Uncertainty-aware Decisions in Cloud Computing: Foundations and Future Directions

H. M. DIPU KABIR and ABBAS KHOSRAVI, Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Australia
SUBROTA K. MONDAL, Faculty of Information Technology, Macau University of Science and Technology, Macao, China
MUSTANEER RAHMAN, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia
SAEID NAHAVANDI, Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Australia
RAJKUMAR BUYYA, Cloud Computing and Distributed Systems (CLOUDS) Lab, School of Computing and Information Systems, The University of Melbourne, Australia

The rapid growth of the cloud industry has increased challenges in the proper governance of the cloud infrastructure. Many intelligent systems have been developing, considering uncertainties in the cloud. Intelligent approaches with the consideration of uncertainties bring optimal management with higher profitability. Uncertainties of different levels and different types exist in various domains of cloud computing. This survey aims to discuss all types of uncertainties and their effect on different components of cloud computing. The article first presents the concept of uncertainty and its quantification. A vast number of uncertain events influence the cloud, as it is connected with the entire world through the internet. Five major uncertain parameters are identified, which are directly affected by numerous uncertain events and affect the performance of the cloud. Notable events affecting major uncertain parameters are also described. Besides, we present notable uncertainty-aware research works in cloud computing. A hype curve on uncertainty-aware approaches in the cloud is also presented to visualize current conditions and future possibilities. We expect the inauguration of numerous uncertainty-aware intelligent systems in cloud management over time. This article may provide a deeper understanding of managing cloud resources with uncertainties efficiently to future cloud researchers.

CCS Concepts: • **Computer systems organization → Dependable and fault-tolerant systems and networks**; • **Computing methodologies → Uncertainty quantification**

Additional Key Words and Phrases: Cloud computing, uncertainty, cloud traffic, QoS, cloud reliability

Authors' addresses: H. M. Dipu Kabir, A. Khosravi, and S. Nahavandi, Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Australia; emails: {hussain.kabir, abbas.khosravi, saeid.nahavandi}@deakin.edu.au; S. K. Mondal, Faculty of Information Technology, Macau University of Science and Technology, Macao, China; email: skmondal@must.edu.mo; M. Rahman, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia; email: mustaneer.rahman@gmail.com; R. Buyya, Cloud Computing and Distributed Systems (CLOUDS) Lab, School of Computing and Information Systems, The University of Melbourne, Australia; email: rbuyya@unimelb.edu.au.

0360-0300/2021/05-ART74 $15.00
https://doi.org/10.1145/3447583

ACM Reference format:
H. M. Dipu Kabir, Abbas Khosravi, Subrota K. Mondal, Mustaneer Rahman, Saeid Nahavandi, and Rajkumar Buyya. 2021. Uncertainty-aware Decisions in Cloud Computing: Foundations and Future Directions. *ACM Comput. Surv.* 54, 4, Article 74 (May 2021), 30 pages.
https://doi.org/10.1145/3447583

1 INTRODUCTION

Cloud computing is gaining popularity through its eye-catching features, such as scalability, elasticity, and pay-as-you-go [1]. Increased popularity results in an enormous growth in the cloud computing industry [2, 3]. Consequently, increased cloud traffic and resources introduce higher uncertainties. Therefore, the consideration of uncertainty is becoming important to manage cloud resources efficiently and to keep a profit margin [4]. The uncertainty exists almost everywhere in the cloud [5]. There are uncertainties in the price, availability, saving progress, and computation time of cloud **virtual machines (VMs)**, network traffics, and so on.

1.1 Background

The uncertainty is everywhere. In a production-based industry, uncertainties exist on cost, quality, production time, transportation, demand, competitors, price, and lifetime [6–8]. Uncertainties also exist in natural events influencing our daily life, such as temperature, raining, sunshine, and so on [9]. Cloud computing itself is one uncertainty-aware approach. Individual users need computations at a varying rate. The demand becomes very high for a small time and the demand becomes very low for the rest of the time. Buying a desktop of average configuration results in a longer execution time during execution and unused capacity rest of the time. The concept of cloud computing allows a user to use a large number of computational resources for short time at a reasonably low cost. Therefore, various uncertainties on the user, computation jobs, traffic, and so on, become a huge concern for cloud providers. Moreover, the user also requires information on the uncertainty of the provider for the efficient completion of the job. There are uncertainties associated with cloud users, traffics, and providers. Numerous approaches have been developing to handle the uncertainties in the cloud. The cloud user and brokers need to know uncertainties associated with different providers. A provider may provide better instances at a cheaper price but another provider may have higher consistency. The provider needs to provide computing resources in a highly fluctuating environment. A prediction for workload containing multiple uncertainty bounds is useful for them to know the exact uncertain condition of the upcoming number of users. Moreover, a proper prediction on their willingness to pay and the nature of tasks help providers in proper pricing and configuring. The knowledge in the uncertainty is useful in both of the safety and the profitable management [10–12]. Therefore, this article discusses the concept of uncertainty, current trends in cloud, and the future of uncertainty in cloud computing.

1.2 Prediction and Uncertainty

Researchers develop models for predicting quantities. The traditional point of view considers the error value as the quality of the prediction model. The error value can be the **root-mean-square-error (RMSE)** or the **mean-square-error (MSE)** in regression problems, the percentage of wrong predictions in classification problems, and so on. People develop a better model to reduce the error value of the model [13]. With the advancement in modeling, researchers reach saturation in terms of error reduction. No matter how well the model training is, there is a certain error probability. Such as, researchers are developing neural networks for handwritten digit recognition. Researchers have concluded that an excellent CNN can provide about 99.7% accuracy. Researchers

have found 99.84% state-of-the-art performance with excellent CNN and extremely lucky training session. The rest 0.16% portion is the uncertainty of the system. Some digits are confusing to both humans and machines [14, 15]. However, we can not assume that the uncertainty is uniform; some handwritten digits are easy to detect, and some handwritten digits have high uncertainty. Therefore, a heteroscedastic uncertainty quantification system is required [16]. Cloud computing is a fast-growing field. Various components of clouds have different levels of uncertainties. Researchers may soon conclude that there exist inherent randomness of the system betterment of model cannot reduce that error probability [17], and they need uncertainty modeling. This article may provide an overview of cloud uncertainty to cloud researchers.

1.3 Related Work

There exist several short-length survey papers on the uncertainty in cloud computing [18–21]. Some research works in several sub-domains of cloud computing also mention about uncertainty on cloud [22–24]. However, there is no paper providing a detailed discussion on uncertainty in cloud computing. Therefore, we write this survey to provide a detailed discussion on uncertainty in cloud computing.

We focus on both human involved decisions and intelligent systems for the uncertainty-aware cloud management. The uncertainty is traditionally quantified as the interval forecast or the error probability in various economic and industrial problems. Engineers and statisticians take decisions based on quantified uncertainties. In cloud management, some uncertainty-aware decisions need to be taken within a very short time, such as checkpointing [25]. The human involvement is inefficient, as it takes a longer time. Therefore, many intelligent approaches have been developing over time to handle uncertainties in the cloud. Some other management allows a higher time for taking a decision, such as capacity management. Automated approaches usually handle situations with a single uncertainty bound corresponding to certain error probability, and humans usually follow the quantified uncertainty for the future planning.

1.4 Our Contributions

- We discuss approaches of humans for managing uncertainties.
- We propose major uncertain parameters in cloud computing. We discuss how numerous uncertain factors may affect major uncertain parameters and how major uncertain parameters affect cloud performance.
- We propose a comprehensive survey of uncertainty-aware decisions in cloud computing.
- We propose a hype curve of uncertainty in the cloud and discuss the future of uncertainty-aware approaches.

