives

We can use the Supply Chain analogy to organize the cloud computing decision space. There are three major players in the Cloud IT Supply Chain:

1. Providers (compute, storage, and network)
2. Consumers (individuals and enterprises)
3. Brokers (equivalent of third party logistics)

Providers includes all players that provide any service which enable the cloud process. This includes; providers like Terremark, Savvis, GoGrid, Amazon, and Rackspace, etc. that provide compute, storage and network Infrastructure as a Service; PaaS and SaaS providers; and managed service providers that offer backup, disaster recovery, and monitoring etc.

Consumers are all those that consume services at any level in the cloud process. This category is dominated by small to medium enterprises looking to replace their data centers with cloud services, and large enterprises wanting to use virtualization technology to share resources within their data centers. At the same time, individuals are also starting to consume cloud services particularly for software and storage. Pfizer, GE, and MorganStanley are some examples of companies heavily using the cloud for their computing needs, and some others like NASA have even setup their own cloud called Nebula.

Brokers, on the other hand, help connect consumers with providers as well as improve the cloud process as much as possible. They are the equivalent of third party logistics (3PL) providers in the traditional supply chain industry. Gravitant, Jamcracker, and Appirio are examples of cloud brokers.

In summary, the operations decisions in the cloud space can be divided based on the perspective of each player in the Cloud IT Supply Chain. All the players mentioned above are real companies that have already adopted cloud computing into their business in some form or another, so the optimization problems identified in the following sections are real problems they face and good solutions would create true business value.

4. Provider Perspective

This section discusses the decision space for cloud providers. The perspective from the provider changes slightly depending on whether the cloud is public or private. We focus on public cloud providers, and include a discussion on private cloud decisions at the end of this section.

Provider costs are primarily tied to their assets and the maintenance of these assets. For example, compute providers have a large number of expensive chassis which hold many servers, known as blades. These chassis are housed in data centers that need to be powered and cooled. Similarly, storage providers have storage arrays containing storage disks, and these arrays are connected to chassis which are all housed in large data centers. So, major provider costs can be categorized as follows (Greenberg et al 2009):

1. Servers cost (compute, storage, software)
2. Infrastructure cost (power distribution and cooling)
3. Power draw cost (electrical utility costs)
4. Network cost (links, transit, equipment)

A number of other costs exist, but this is what most cloud data centers spend money on. To keep things in perspective, suppose we have a cloud data center with 50,000 servers; the cost of servers is approximately $52.5 million per year (assuming $3000 per server, 5% cost of money, and 3 year amortization); the infrastructure costs approximately $18.4 million per year (assuming $200M in infrastructure amortized over 15 years); the cost of power for such a data center would be approximately $9.3 million per year (for a price of $0.07 per KWH and each server drawing 180W); and network costs of about $5 per Mbps per month. All these estimates are based on Greenberg et al (2009).

In addition to cost, another major issue is security, specifically with respect to intrusion detection and intrusion prevention. Customers are always worried about intrusions into their applications and data in the cloud. While some may rightly argue that these customers do not have anything to prevent similar intrusions in their current data centers anyway, providers still need to show additional intrusion detection and prevention capabilities to sustain and increase their business.

A further important issue is the environmental impact of a provider's business operations. Energy consumption related to cloud computing is a significant part of the national energy consumption. The cloud computing paradigm facilitates a general push for Green IT, for instance Google presently claims to have no carbon footprint from its data centers. In the decade to come, "green" considerations are likely to play a role in all major decisions facing cloud providers.

Optimization is of utmost importance for providers to offer competitive prices to prospective customers. Table 2 is a list of key optimization problems to be solved by the providers.