1.5 Article Structure

The rest of the article is organized as follows: Section 2 presents the uncertainty—the concept of quantification, managing uncertainties, and the uncertainty in the cloud. Section 3 presents major uncertain parameters and effects of various activities on them. Section 4 presents the effect of major uncertain parameters in the cloud QoS. Section 5 presents existing uncertainty-aware systems in the cloud. Section 6 presents the current condition and future directions with the help of the Hype Curve. Section 7 summarizes and concludes the article.

2 THE UNCERTAINTY

The uncertainty is simply known as the lack of certainty. More or less, we have uncertainties associated with predictions almost everywhere. Examples are chances of rain, equipment failures, and flight delays.

2.1 Types of Uncertainties

Uncertainty categorization depends on the perspective. Based on the level of uncertainty, the uncertainty can be classified into the following categories [26]:

- Ignorance
- Severe uncertainty
- Mild uncertainty
- Certainty

Name of each category also presents the definition. Ignorance is such a situation when the person has no idea about the outcome. For example, one user wants to see a number less than four in single dice rolling but he does not know how many sides the dice have. In ignorance, the person has no idea about important determinants of the outcome. There exists a high risk with severe uncertainty. An individual wants to see heads while tossing one coin. The person knows the probability of success but that probability is not too high (>>0.5) or not too low (<<0.5). Therefore, the person cannot say clearly whether the event will occur or not. The example of the mild uncertainty can be rolling one 16-sided dice once and observing a number lower than 16. The certainty can be exemplified as the expectancy of both heads or tails while tossing. An outcome with a very slight success or unsuccess probability can also be treated as the certainty. Such as expecting at least one head in the toss of 100 coins. Uncertainty can also be presented by numeric limits, popularly known as the interval forecast. With historical data and several related parameters, one may predict the temperature of a certain location to be between $20^oC–22^oC$ for the upcoming hour. That event has mild uncertainty. However, one-day ahead forecast with insufficient information may result in a wider interval ($10^oC–30^oC$), possessing a severe uncertainty.

Statisticians and mathematicians also try to reduce the level of uncertainty through modeling improvement. In this context, the uncertainty can be classified as follows [27]:

- Aleatory uncertainty
- Epistemic uncertainty

The aleatory uncertainty is also known as the inherent randomness. The consequence of the same action with the same circumstances can be different due to the aleatory uncertainty. A signal can vary largely from its common historical pattern due to the aleatoric uncertainty. Day-to-day temperature curve may vary largely on a day due to an unpredictable event caused by the aleatoric uncertainty. The epistemic uncertainty is known as the modeling error. The epistemic uncertainty occurs when secondary or tertiary effects are overlooked during the modeling. The model designer can reduce the epistemic uncertainty through improving the modeling process.

2.2 Uncertainty and Risk

Many people consider the uncertainty quantification as the risk analysis. However, risk analysis is a special case of uncertainty quantification. The risk is a type of uncertainty where some possible outcomes a significant loss [29]. Uncertainty in the production time may indicate some risk due to a nearby deadline. An increase in production time usually causes a linear loss, increasing effective running costs. A substantial loss can occur when the production misses a major event of the shipment date.

2.3 Uncertainty Quantification (UQ)

The probability density function can present the exact uncertain condition. However, the probability density or the cumulative probability density cannot be expressed with a few words or numbers. Therefore, the uncertainty is usually quantified as the interval forecast [30]. The

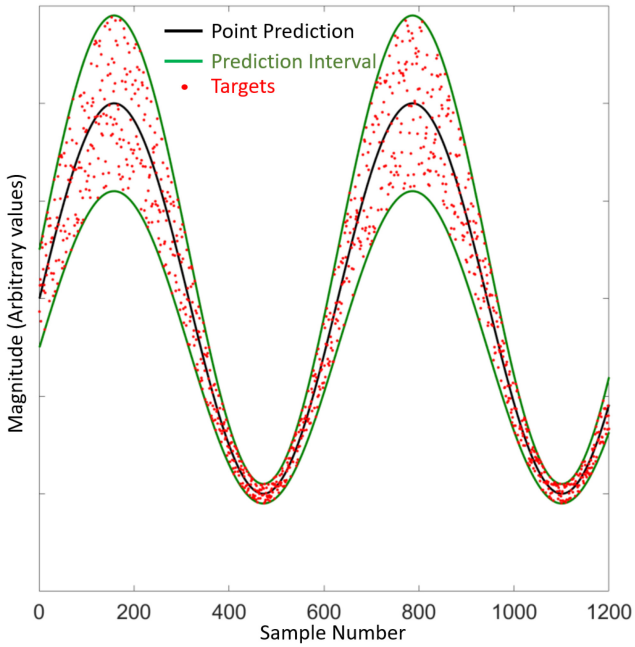

Fig. 1. A rough sketch presenting the importance of uncertainty quantification. Red dots present targets and the black solid line presents the point prediction. The point prediction is a value corresponding to the mean or median of the probability distribution. It does not convey any message about the uncertainty. The uncertainty is low near sample 450 and the uncertainty is high near sample 800. The prediction interval with the green line is representing uncertainty. The interval is sharp for a low uncertainty and the interval is wide for a high uncertainty.

interval forecast is presented by three numbers: the upper bound, the lower bound, and the coverage probability. The upper bound and the lower bound are the predicted upper limit and predicted lower limit of the future quantity, respectively. The coverage probability is the probability that the target will be within the upper and the lower bounds. To quantify uncertainties, researchers propose prediction intervals of different coverage probability and probabilistic forecasts [31, 32]. Figure 1 presents a rough sketch showing the importance of uncertainty. The point prediction provides a numeric value that is derived from the minimum statistical error. The point prediction conveys no evidence of the heteroscedastic uncertainty. An unavoidable aleatoric heteroscedastic uncertainty may cause different deviation from targets at different positions. A heteroscedastic prediction interval can represent that uncertain condition.

The interval-based uncertainty quantification may seem inefficient for a single event. However, the interval provides a smart indication of the uncertainty while numerous factors are responsible for the uncertainty. For example, while rolling one six-sided dice the probability of getting one to six is equal. An interval of 90% coverage probability extends the entire output range. While rolling five dices and observing the sum of outcomes, the probability of getting a number between 12 to 23 is 88.244% [33]. The range of output in rolling five dices is 5 to 30. Therefore, the width of an interval of 88.244% confidence is (12/26=) 46.15% of the range, as shown in Figure 2. The interval becomes narrower compared to the range with a larger number of dice-rolling. Our real-life events are influenced by numerous probabilistic events and the effect of all probabilistic events can be predicted by a narrow interval of high coverage probability; most of the situations. The

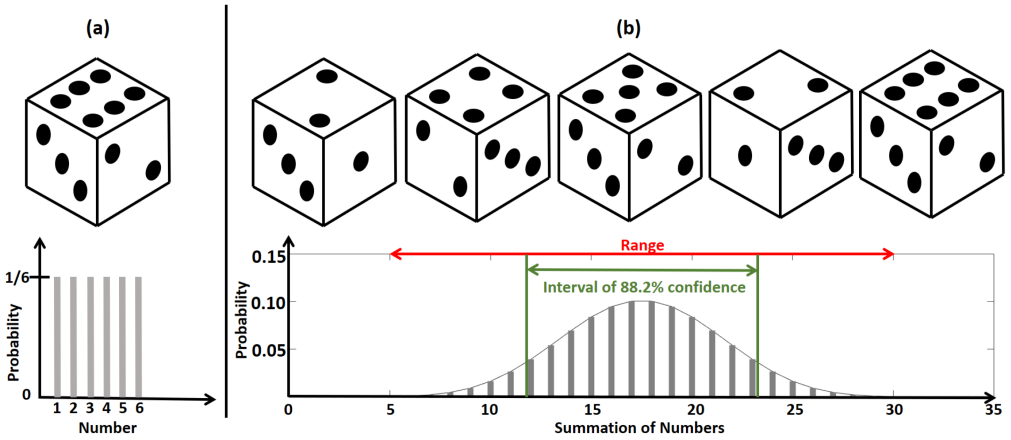

Fig. 2. The formation of high and low probable regions from events of a uniform probability distribution. Rolling a six-sided dice has an equal probability of getting a natural number from one to six. An interval of 90% confidence expands the entire range. The situation is presented in (a). Summation of numbers of rolling five dice has a non-uniform probability distribution. An interval of 88.24% coverage has 46.15% width of the range. The situation is presented in (b). The interval becomes narrower compared to the range with a higher number of dice rolling. Considering the multiplication of numbers results in even narrower intervals compared to the range [28].

statistical outcome of a large number of dice rolling is highly deterministic with a small error and the outcome becomes dependent on the property (number of sides) and the fairness of the dice. The usage of an individual cloud user is difficult to predict but the total usage of million users is predictable with higher confidence. Moreover, the total usage becomes a function of major events (vacation, sports, weather, etc.) for a large number of users.

2.4 Managing Uncertainties

Uncertainty is unavoidable for human beings [34]. People make a number of precautions to handle uncertain situations. Common precautions are as follows:

- Assume future is like the past and modeling
- Rules, norm, and conventions to eliminate some worst possibilities
- Buffers and redundancies
- Trial and error
- Routine inspection and maintenance
- Regulatory institutions
- Consideration of alternatives

Uncertainty is growing in emerging engineering and economic issues. For example, the large-scale inauguration of renewable resources in the power grid has made the power generation more unpredictable [35, 36]. Numerous prediction algorithms have been developing proposing optimal prediction systems [37, 38]. Online auction-based cloud computing services of low reliability have put users in an uncertain condition [39]. They are designing intelligent approaches to finish their computation jobs in a cost-efficient way [40]. The enormous growth of the cloud industry has increased uncertainty in cloud traffic. Increased research in designing autonomous systems has raised the importance of understanding uncertainty [41]. In all of these situations, researchers are developing models for both prediction and uncertainties.

Rules, conventions, and intelligent strategies also applied to manage uncertainties. When everyone obeys rules, fewer uncertain situations arise. Fewer accidents happen while drivers and pedestrians follow rules and conventions. Such as saving the progress of the computation job results in less damage due to the unexpected termination of the computation jobs. The creation of a historical log may help future users in understanding consequences. Keeping a good buffer is a must to overcome uncertainty. The buffer can be excess time or money or any other utility. A deadline constrained computation job can be finished economically when the deadline is much larger than the required computation time. Trial and error are required to collect initial data for the modeling when the consequence is unknown.

Routine inspection and maintenance are also prescribed to avoid unexpected situations. All vehicles need routine maintenance to check several degradations that may lead to a major malfunction. Moreover, the user needs more frequent personal inspections, such as checking oil and coolant levels. People also consider alternatives to avoid uncertainty. People often consider different shops when one shop is too busy. There are multiple counters in busy shops and banks. Upcoming customers usually stand in a shorter queue. Customers also observe the movement in different queues and switch to the queue that moves first. Through the process, most customers get the service within a reasonable time.

2.5 Uncertainties in Cloud

Cloud specialists agree that 100% reliability target is wrong for almost everything in the cloud. Notable exceptions are pacemakers and anti-lock brakes. It is possible to cover 99.999% situations with a reasonable amount of precautions but covering the rest 0.001% requires much more precautions [42]. In various fields, such as economics, energy generation, and demand predictions, people consider 95% certainty [43]. The expected uncertainty in the cloud depends on the following factors: the user's satisfaction, the uncertainty of alternatives, and the consequence of different actions.

Cloud computing provides three basic services named, **Infrastructure-as-a-Service (IaaS)**, **Platform-as-a-Service (PaaS)**, and **Software-as-a-Service (SaaS)** [44]. IaaS provides only fundamental resources, such as processing power, network, and storage. Popular examples of IaaS are **Amazon Web Services (AWS)** and **Google Compute Engine (GCE)**. PaaS provides all facilities of IaaS with operating systems and middlewares. Popular examples of PaaS are Apprenda, Pivotal CF, and Red Hat OpenShift. SaaS applications usually run directly through the web browsers and do not require any downloads or installations of software. Popular examples of SaaS are Google Apps and Dropbox.

There exists numerous work in predicting the quantities associated with the cloud computing. Although many works on predicting clouds do not consider the uncertainty, researchers know two major facts. First, the error probability in one-minute ahead prediction is much lower than the error probability in one-hour ahead prediction. Secondly, the error probability in predicting some quantities are significantly higher than some other quantities. The error probability is defined by some statistical error values between the model output and observations. Popularly applied statistical error values are the **root mean square error (RMSE)**, the **mean square error (MSE)**, the **sum squared error (SSE)**, and the **mean absolute percentage error (MAPE)**.

Tchernykh et al. agree that uncertainty is the main hassle of cloud computing and it brings challenges to brokers, resource providers, and end-users [20]. Vredeveld et al. develop a model for online scheduling with the consideration of uncertainty [45]. Mendoza et al. propose a model for VoIP cloud environment considering uncertainty [46]. Bychkov et al. consider failure probability and possible financial results [47]. Fard et al. consider the lower and upper bounds of the processing time for executing workflow applications on the cloud [48]. The work of Fabio et al.

consider service elasticity, which includes scaling of cloud computing services and overbooking [49]. Roland et al. propose a realistic cloud workflow simulation with noisy parameters [50]. Very recently Aranitasi et al. quantify uncertainties for preemptive resource provisioning in the cloud [51]. Bhargavi et al. present a novel soft-set-based optimal scheduling of cloud tasks [52] under uncertainty. According to Basset et al. [53] imprecision latent in the estimation process is one of the three major challenges in implementing cloud services in an organization. They develop the **neutrosophic multi-criteria decision analysis (NMCDA)** approach to estimate the quality of services under uncertainty. Section 5 presents a detailed survey of uncertainty-aware decisions in cloud computing.

3 MAJOR UNCERTAIN PARAMETERS IN CLOUD COMPUTING

The cloud is connected to the entire world through the Internet. Millions of parameters may slightly affect the cloud-job. Our environment is also similarly connected to the entire world. According to weather prediction specialists, flapping a butterfly in one country can be related to rain in another country [54]. A small occurrence in a country may affect a cloud user in another country greatly. Therefore, we sort out five major uncertain parameters of cloud computing those directly affect a cloud user. These parameters are price, availability, traffic, workload, and security. Numerous occurrences change these uncertain parameters. Also, these five uncertain parameters combine with the provider's attribute and determine the **quality of service (QoS)** [55].

There can be a large number of influencing factors that affect cloud computing. A larger number of users than the capacity can hamper traffic, availability. A security breach or any inconstancies in another datacenter in a different country can also affect a datacenter. Many users may switch datacenter. The effect will be on availability, traffic, and workload. Being driven by the internet, cloud computing has many influencing factors. However, all factors are directly influencing five major uncertain parameters. In Table 1, we present how common influencing factors affect major uncertain parameters. Table 2 shows that all major parameters performance parameters. Many factors are not one of the major uncertain parameters, but they often directly affect performance parameters. Such as, a provider can provide a different level of freedom to a certain user. A provider may not allow a new user to book a large number of servers. In fact, that policy affects availability. Similarly, all other policies and factors affect major uncertain parameters.

Common occurrences directly influencing five uncertain parameters are presented as Table 1. A server request may come from different locations. The location of a cloud user directly affects the network traffic. That location can also raise concerns about the security due to the probability of the leakage of cloud information. However, the location of the user does not affect the price, availability, and workload of instances directly. Data size directly affects cloud traffic. More data is transferred through the network when the data size is larger. More data does not directly affect price, availability, workload, and security. However, the traffic congestion due to the size of data may potentially cause a long time to save the progress affecting the availability and the future workload. Moreover, any unsaved data can be lost. Everything directly or indirectly influences everything in the cloud. Therefore, we mark only direct influences. The arrival of new jobs influences price, availability, traffic, and workload.