Table 2. List of Provider Optimization Problems

Models	Decision Variables	Objective	Constraints
Data Center Location Planning	• Geographical locations • Physical sizes	*Minimize* • Infrastructure investment • Operations cost • Non-renewable energy consumption	*Subject to* • Service levels
Data Center Capacity Planning	• Compute chassis requirements • Storage array requirements • Bandwidth requirements	*Minimize* • Infrastructure investment • Operations cost	*Subject to* • Strategic demand • Service level agreements
Data Center Layout Planning	• Hardware layout • Power and cooling vent layout • Space requirements	*Minimize* • Energy and cooling cost • Greenhouse gas emissions	*Subject to* • Tactical demand • Security restrictions
Data Center Scheduling	• Hardware on/off schedule • Power on/off schedule	*Minimize* • Energy and cooling cost • Greenhouse gas emissions	*Subject to* • Tactical demand variability
Hardware Load Balancing	• VM-Hardware assignment	*Maximize* • Hardware utilization	*Subject to* • Tactical demand variability • Load balancing rules
Partner Selection	• Primary partner • Secondary partner	*Minimize* • Partnership cost	*Subject to* • QoS required • Outage risk
Intrusion Detection / Prevention	• Firewall locations	*Minimize* • Evasion probability	*Subject to* • QoS required • Security budget
Product/Services Pricing	• Price per processor unit • Price per memory unit • Price per storage unit • Price per bandwidth unit	*Maximize* • Profit	*Subject to* • Customer satisfaction • Competitor pricing

* Note that the optimization problems described above are relevant for all providers, regardless if they provide IaaS, PaaS, or SaaS.

Data Center Location Planning: Providers need to decide where to house their data centers given their global demand and projected energy profile in each geographical location. This could be modeled as a facility location problem with service level constraints. The data centers would be analogous to the "facilities", the customer locations play the same role in both cases, and the latency can be related to the "distance to the facility". Moreover, if a consumer decides to put their main VMs in one location and have a backup in a different location, then the latency between data centers comes into play as well. This is analogous to lateral shipments between facilities. For a further discussion, see (Greenberg et al 2009).

An increasingly important factor in today's provider environment is how "green" the data center is. For instance, data centers can use a considerable percentage of energy from renewables (wind, solar, etc.) or they can use natural resources to assist in cooling (sea water, wind). Going green has become an important part of the optimization objective for providers to assert leadership and as a marketing tool, and there is always a trade-off between the extent of going green and the energy consumption expenditure.

This desire to increase the amount of renewables in a provider's energy consumption is also present in existing data centers. Since the big players have started to do much of their power management in-house to gain as many energy savings as possible, the resulting centralized control brings opportunities for optimization.

Data Center Capacity Planning: Providers need to purchase hardware resources for their data centers, and these decisions are typically made once a year based on expected long term demand. The main resources to be purchased are compute chassis, storage arrays, and bandwidth cables. Compute chassis are containers that hold many servers (blades). Similarly, storage arrays are containers that hold many storage disks (which could be flash, fiber, or SATA). Compute chassis and storage arrays come with pre-installed network fabric for communication between blades and storage disks, but providers still need to purchase fiber optic cables for inter-cloud communication.

An important aspect of capacity planning is to select the number of resources required and a timeline for purchasing them. Resource costs are in the range of millions of dollars and take 1-2 months for setup, so the timing of procurement is critical to satisfy demand on time while not exceeding the procurement budget. Providers setting up cloud for the first time depend on demand forecasts obtained from marketing and sales to setup a procurement plan. Providers who already have a cloud can use time series techniques or machine learning tools along with input from marketing and sales to setup a procurement plan for the next 6-12 months. Depending on the demand profile of the customers and the size of the provider, it may or may not be possible to get a high-fidelity estimate of the demand after this initial period. If this is possible, then the stochastic nature of demand can be aggregated and simplified so the procurement plan can be adjusted using deterministic mathematical programming with time-varying demand and service level constraints. If not, then the procurement plan may need to change over time as the confidence level in the demand forecast increases. This leads to a multi-period decision capacity planning problem under stochastic demand, somewhat akin to the newsboy problem. Note that cloud resources are reusable in the sense that they become available again as soon as requests have been fulfilled. This problem is reminiscent of workforce management (Mojcilović and Connors 2010, Levi and Radovanović 2010), but a difference is that project admission control plays a much smaller role than capacity selection in cloud computing.