The price of spot EC2 instances instantly varies depending on the prices of bids [56]. The price of the on-demand EC2 instances also changes based on the arrival of jobs but that change is not too frequent. As a result, checkpointing increases cloud traffic and the cloud workload of corresponding machines. However, checkpointing is mandatory to prevent significant loss of data or computation progress. Communication between server also increases workload and traffic. The total number of available computing resources is also a determinant of price and availability. The network capacity influences the traffic directly. Many users parallelize their task and run them

Table 1. Factors Directly Influencing Major Uncertain Parameters in Cloud Computing

Common Influencing Factors	**Major Uncertain Parameters of Cloud**				
	Price	Availability	Traffic	Workload	Security
Location of Users	×	×	✓	×	✓
Data Size	×	×	✓	×	×
Jobs Arrival	✓	✓	✓	✓	×
Checkpointing	×	×	✓	✓	×
Communication	×	×	✓	✓	×
Capacity of Provider	✓	✓	×	×	×
Network Capacity	×	×	✓	×	×
Task Parallelization	✓	✓	✓	✓	×
Overbooking	✓	✓	×	✓	×
Task Execution Time	×	✓	×	✓	×
Execution Failure in VM	×	✓	×	✓	✓
VM Interruption	×	✓	×	✓	✓
Network Failure	×	✓	✓	✓	✓
Cyber Attacks	×	×	✓	✓	✓
Nearby Datacenters	×	✓	✓	✓	×

✓- Directly influencing; × - Not directly influencing.

Table 2. Influence of Major Uncertain Parameters to Performance Parameters of Cloud Computing

Performance Parameters [55]	**Major Uncertain Parameters Affecting**
Service Response Time	Availability, Traffic, Workload.
Sustainability	Availability, Traffic, Security.
Suitability	Price, Availability, Traffic, Workload, Security.
Accuracy	Availability, Traffic, Workload, Security.
Transparency	Availability, Security.
Interoperability	*
Availability	Availability
Reliability	Availability, Security.
Stability	Availability, Traffic, Workload.
Effective Cost	Price, Traffic, Workload.
Adaptability	Availability, Traffic, Workload.
Elasticity	Availability, Traffic, Workload.
Usability	Price, Availability, Traffic, Workload, Security.
Throughput and Efficiency	Availability, Traffic, Workload.
Scalability	Availability, Traffic, Workload.

* - Providers' Attributes.

in different servers. Task parallelization changes the workload pattern over time. It also affects availability, the price of instances, and the network traffic. Cloud service providers often overbook their resources [57]. Overbooking can affect availability, workload, and price of instances.

The task execution time affects the availability of the resource and the workload directly. Execution failure in a cloud instance may cause a loss of simulation progress. The instance may become unavailable and that may affect the availability of instances in a data center. Re-performing the

simulation results in an increased workload. Cloud service providers often interrupt the progress of low-cost preemptible instances to facilitate premium users [58, 59]. The instance interruption directly affects availability, workload, and security. Any failure at the network also changes availability, traffic, workload, and security. Cyber attacks influence traffic, workload, and security. However, influences may vary based on attacks. The attacker may try to access the same server from different locations with the help of spyware. The data is secured in such a situation but traffic and workload are affected. While the attacker steals the data, the security is breached. Nearby datacenters directly affect availability, traffic, and workload. The price is also indirectly affected.

3.1 Price of Cloud Instances

The price of cloud instances varies depending on time, location, provider, and types of instances. As cloud computing is becoming popular over time, numerous companies have begun to provide cloud services. Popular cloud providing companies are Amazon Web Services, Microsoft Azure, Google Cloud, IBM Cloud, Adobe, VMware, Rackspace, Red Hat, Salesforce, Oracle Cloud, SAP Cloud, and Dropbox. Some instances have higher flexibility and reliability than others, such as AWS on-demand instance [60] or Azure Pay-as-you-go instance [61], which are reliable and allow the user to leave the instance without any penalty. The price of such instances does not change frequently over time but varies from location to location. The user may require to wait for the instance during the provisioning or switch the region to get the instance. Reservation of instances can be up to 40% cheaper compared to highly flexible and reliable AWS on-demand or Azure Pay-as-you-go instances. The price is determined during the provisioning and there are uncertainties associated with the price. There are also low cost and highly unreliable instances, which can be 80% cheaper than highly flexible reliable instances. Four companies are currently providing such instances. Instances are AWS Spot Instance [56], Azure Low Priority VM [62], Google Preemptible VM [63], and IBM Transient Virtual Servers [64]. The price of such instances is highly uncertain. Low cost and highly unreliable instances are the spare capacities after providing highly flexible and reserved instances. The price of AWS spot instances may increase or decrease anytime. The user may lose the instance due to the price hike.

3.2 Availability of Cloud Instances

The availability of a cloud service is the percentage of time a user can access the service [55]. The following equation defines the availability:

$$Availability = \frac{total\ time\ for\ which\ service\ was\ available}{total\ service\ time}. \tag{1}$$

Lu et al. incorporated uncertainty for the cloud application development decisions in 2013. They perform an availability analysis from the cloud consumer perspective [4]. To obtain uncertainties, they propose a set of availability analysis models applying **Stochastic Reward Nets (SRNs)** [65]. They also provide insights into some deployment decisions. Several reasons are affecting the availability of cloud instances. These include runtime failures [66], workload spikes [67], rare and hardly predictable events [68], interference [69], and combined effects. To address challenges such as the inherent uncertainty in the mobile cloud, Viswanathan et al. [70] proposed role-based resource provisioning framework with self-healing, self-optimization, and self-organization.

The recent popularity of low cost and highly unreliable instances has bought extensive research on their availability [71]. Azure low-priority-VMs and Google's preemptible VMs are 60% to 80% cheaper. The price of such VM does not change rapidly, but the user may lose the instance with a 30-second notice due to the shrink in spare capacity. IBM's transient virtual servers can be reclaimed without any notification.

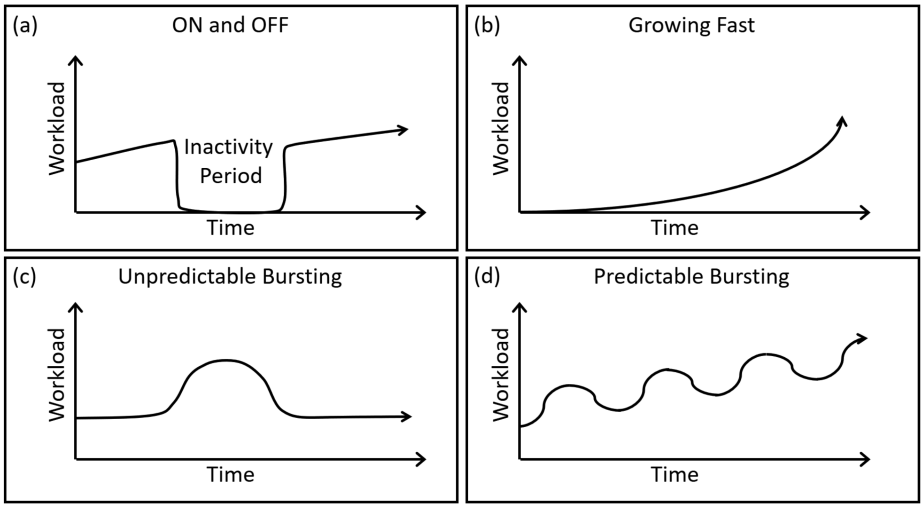

Fig. 3. Rough sketches presenting four workload patterns in cloud computing (a) ON and OFF, (b) Growing Fast, (c) Unpredictable Bursting, and (d) Predictable Bursting.