A further problem in data center capacity planning is to select the types of resources to purchase. Typically, less reliable resources are cheaper than more reliable resources, and this choice impacts the maintenance policy and cost for each resource purchased. Most providers currently make this decision by choice, with large providers generally choosing the most reliable resources.

Another relevant research topic is the relation between resource capacity and Quality of Service (QoS). Resource capacity impacts performance in a nonlinear manner. Since random variations in demand causes QoS constraints not to be met, stochastic models and queueing theory are suitable tools for investigating this relation, see for instance (Chen and Yao 2001) for general tools. Work in this direction has already begun (Dieker et al 2012, Ghosh et al 2011, Zheng et al 2011, Tsitsiklis and Xu 2012). A particular challenge that arises in the setting of cloud computing is that the number of network components is very large, which renders classical models computationally infeasible. However, scaling methods and asymptotic theory are suitable tools for these cloud models.

Data Center Layout Planning: Given that power and cooling is a major contributor to provider cost as well as greenhouse gas emissions, the hardware needs to be placed in such a way that parts of the data center can be powered down while other sections remain powered up. This needs to correspond with the power cables layout as well as the cooling vents layout. While it is preferable to minimize the total area of the data center, security restrictions might prevent chassis and storage arrays from being stored very close to each other. Flexibility of the design is also important in view of uncertain scope and usage pattern of future additions to the data center. One method is to partition the data warehouse into sectors, each of which can be optimized in some way, and to start using more sectors as more hardware is needed. There are many ways to partition and/or optimize, e.g., based on geographical location or variability profile of the demand. We believe that mathematical programming is well-suited for problems such as these.

These problems may seem closely related to traditional warehousing design (e.g., Bartholdi and Hackman 2011), but they are fundamentally different in nature. Although uncertainty in demand plays a major role in both settings, there is no consideration of travel time and restocking within a data center. Compute chassis and storage arrays are pre-wired internally, and cables for wiring between these units are currently cheap. Moreover, the communication bottleneck lies outside of the data center.

Data Center Scheduling: The provider can establish a schedule for turning the hardware on or off to minimize power and cooling cost as well as greenhouse gas emissions, leading to a fundamental tradeoff between performance and energy usage. If the hardware is clustered by usage pattern, then entire portions of the data center can be powered down. In many cases customer demand is highly variable and patterns may emerge over time. These trends can be used to generate hardware on/off schedules using mathematical programming techniques. In other cases, it is important to take randomness into account. Work in this direction has already begun, see for instance (Al-Daoud et al 2012, Gandhi et al 2010, Lefevre et al 2010, Lin et al 2011, Mateus and Gautam 2011) and references therein.

Hardware Load Balancing: Load balancing is the problem of assigning virtual machines to servers, which happens at run-time. As demand changes, virtual machines need to be reassigned so that servers do not build up a large backlog. One possibility is to use constraint programming for these problems (Bin et al 2011). Stochastic models are suitable tools as well; since resource usage on a server is unpredictable, the utilization of servers is varying randomly over time. Much existing scheduling theory focuses on algorithms which establish system stability (e.g., Dai and Lin 2005, Dieker and Shin 2012, Maguluri et al 2012), without considering delays (response times). However, response times and load balancing are of great practical importance in the cloud.

Partner Selection: Many customers require that their data and virtual machine images be replicated across locations and across providers. Therefore, providers need to identify partner providers to offload demand in case of natural disasters or crises. In fact, partnering with other providers may also provide benefit because one provider can refer to the other for those services that they cannot themselves provide. This is analogous to partner alliances in the airline industry.

Related problems arise within a data center, where service failure can be mitigated using backups of critical information. It is thus of interest to study the relationship between backup mechanisms and failure probabilities (Undheim et al 2011, Vishwanath and Nagappan 2010).

Intrusion Detection / Prevention: With a large number of servers and continuously changing technology, data centers could be vulnerable to hackers and other threats. Therefore, providers need to minimize the probability of hacker evasion by optimally placing firewalls and other intrusion detection devices in all the key areas. While this is mainly a software issue, the provider still needs to spend resources deploying these devices and constantly testing for intrusions. This problem can be modeled as a network interdiction model, see (Morton et al 2005) for a similar application in the nuclear space.