3.3 Cloud Traffic

The cloud traffic prediction is required for the proper management of computer networks [72]. Resource allocation and checkpointing may consume a much longer time than usual due to a wrongly predicted traffic [73]. Benson et al. investigate [74] data center traffic patterns. However, one research cannot cover all traffic patterns and there are always uncertainties. Users of less-reliable cloud instances may save progress during the price hike [75]. However, if all users use the same algorithm to save progress, then the network becomes busy during that time. The user may fail to save progress within the expected time due to the traffic. That may result in the loss of computed results.

Wolski et al. developed one of the very first network traffic prediction systems in 1997 [76], which aimed to predict cloud traffic with different prediction algorithms and selected one with the lowest statistical error. Later, Xinyu et al. proposed [77] error-adjusted LMS method. The prediction of cloud traffic is still challenging and predictions results in high MSE or MAPE error values [78]. Cloud users may require uncertainty quantification of the cloud traffic to select VMs of different providers and locations near future.

3.4 Cloud Workload

The word workload means the amount of work needs to be accomplished by someone or something [79]. However, different cloud researchers define cloud workloads differently. Yang et al. [80] and Liang et al. [81] define workload as the number of requests of the application. Song et al. [82] and Jiang et al. [83] define workload as the future demand of VMs. Garg et al. [84] and Jheng et al. [85] define workload as the resource utilization of VMs. Rodrigo et al. derive a model for VM provisioning under the uncertain workload [86]. Their adaptive resource provisioning model maintains the required QoS and the utilization threshold.

Steve et al. classified cloud workload patterns into four categories, as shown in Figure 3 [87]. The cloud workload increases slightly and faces inactivity periods in the ON-and-OFF pattern. A good example of such a pattern is the verification department of pharmaceutical R&Ds. They perform numerous measurement and analysis before launching any product. Once those simulations are

complete, they release instances until the next time. The exponential increase in the demand is categorized as the growing fast. That is the situation of a growing company using the cloud VMs or a growing cloud service provider. The unpredictable bursting is the third category, which happens due to a huge burst in demand for a short time. Popular examples are workloads of newspapers, social media, and search engine when anything goes viral. The demand fluctuates periodically or following events in a predictable bursting category. The use of computational resources by offices are mostly limited by office hours and follows daily patterns and changes over holidays.

Gilles et al. propose several models for the workload prediction and compare them [88]. According to Gilles, the NN-based prediction system provides better performance, and constraint programming is better for the trace generation. Cao et al. predict cloud workload with the NN [89]. Their data is real, large-scale, and enterprise-class collected from a database-based data center.

3.5 Security

The cloud security is another broad field [90–92]. We mention common security concerns in the current work. Common security concerns include: the nefarious use of cloud computing, unauthorized access, data loss, identity hacking, leakage of user information, service interruption, and so on. Cyber-attacks include denial of service attack, service injection attack, virtualization attack, the user to root attack, port scanning, man-in-middle attack, metadata spoofing attack, phishing attack, and backdoor channel attack [93, 94].

Malicious attacks and software errors are increasingly common. Software errors are increasing due to the growth in size and complexity of software and novel applications. Malicious attacks and software errors can cause faulty nodes to exhibit Byzantine (i.e., arbitrary) behavior [95, 96] in which components of a system fail in arbitrary ways, i.e., not just by stopping or crashing but by processing requests incorrectly, corrupting their local state, and/or producing incorrect or inconsistent outputs. Consequently, it is mandatory to have **Byzantine Fault Tolerant (BFT)** mechanism to defend against Byzantine failures so a system can continue to operate accordingly even if some of its components exhibit arbitrary, possibly malicious behavior. Usually, using BFT mechanism helps ensure not to preempt each particular fault, however, the number of system components that can fail at a time is bounded. Note that BFT mechanism adopts replication technique to defend Byzantine fault as suggested by Byzantine Generals' Problem [95, 96]. In the mechanism, we would have at least $3f + 1$ replicas where f be the maximum number of replicas that may be faulty. For example, **Hadoop distributed file system (HDFS)** uses the default replication factor of 3 for enhancing/ensuring fault tolerance [97]. As we know, redundant resources incur cost even though dependability and security are enhanced. The authors in Reference [98] analyze the tradeoff of redundant resources usage in terms of unavailability metric, cost of cloud service deployment, and security of the service deployed.

A major concern in the cloud is the loss of data or computation progress. It is more frequent in **high-performance computing (HPC)** jobs in preemptible cloud instances compared to the security breach [99]. Efficient checkpointing or local storing capability can reduce the loss due to the unexpected termination of instances.

4 EFFECTS OF MAJOR UNCERTAIN PARAMETERS ON THE PERFORMANCE OF CLOUD SERVICES

Cloud performances, such as the **Quality of Service (QoS)** and **Service Level Agreements (SLAs)** can be different for different cloud specialists. Garg et al. define **key performance indexes (KPIs)** for evaluating cloud computing services [55]. That paper is highly cited and, therefore, we consider their KPIs as the performance parameter of the cloud. They mention 15 KPIs, presented

in Table 2. We find that some of the KPIs are quantitative and some of the KPIs are qualitative. Moreover, some of the KPIs are varying, and some other KPIs are constant for a provider.

4.1 Service Response Time

A user requests for a VM or a service. Based on availability, traffic, and workload, the user faces a delay [100]. Garg et al. [55] mention three response parameters as evaluation indexes for the response time: average response time, maximum response time, and response time failure. According to a recent study, the average response time of cloud services is 50.35 milliseconds [101]. The response time failure occurs when the provider fails to serve within the promised response time [102].

4.2 Sustainability

Sustainability refers to the environmental impact of cloud service. The sustainability mostly depends on the location of data centers and renewable installations near data centers. These factors depend on the provider. Several major uncertain parameters of the cloud can also affect sustainability. When servers at sustainable locations are not available, the user may launch their job in a different location. The traffic conditions may also cost packet loss during the communication, which results in higher power consumption. Security issues such as loss of computation progress or denial of service degrade sustainability. The attributes of the provider are certain and usually known to the user. Therefore, Table 2 presents only the effect of major uncertain parameters.

4.3 Suitability

A cloud provider may fail to achieve one customer's suitability while another cloud provider may succeed. A user can be dissatisfied due to price, VM configuration, latency, or anything else. All of the five uncertain parameters can directly affect the suitability. There can be a drawback of the infrastructure of the cloud provider, but the infrastructure influences major uncertain parameters, and several major uncertain parameters affect the suitability.

4.4 Accuracy

The first indicator of the accuracy is the percentage of time the provider maintains the promised SLA [103–105]. The SLA is a statement from the provider where the provider states the minimum quality of service that the provider should provide. The accuracy depends on availability, traffic, workload, and security.

4.5 Transparency

Transparency in cloud computing is the quality of the provider that allows good usability during the change of circumstances [55]. The transparency is often inferred as the time for which the performance of the application is postponed due to the change of service. Transparency is an important factor, as cloud services are changing rapidly. The statistical accuracy is the ratio between the summation of delays and the summation of such occurrences. The transparency depends on the infrastructure of the provider and the nature of change. The provider affects availability and security, and these uncertain parameters affect transparency.

4.6 Interoperability

Interoperability is defined as the ability of a cloud-service in interacting with other cloud services. Other cloud services can be from the same cloud provider or can be from a different cloud provider.

The interoperability is approximated as follows:

$$Interoperability = \frac{NPO}{NPR}, \tag{2}$$

where NPO is the number of platforms offered by the provider. NPR is the number of platforms required by users for interoperability.