Product/Services Pricing: Cloud services can be priced based on; component usage such as processor GHz/hr, memory GB/hr, or storage GB/month, virtual cpu (VCPU) usage such as VM hours or VCPU hours, or based on custom packages such as cloud compute units (CCUs) or elastic compute units (ECUs). Providers need to identify the best pricing model that suits their customer base, and then set the prices such that they maximize profit while maintaining customer satisfaction, taking into account competitor pricing as well. Pricing questions such as these are well-studied in various industries, such as for airline products, hospitality products (hotels), and car rentals (Talluri and Van Ryzin 2005, Phillips 2005). Other applications of revenue management and pricing include self-storage, apartment/office space leasing, and air cargo. For cloud computing, as in traditional revenue-management areas, capacity overbooking is a critical aspect of exploiting variability in demand (Meng et al 2010).

There are several critical differences between these traditional revenue management industries and pricing for cloud computing. First, the hotel, airline, and rental car industry can operate at full capacity without significant quality of service implications, whereas in cloud computing there is a trade-off between service level (performance) and capacity utilization. Second, capacity units in these traditional industries are well defined (seat, room, car) and typically discrete. Cloud computing capacity units differ and could be discrete (such as number of virtual machines) or continuous (such as virtual machine hours). Third, these industries only have a "reserved price" whereas cloud computing has "reserved prices" as well as "overage prices" also known as bursting prices. Fourth, in these industries price is a one-time charge, and it depends on expected demand and remaining availability at the consumption date (arrival date, departure date, rental date). Cloud computing prices are charged every month, so it depends on the expected demand over many months of customer engagement.

Private Cloud Decisions

We next discuss in what sense private cloud decisions are different from the public cloud decisions, using the above public cloud decision problems as a basis for comparison.

- Data Center Location Planning – Primarily of interest to large private clouds.
- Data Center Capacity Planning – Private clouds would be more concerned than public clouds, because they may not have as much capital for equipment purchasing.
- Data Center Layout Planning – Private clouds would be less concerned than public clouds, because they typically have a small number of chassis while public clouds typically have hundreds of chassis in the data center.
- Data Center Scheduling – Private clouds would be less concerned than public clouds, for the same reason as in the preceding point.
- Hardware Load Balancing – Both public and private clouds could benefit from solving this problem, because this is where "sharing" helps reducing cost.
- Partner Selection – Primarily of interest to public clouds.
- Intrusion Detection / Prevention – This is more of a problem for public clouds than private clouds because private clouds are housed within their own premises and are not accessible through public IP addresses.
- Product/Services Pricing – Primarily of interest to public clouds.

5. Consumer Perspective

This section discusses the decision space from the cloud consumer's perspective. While providers have rapidly implemented virtualization and cloud computing into their business, consumer adoption has not been as fast. This is largely due to confusion and misinterpretation of what cloud computing is and what can be expected from it. Some consumers underestimate the cloud due to security and cost issues and other consumers overestimate the cloud due to ease of use.

As first steps, consumers need to check if the cloud is right for them, order the right set of cloud resources and then properly manage the cloud resources once provisioned. These problems have been defined from an operations research point of view in Table 3. Failure to solve these problems would result in high operating costs for consumers due to VM sprawl (large number of VMs not de-provisioned after use). The abbreviation VDC stands for Virtual Data Center, i.e., the collection of virtual machines at the disposal of the customer.

Table 3. List of Consumer Optimization Problems

Models	Decision Variables	Objective	Constraints
Cloud Feasibility and Benefit Analysis	• Migrate or not • Organizational structure	*Maximize* • Scalability	*Subject to* • Budget
VDC Capacity Planning	• VM requirements • VM configuration • Storage requirements	*Minimize* • Infrastructure subscription cost	*Subject to* • Tactical demand
VDC Scheduling	• VDC active/inactive schedule • VM active/inactive schedule	*Minimize* • Infrastructure subscription cost	*Subject to* • Operational demand variability
VM Load Balancing	• Application-VM assignment • Load balancing rules	*Maximize* • VM utilization	*Subject to* • Operational demand variability
Product/Services Package Selection	• Allocated capacity • Reserved capacity • On-demand capacity	*Minimize* • Infrastructure subscription cost	*Subject to* • Tactical demand • Operational demand variability • QoS required

* While the optimization problems described above are all relevant for IaaS consumers, PaaS and SaaS consumers would be primarily focused on the Product/Services Package Selection problem.

Cloud Feasibility and Benefit Analysis: Not all applications are a good fit in the cloud. When a consumer is considering migrating an application to the cloud, they need to first check if the platform and operating system used by their application is available on the cloud. Then they need to test dependencies between this application and others to determine if it can be migrated independently with minimal impact to the other applications or if all the dependent applications need to be migrated as well. Even if the application is feasible on the cloud, if demand is very stable and there is not much need for scaling capacity on-demand then there may not be much benefit in migrating to the cloud. Therefore, consumers need to optimize the decision of whether to migrate the application or not. This decision also impacts the organizational structure and budget structure of the consumer (because cloud based applications are operational cost intensive while traditional data centers mainly involve capital expenditures).

VDC Capacity Planning: Once a consumer decides to migrate an application to the cloud, they need to know how many VMs and how much cloud storage to order. This is the problem of translating physical capacity into cloud capacity. Another problem is the arrangement of this virtual capacity into multiple VDC containers. Since each VDC is tied to a specific provider, there is a cost for each VDC. The tradeoff here is between placing all resources within a single VDC (to minimize VDC cost) and splitting resources across VDCs for added redundancy. Having multiple VDCs helps when one provider's cloud goes down and demand can be satisfied by another VDC which is handled by a different provider.

VDC Scheduling: In the same way that providers try to reduce power and cooling costs by turning hardware off whenever possible, consumers also can reduce operations costs by de-activating VDCs. These schedules can be optimized for cost based on demand patterns. Another option for consumers is to setup high demand VDCs in hot sites (fully active all the time), move failover demand VDCs to warm sites (in sleep mode but can be activated within a few hours) and arrange disaster recovery VDCs at cold sites (where hardware images can be instantiated within a day). In addition to scheduling VDCs, VM activation schedules should be optimized as well, especially in the case where the consumer is charged based on VM usage by the hour.

VM Load Balancing: When the consumer has multiple identical VMs in a cluster, then the VM activation schedule also needs to include the number of parallel VMs activated when turned on. This is governed by load balancing between the parallel VMs. Just like VM load is balanced across servers (as discussed in the previous section), application load can be balanced across VMs as well. Round robin and least utilization are examples of load balancing rules.

When demand is low, some VMs could be deactivated and its demand rerouted to the active VMs which will increase VM usage and decrease VM cost. Similarly, when demand is high, an extra VM could be activated and some portion of demand from all the other VMs can be rerouted to this newly activated VM. This routing and rerouting is possible due to virtualization technology. In this way, application demand is assigned to VMs at run-time based on demand variation to maximize VM usage, and over time this can be used to derive load balancing rules. Even if the load balancing rule is fixed, there are parameters that could be optimized. For example, in a utilization based load balancing rule, low utilization threshold and high utilization threshold values need to be identified.

Product/Services Package Selection: Providers have many different packages, and consumers have to select the package that best suits their needs. Some packages have low rates for reserved capacity but very high rates for on-demand capacity, whereas other packages have higher rates for reserved capacity with lower on-demand rates relatively. Consumers expecting large amounts of demand variation may choose the package with higher reserved capacity rates to avoid paying very high amounts for on-demand capacity (in the other packages). Even within a package, consumers need to identify the optimal level of capacity to be reserved. This is a stochastic problem driven by demand variability over time.

These are the major optimization problems faced by the consumer in the planning stages of cloud adoption as well as when the consumer is managing and governing the cloud, and some work has begun in this space (Iyoob et al 2011a). Since the cloud is meant to be a subscription based model, all capacity and scheduling decisions can be revised every month for continuous recalibration.