The interoperability is mostly dependent on compatibility issues. Therefore, the interoperability is not time-varying unless the provider brings some upgrade. Therefore, interoperability depends only on the provider.

4.7 Availability

We consider the availability as both the major uncertain parameter and performance criteria. Many other performance criteria are directly dependent on availability. If we do not consider the availability as the major uncertain parameter, then we have to find indirect relations. Such as, a workload may vary availability that may affect the sustainability. However, everything in the cloud is indirectly related and we consider only direct relations.

4.8 Reliability

Reliability in cloud computing is defined as the expected length of uninterrupted service. Reliability in cloud service failure is expressed as follows:

$$Reliability = P_{us} \times PMTTF, \tag{3}$$

where P_{us} in the Probability of uninterrupted service. $PMTTF$ is the Promised mean time to failure.

The probability of uninterrupted service depends on the number of failures and the total number of trials (N). Therefore, the reliability is also expressed as follows:

$$Reliability = \left(1 - \frac{number\ of\ failures}{N}\right) \times PMTTF. \tag{4}$$

The reliability of a system is directly dependent on availability and security. The user may not select a server at a premium cost that has availability issues. Such as Amazon's on-demand cloud instances have higher availability, and the chance of losing availability in the middle of the job is lower compared to spot instances. Therefore, on-demand cloud instances are more reliable. Reliability also depends on security. A user may not perform an important job or put any confidential data on insecure servers.

4.9 Stability

The variation in the performance of a service determines the stability in cloud computing. It is the variance in computation time in computing and the variance in the average read-write time in storage. The stability largely depends on the availability, traffic, and workload. Security also hampers stability, but the failure of security is rare in cloud computing and has a negligible effect in variation.

4.10 Effective Cost

Cloud providers offer different instances of different attributes and attributes do not follow a linear relation. While comparing different instances, we often observe that the size of **random access memory (RAM)** of one instance is double of another but the number of cores of the processor is the same and one processor is 1.1 times faster than another processor. It is difficult to compare the price of a processor of one configuration from the price of a processor of another configuration

[106]. Garg et al. define the effective cost as follows [55]:

$$Effective\ Cost = \frac{Price}{CPU^a \times net^b \times data^c \times RAM^d}, \tag{5}$$

where a - d are weights and $a + b + c + d = 1$. As an average user cannot compare the effective cost, this equation is useful to compare servers in terms of the effective cost.

4.11 Adaptability

The user may try to upgrade the service to a higher level due to an urgent requirement or switch to a lower level instance for cost efficiency. The provider needs to provide another resource based on customers' requirements. The adaptability of the provider is the time required to switch between services [106]. Depending on the situation, the provider may require different times to switch between instances. Availability, traffic, and workload affect adaptability. When the required instance is not available, the user needs to wait. Large traffic may also increase the required time.

4.12 Elasticity

Elasticity is the scalability of the provider during a sudden demand. Two terms determine the elasticity: the maximum capacity of service and time required to expand the usage [107, 108]. Elasticity depends on availability, traffic, and workload. The provider can maintain a good elasticity when the amount of job is lower than their capacity. However, the elasticity is low when a small provider is serving a large number of jobs [109]. High traffic and workload can also potentially create congestion resulting in a lower elasticity.

4.13 Usability

Usability in cloud services is the ease of using the cloud service. It is solely dependent on the provider. The user may feel more comfortable with one providers' interface than another provider. However, the provider is a certain parameter. Different policies of providers also affect all major uncertain parameters, resulting in a significant influence on usability. The provider as well as major uncertain parameters determine the usability. As the providers' configuration is constant, all major uncertain parameters are responsible for the uncertainty in usability.

4.14 Throughput and Efficiency

The throughput of a cloud service is the number of tasks completed by that service in unit time. The throughput depends on availability, traffic, and workload. The efficiency is the effective utilization of leased resources. The efficiency depends on how instances are designed and offered by a provider and the requirement of the application. However, the provider is constant. Therefore the uncertainty on the throughput and efficiency depends on availability, traffic, and workload.

4.15 Scalability

The scalability is the ability of a cloud service to handle a large number of requests simultaneously. The scalability has two dimensions: horizontal and vertical scalability. Horizontal scaling is the initiation of more virtual machines. Vertical scaling is the increase in the resource; such as physical memory, CPU speed, or network bandwidth. The provider can limit the scalability of an individual user. A new cloud user cannot use a large number of Amazon Spot Instances. The vertical scaling is dependent on availability, traffic, and workload.

5 EXISTING UNCERTAINTY-AWARE INTELLIGENT SYSTEMS IN CLOUD

5.1 Pricing Models

The price of non-preemptible servers changes less frequently. Only four major cloud providers are providing preemptible instances. Among them, the price of Amazon **Spot Instance (SI)** changes more rapidly and it is possible to get higher reliability with the willingness to pay high. Therefore, a numerous bidding framework has been developed by researchers [110–112]. However, all factors determining the price and methods for computing the price for both preemptive and non-preemptive instances are not disclosed by providers. Several researchers develop price optimization functions based on historical data of the Amazon SI price. The optimization function consists of two major parts [113], commonly known as revenue maximization and capacity utilization [114, 115]. The revenue maximization function is the multiplication of the number of accepted SIs and the price of SI. To increase the user-friendliness of the EC2 bidding system, Amazon is also considering a utilization maximization function. Utilization optimization function increases logarithmically with the increment of the number of accepted bids. Equation (6) presents the profit function, Equation (7) presents the utilization function, and the Amazon EC2 SI provider's probable optimization function is the maximization of the sum of these functions, presented as Equation (8).

$$Profit\ Function = \pi(t)N(t), \tag{6}$$

$$Utilization\ Function = log(1 + N(t)), \tag{7}$$

$$\max_{\pi(t)} \quad \pi(t)N(t) + \beta log(1 + N(t)), \tag{8}$$

where, $N(t)$ is the number of accepted SIs, $\pi(t)$ is the price per accepted SI, and β is the weight of the utilization term. The price per accepted SIs ($\pi(t)$) is determined by numerous uncertain bid values coming from numerous users and the available SI capacity. Amazon has recently in 2018 announced that they are not considering bid prices. They are only considering the available capacity and acceptable bids [116]. However, that encourages users to bid high although they are not willing to pay a high price. Moreover, workload prediction for the pricing of SIs is important to reduce frequent allocation and termination of servers [117, 118]. From discussions on preemptible instances of popular providers [197–201], we have sketched a rough cloud providers' diagram representing the information flow in cloud management in Figure 4.

5.2 Capacity Planning to Ensure Availability

The prediction is widely applied to the capacity planning of cloud datacenters [119, 120]. However, that prediction can never be the traditional prediction that results in a value close to the mean or median of the probability density function. Amazon and Google SLAs promise 99.95% uptime [121]. They need to consider the highest probable value of the demand to ensure 100% availability. However, installing a very large number of servers to meet the highest possible demand is inefficient. That results in a large number of unused capacity and a very large installation cost [122]. Therefore, the cloud provider needs to consider the uncertainty upper bound. Such as, they need a value that is higher than 99.95% cumulative probability. The uncertainty **upper bound (UB)** can be defined as follows:

$$P\Big(Demand \leq UB^{(99.95\%)}\Big) = 99.95\%, \tag{9}$$

where, $P(Condition)$ is the probability function. Recently Amazon has announced that they are applying AI for the capacity planning [123]. We expect that other major cloud service providers are switching towards AI-based uncertainty-aware decisions in the near future.