6. Broker Perspective

This section discusses the decision space from the perspective of cloud brokers. Some (prospective) cloud consumers choose to outsource researching the intricate differences between cloud providers, their offerings and prices. Cloud brokers offer their expertise on these differences as a service. In most cases, these third-party companies work on behalf of the consumer and get paid by the consumers, so their objective is to guide the consumers through the cloud adoption process and beyond.

It is critical for cloud brokers to solve the problems discussed in this section, because they are the drivers of cloud adoption and this drives their business. On the one hand they solve planning and migration problems for consumers and help them get on the cloud with the right providers, and on the other hand they help providers dispose of their excess capacity in the best way possible. Therefore, cloud brokers play an integral role in the future of cloud computing.

A few key problems to be solved by brokers can be seen in Table 4.

Table 4. List of Broker Optimization Problems

Models	Decision Variables	Objective	Constraints
Provider Matching and Selection (for consumers)	• Primary providers • Secondary providers	*Minimize* • Infrastructure subscription cost	*Subject to* • QoS required • Outage risk
Resource Migration Planning (for consumers)	• Legacy resources salvaged • Cloud resources provisioned with different providers	*Minimize* • Infrastructure subscription cost	*Subject to* • Tactical demand
Transformation Scheduling (for consumers)	• Transformation schedule	*Minimize* • Completion time	*Subject to* • Budget • Operations disruption
Capacity Reverse Auctions Optimization (for providers)	• Consumer for each VM auctioned • Consumer for each TB of storage auctioned	*Maximize* • Revenue	*Subject to* • Consumer performance

* The optimization problems described above are all relevant for IaaS, PaaS, and SaaS brokers.

Provider Matching and Selection: The most common problem faced by brokers is matching providers to fit consumer requirements. Not one provider will match all the requirements, so a set of providers needs to be selected for the consumer. Then another set of providers needs to be selected as backup. The tradeoff here is between match index and cost. The broker can either select a small number of expensive providers with comprehensive offerings (higher match index) each or a larger set of providers with more specific offerings (lower match index) each. The match index can be modeled qualitatively through customer experience parameters (Klancnik et al 2010) or by a quantitative approach using provider features and functions that match customer needs (Iyoob et al 2011b).

Resource Migration Planning: The typical consumer has considerable investment tied to their current IT infrastructure. So, they need assistance in planning how much of their legacy resources to keep and how much to discard (and replace with cloud resources). Brokers can solve this problem by optimizing the tradeoff between operations cost and agility for consumers. As more legacy resources are replaced with cloud resources, agility increases but operations cost increases as well (since the legacy resources are already paid for, while cloud resources are charged for on a monthly basis).

Transformation Scheduling: In addition to planning the resource migration, consumers also need to make organizational changes for the transformation to be successful. Brokers can perform an assessment of the consumer and then solve this transformation problem using project planning models. Each activity for cloud transformation has workforce resource and skill requirements, and the activities need to be scheduled so that the time to complete the migration is minimized without exceeding time-dependent budget constraints and without disrupting current operations. The schedule also needs to incorporate dependencies between activities.

There are two reasons why cloud transformation scheduling is different from classical scheduling theory as discussed in for instance (Pinedo 2008). First, multiple activities can be performed in parallel, in which case they would take longer to be completed using the same workforce; this feature also appears in other settings (Mojcilović and Connors 2010). Second, resource requirements for activities can be a function of the schedule. Some activities may require fewer resources when automation from previously completed activities can be exploited. An example is automatic data collection from one activity that makes monitoring and intrusion detection in another activity faster and easier to carry out.

Capacity Reverse Auctions Optimization: While brokers help consumers to select providers in the *provider matching and selection* problem described above, they can also help providers select consumers. When a provider has excess capacity, it is sold in spot markets such as spotcloud.com. However, brokers can setup auctions where providers can auction this excess capacity to consumers (Bapna et al 2011). In this way, providers can get a higher salvage value for their excess capacity. This is known as a reverse auction and can be solved as a bid evaluation optimization problem, see for instance (Kwasnica et al 2005) for background. This problem is also relevant for the supply chain industry and is being solved by third-party logistics providers. Game theoretic techniques seem particularly suitable to shed light on this problem.

7. Prospects

Having described a multitude of decision problems facing today's cloud computing industry, it is the aim of this section to provide an outlook for future research in this area. We also list the areas where quite a bit of research is being done and those which have not received much attention, as identified in the preceding sections.

All decision problems described in this paper are practically relevant to various parties in the cloud computing industry, but some problems intrinsically have less of a qualitative component or seem relatively straightforward applications of existing theory. Some of these problems could be viewed as "analytics" rather than OR, as they require data-driven decision making. Others are likely to give rise to new OR methodology and a few questions have already witnessed some initial progress. All problems need to be solved regardless of whether a consumer adopts IaaS, PaaS, or SaaS, but the perspective with which the problem is solved may be different. This could mean that different OR tools need to be employed.

We note that quite a bit of research is being done in the areas of: data center capacity planning (provider perspective), data center scheduling (provider perspective), and provider matching and migration planning (broker perspective). On the other hand, other areas have not yet received much attention: pricing (provider perspective), intrusion detection and prevention (provider perspective), and capacity reverse auctions (broker perspective).

Our list of cloud computing decision problems may not be all inclusive, and new technology is sure to give rise to further interesting decision-making questions. Still, we advocate for a broader participation by OR researchers in this interdisciplinary research field of increasing importance, and we believe that there are opportunities to study and develop tools from across the full methodological spectrum of Operations Research.

References

1. Al-Daoud, H., Al-Azzoni, I., Down, D. (2012). "Power-aware Linear Programming Based Scheduling for Heterogeneous Server Clusters", *Future Generation Computer Systems*, 28, pp. 745–754.
2. Armbrust, M. Fox, A., Friffith, R., Joseph, A. D., Katz, R., Konwinskii, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M. (2010). "A View of Cloud Computing", *Communications of the ACM*, 53(4).
3. Bapna, R., Das, S., Day, R., Garfinkel, R., Stallaert, J. (2011). "A Clock-and-Offer Auction Market for Grid Resources When Bidders Face Stochastic computational Needs", *INFORMS Journal on Computing*, 23(4), pp. 630–647.