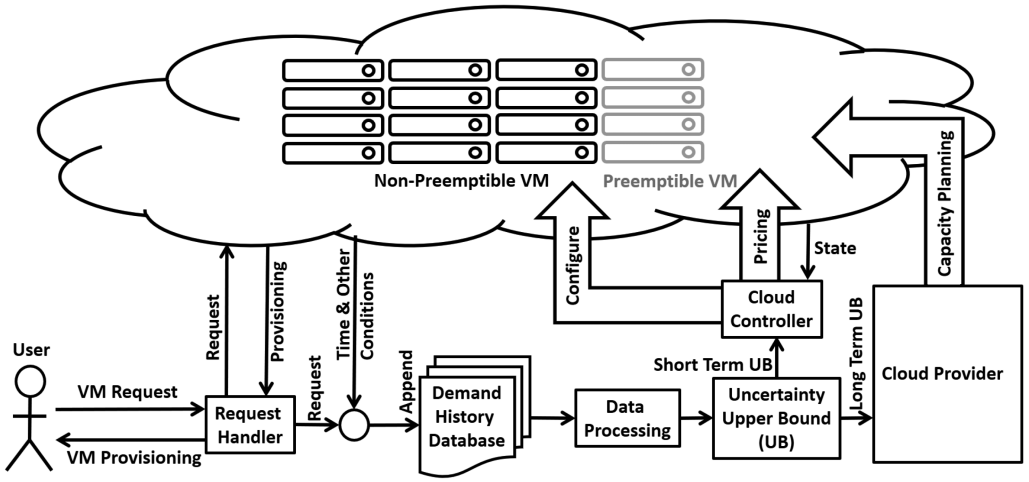

Fig. 4. A rough diagram representing the information flow in cloud management. The cloud controller requires one or more short-term uncertainty bounds for proper control. Long-term uncertainty bounds are useful in capacity planning for the future.

5.3 Traffic Management

Traffic uncertainty models are basically of two types: offline design and online routing [124]. Applegate et al. propose an offline design for network traffic management [125]. Their objective is to minimize the maximum link utilization. The objective of M. Kodialam et al. [126] offline model is to maximize the throughput. Several other uncertainty-aware offline network models aim to minimize the cost [127–130]. Several other researchers develop online routing mechanisms for network management [131, 132]. Both offline design and online routing can consider the probability distribution of incoming traffic. D. Xiao et al. [133] consider a Gaussian probability distribution of traffic. However, the probability distribution can be non-Gaussian and there are a lot of opportunities to improve.

5.4 Configuring Cloud for Workload Management

The auto-scaling of cloud resources is performed in a loop consisted of Monitoring, Analysis, Planning, and Execution steps [108]. The monitoring step monitors some performance indicators, such as: application characteristic, monitoring cost, and SLAs. The analysis step determines the volume of resource allocation for each category. This step computes workload prediction, scaling time, adaptivity to mitigation, and oscillation changes. The planning step performs resource estimation and resource combination based on the outputs of the analysis step. Finally, planned actions are executed. These steps are performed repeatedly in an interval called the monitoring interval.

M. Tajvidi et al. develop an uncertainty-aware system model for the optimized resource provisioning [134]. P. Jamshidi et al. design uncertainty-aware automatic elasticity controller based on fuzzy logic and machine learning [135]. The proper scheduling of cloud resources can reduce the cost and the peak-hour demand for cloud resources. However, the task scheduling in the cloud is challenging due to high fluctuations in workload patterns and unstable performance of the infrastructure. Marco et al. develop a model for efficient cloud-resource provisioning and scheduling [136]. They consider the minimization of the overall monetary cost. Their execution time of an application does not exceed the specified deadline with a given probability even in presence of high uncertainties.

5.5 Security Enhancement Techniques in Cloud Computing

5.5.1 Confidentiality and Authentication. The data confidentiality in cloud computing is the quality of the provider of protecting data from illegal or unwanted access [137–139]. Confidentiality is becoming a challenging issue with the added features of cloud computing platforms. Computational resources in cloud computing are often shared with a group of people [140]. The owner may add or remove users over time. Confidentiality is not only limited to data security but also relies on several circumstances and providers' policies. When the owner removes access of certain people, they may lose the document [141], or they may have the permission of seeing the older version of the document. In some situations, the owner may not have enough rights to remove access from one collaborator. Although network security is one of the oldest problems of the internet, still, researchers are facing new problems, developing algorithms, and simulating before applying those algorithms [142, 143].

5.5.2 Secured Domain and Data Encryption. Many datasets contain confidential industrial, political, or health-related information. It is also discouraged to keep confidential and sensitive datasets on the cloud. The provider needs to provide a secure domain for confidential datasets to attract users. Providing a secured domain and data encryption is a quite saturated field [144, 145]. However, maintaining a good tradeoff between high security and swiftness of the domain is challenging. Researchers are developing new algorithms and strategies to solve novel problems [146–148].

5.5.3 Reducing the Number of Dropped or Lost Jobs. Traditional dependability metrics such as availability, reliability, continuity, and maintainability are defined from a system-oriented perspective and may not adequately capture the dependability experience from a user's perspective [149]. Bauer et al. [150] state that it is better to focus on the much smaller number of unreliable service events or service defects, since most of the components in a modern cloud computing system is reliable. These service defects are conveniently normalized as the number of customer demands or user requests or jobs not served or dropped or lost, per million attempts—referred to as **defects per million (DPM)** [150–153]. The authors in Reference [153] analytically show the effectiveness of the checkpointing and replication scheme to the minimization of DPM, i.e., to optimize the uncertainty of cloud services.

5.6 Uncertainty from the Perspective of Users

Several companies are providing cloud services of different cost and different QoS. A user requires performance indication in the uncertain cloud environment to choose a proper resource. Moreover, the user of preemptible instances needs to consider the heteroscedastic reliability to save the computation progress efficiently [154, 155].

5.6.1 Server Selection. Zheng et al. predict the QoS ranking for the selection of cloud server [156]. They consider past user experiences in terms of response time and throughput. Instead of considering the average value, they consider minimum, maximum, mean, and the standard deviation. These parameters indicate uncertainties associated with the corresponding cloud server. Rehman et al. applied **Multiple Criteria Decision Making (MCDM)** to historical QoS data and importance weights from users to rank servers [157]. Qian et al. present a system cloud service selection for IaaS platforms considering usage, performance, and geographical location [158]. The server selection on Amazon SI's also depends on the price. Sabyasachi et al. quantify uncertainties associated with the SI price [40]. They consider the condition of user and price for efficient bidding. Benouaret et al. are using statistical human decisions for the selection [159]. Human brains can understand more quality parameters than machines. Such as the background color of a website may

affect its popularity, and human voting can indicate the overall popularity where mathematical equations fail.

5.6.2 Checkpointing. The cloud is widely used for the progress monitoring of different tasks. Therefore, the approach of monitoring a cloud computing job and saving the progress is popularly known as the checkpointing or application checkpointing. The checkpointing consists of three steps: pausing the computation, saving the current state, and resuming the computation. Checkpointing can be periodic or based on uncertainties. Yi et al. [160] develop a checkpointing scheme based on the uncertainty on the spot instance price. A checkpointing is performed with the price increment. Sui et al. [161] propose a learning-based adaptive checkpointing strategy. Different computation nodes may have different reliability. Therefore, different checkpointing frequencies can be applied to different servers [162].

5.6.3 Keeping Margin. There are time-gaps between buying and getting instances. Although the gap is mostly limited by the SLA agreement, SLA violation may happen anytime. The user needs some time margin while applying for cloud instances or migrating cloud instances [160, 163]. The user of preemptive instances needs to have one stable machine to withstand the preemption [40]. If there are insufficient backup servers, then he may lose the progress. The performance of cloud servers can be slightly lower than a physical server of the same capability. The difference occurs due to communication and monitoring. The user needs to select a proper instance or instance groups to finish the task before the deadline.