4. Bartholdi, J. J., Hackman, S. T. (2011). "*Warehouse and Distribution Science"*, Version 0.95.
5. Bin, E., Biran, O., Boni, O., Hadad, E., Kolodner, E. K., Moatti, Y., Lorenz, D. H. (2011). "Guaranteeing High Availability Goals for Virtual Machine Placement", *Proc. ICDCS conference*, pp. 700-709.
6. Chen, H., Yao, D. D. (2001). "Fundamentals of Queueing Networks", Springer.
7. Dai, J., Lin. W (2005). "Maximum Pressure Policies in Stochastic Processing Networks" *Operations Research*, 53(2), pp. 197–218.
8. Dieker, A. B., Ghosh, S., Squillante, M. S. (2012). "Optimal resource capacity management for stochastic networks", *preprint.*
9. Dieker, A. B., Shin, J. (2012). "From Local to Global Stability in Stochastic Processing Networks through Quadratic Lyapunov Functions", to appear in *Mathematics of Operations Research*.
10. Foster, I., Zhao, Y., Raicu, I., Lu, S. (2008). "Cloud Computing and Grid Computing 360-Degree Compared", *Proc. Grid Computing Environments Workshop*, pp. 1–10.
11. Gandhi, A., Gupta, V., Harchol-Balter, M., Kozuch, M. (2010). "Optimality Analysis of Energy-performance Trade-off for Server Farm Management", *Performance Evaluation*, 67(11), pp. 1155–1171.
12. Gentry, C. (2010). "Computing Arbitrary Functions of Encrypted Data", *Communications of the ACM*, 53(3).
13. Ghosh, R., Naik V., Trivedi, K. (2011). "Power-Performance Trade-offs in IaaS Cloud: A Scalable Analytic Approach", *Proc. IEEE DSN-W conference*, pp. 152–157.
14. Greenberg, A., Hamilton, J., Maltz, D.A., Patel, P. (2009). "The Cost of a Cloud: Research Problems in Data Center Networks", *ACM SIGCOMM Computer Communications Review*.
15. Iyoob, I., Zarifoglu, E. (2011a). "Optimal Capacity Management for IaaS Consumers", *working paper*.
16. Iyoob, I., Zarifoglu, E., Modh, M., Farooq, M. (2011b). "Optimally Sourcing Services in Hybrid Cloud Environments", *US Patent Filed 13/373,162*.
17. Klancnik, T., Blazic, B.J. (2010). "Context-Aware Information Broker for Cloud Computing", *International Review on Computer and Software*, 5(1).
18. Kwasnica, A. M., Ledyard, J. O., Porter, D., DeMartini, C. (2005). "A New and Improved Design for Multiobject Iterative Auctions", *Management Science*, 51(3), pp. 419–434.
19. Lefevre, L., Orgerie, A.C. (2010). "Designing and Evaluating an Energy Efficient Cloud", *Journal of Supercomputing*, 51.
20. Levi, R., Radovanović, A. (2010). "Provably Near-Optimal LP-Based Policies for Revenue Management in Systems with Reusable Resources", *Operations Research*, 58(2), pp. 503–507.
21. Lin, M., Wierman, A., Andrew, L., Thereska, E. (2011). "Online Dynamic Capacity Provisioning in Data Centers", *Proc. Allerton Conference on Communication*, *Control, and Computing.*
22. Maguluri, S. T., Srikant, R., Ying, L. (2012). "Stochastic Models of Load Balancing and Scheduling in Cloud Computing Clusters", *Proc. IEEE INFOCOM conference,* pp. 702–710.
23. Mateus, C. R., Gautam, N. (2011). "Efficient Control of an $M/M_t/k_t$ Queue with Application to Energy Management in Data Centers", *preprint.*
24. Mell, P., Grance., T. (2011). "The NIST Definition of Cloud Computing", *Special publication 800-145*, National Institute of Standards and Technology.
25. Menascé, D. A, Ngo, P. (2009). "Understanding Cloud Computing: Experimentation and Capacity Planning", *Proc. Computer Measurement Group conference.*
26. Meng, X., Isci, C., Kephart, J., Zhang, L., Bouillet, E., Pendarakis, D. (2010). "Efficient Resource Provisioning in Compute Clouds via VM Multiplexing", *Proc. ICAC conference*, pp. 11–20.
27. Mojcilović, A., Connors, D. (2010). "Workforce Analytics for the Services Economy" in: Maglio, P. et al (eds), *Handbook of service science*, Springer.
28. Morton, D.P., Pan, F., Saeger, K.J. (2005). "Models for Nuclear Smuggling Interdiction", *IIE Transactions*, 38(1), pp. 3–14.
29. Phillips, R. (2005). "Pricing and Revenue Optimization", Stanford University Press.

30. Pinedo, M. (2008). “Scheduling: Theory, Algorithms, and Systems”, Springer.
31. Talluri, K., and Van Ryzin, G. (2005). “The Theory and Practice of Revenue Management”, Springer.
32. Tsitsiklis, J. N., Xu, K. (2012). “On the Power of (even a little) Resource Pooling”, to appear in *Stochastic Systems*.
33. Undheim, A., Chilwan, A., Heegaard, P. (2011). “Differentiated Availability in Cloud Computing SLAs”, *Proc. GRID conference*, pp. 129–136.
34. Vishwanath, K., and Nagappan, N. (2010). “Characterizing Cloud Computing Hardware Reliability”, *Proc. ACM SoCC conference*, pp. 193-204.
35. Zheng, Y., Shroff, N., Sinha, P. (2011). “Design of a Power Efficient Cloud Computing Environment: Heavy Traffic Limits and QoS”, *preprint*.