5.7 Artificial Neural Network for Cloud Management

The value of many real-world quantities depends on a large number of factors. The service response time, for example, depends on Availability, Traffic, and Workload. These major uncertain parameters depend on many influencing factors. It is laborious and time-consuming to derive mathematical equations between all factors and the quantity [164–166]. Moreover, the pattern of cloud data changes over time. Therefore, many researchers train Neural Networks with an error-optimization method to find relations between influencing factors and the quantity [167–169]. Several researchers have proposed AI-based solutions for cloud management to achieve better QoS with optimal power consumption and overall cost [170, 171].

The relations among AI, cloud, and uncertainties are deeply rooted. The cloud provides highly scalable computing resources for AI training. Recently, AI is an efficient means of cloud management. AI can compute the level of heteroscedastic uncertainty [165, 172], and uncertainty exists with outputs of NNs. Several works apply AI to improve the performance of fog and cloud systems [173–176].

6 FUTURE DIRECTIONS

6.1 Improved UQ

The quantification of uncertainty is still a debating issue. The probability density expresses the exact uncertain condition. However, there is no suitable algorithm to calculate the heteroscedastic probability distribution. Prediction Interval is a popular approach to express uncertainty [177, 178]. The width of the interval represents the level of uncertainty. However, the interval does not represent the skewness of the distribution. Multiple probabilistic forecasts can represent the shape of distribution [31]. There is no widely accepted formula or NN-training approach for both prediction interval and probabilistic forecast. Researchers may find improved approaches for quantifying uncertainties near future.

6.2 UQ in Emerging Cloud Fields

Cloud computing is still expanding. Many services will be added to the cloud, especially in the **Software as a Service (SaaS)** [179, 180]. Providers will face more difficulties in maintaining the availability of different SaaS instances with cost efficiency. Moreover, there is ongoing research on efficient hardware architectures for different software applications and tasks. Some applications such as neural network training can be performed efficiently with different architectures [181, 182]. Currently, Xilinx and AWS are providing FPGAs as cloud instances. Gradually, all major cloud providers will provide FPGAs and other customized instances for special applications. Improved computation will increase the volume of traffic and traffic uncertainty. The expansion of the fog computing and the edge computing will also increase the uncertainty in traffic.

6.3 Improved Artificial Neural Network Training and GPU Processing

Several major cloud providers have switched to **Artificial Neural Network (ANN)**-based cloud management [123]. ANN is trained with an initial dataset. The ANN performs poorly with a different set of data. As cloud uncertainties and pattern of uncertainties are changing over time, ANNs need to retrain with newer sets of data over a certain interval. Moreover, researchers may apply novel learning approaches for quantifying uncertainties of cloud computing. For example, adversarial training for cloud management or dropout for predicting cloud uncertainties [183, 184]. Hardware accelerators, such as the **graphics processing unit (GPU)** and the **tensor processing unit (TPU)** are becoming popular for remarkably faster processing [185]. Researchers may apply these hardware accelerators to train uncertainty quantification NNs.

6.4 Hype Cycle for Uncertainty in Cloud

The Hype Cycle is the representation of the maturity of applications or approaches of a specific domain [186, 187]. Therefore, we construct a hype cycle to visualize conditions of uncertainty-aware approaches in cloud computing. Figure 5 presents our proposed Hype Cycle for uncertainty in the cloud.

UQ for cloud management, AI for uncertainty in the cloud, and availability analysis are in the innovation trigger region in the hype curve. There are a few works on the UQ of cloud parameters and the term is unfamiliar to the majority in the cloud community. Besides a few recent discussions and announcements from several IT companies [188], very few scholarly works have published on AI-based cloud management. There will be a lot of work on AI for cloud management in coming years. Moreover, training approaches and training cost-functions will also evolve. Therefore, it will take more than 10 years for AI for uncertainty in the cloud to reach the plateau of productivity. The availability analysis of preemptible instances is another rapidly growing approach. More cloud providers will start to provide preemptible instances over time, and the growing popularity will increase the uncertainty in the availability.

QoS ranking and UQ in cloud applications are at the peak of inflated expectations. As several providers are offering similar cloud servers, a QoS rank prediction is becoming crucial for proper selection of servers [53, 156]. Numerous approaches have been developing for the selection of a better server from possible combinations [159, 189–192]. Therefore, there is a growing debate on the relative effectiveness of different QoS rankings. There will be a significant improvement in the selection process in the upcoming years. Users may find a few widely accepted server selection techniques within a few years. After achieving widely accepted techniques, the uncertainty-based server selection will go to the next stage of the Hype Curve. UQ in cloud application has reached its peak of inflated expectations. Researchers are quantifying uncertainties for various quantities [193]. However, current UQ approaches have limitations. These limitations may cause trough of

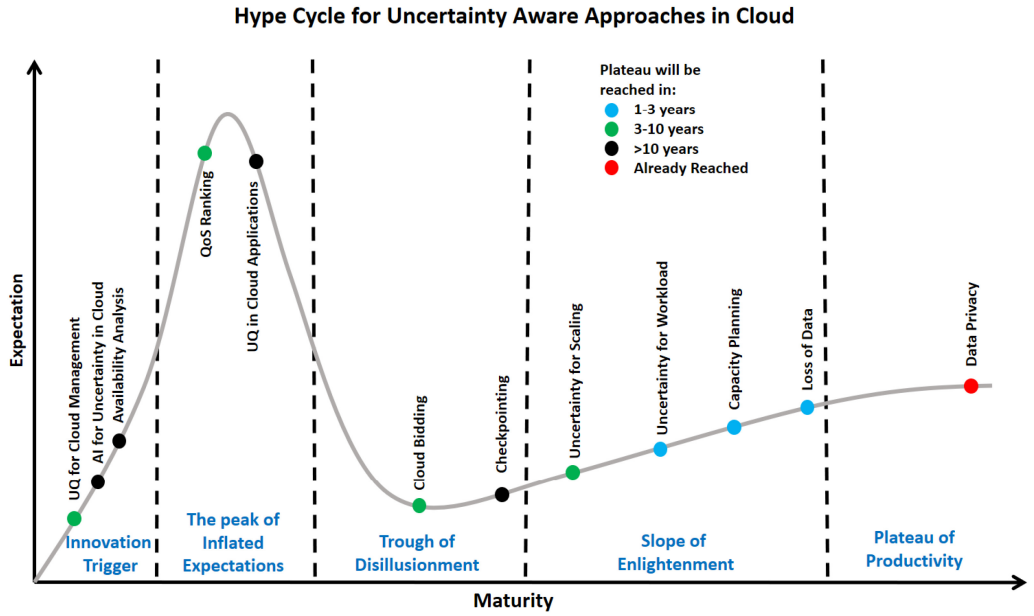

Fig. 5. The proposed Hype Cycle for uncertainty in cloud.

disillusionment in the upcoming years. New applications of UQ will reveal over time. The algorithm improvement will also improve the expectation. It will take more than 10 years for UQ in cloud applications to reach the plateau of productivity [194, 195].

Cloud bidding and checkpointing are at the trough of disillusionment step of the Hype Curve. AWS has recently changed its pricing model [116]. AWS is not considering the profit maximization from available bids at this moment. AWS spot prices are more predictable now. Therefore, less research is ongoing on the development of an efficient bidding system. However, uncertainty in the SI price will increase over time with the increased number of users. Moreover, AWS may change their rules anytime to increase profit and to punish careless bidders. Cloud servers are becoming more reliable over time, and that is decreasing the expectation from new checkpointing approaches. However, many other factors can cause preemption or shutdown of instances and the expectation on checkpointing will increase over time.

Uncertainty-aware decisions for workload, capacity planning, and loss of data are at the slope of enlightenment region and will reach the plateau of productivity within a few years. Major cloud providers are applying machine learning for efficient management of cloud resource. Workload and capacity planning research will come to its maturity within a few years. They are also offering backup memories with the hibernation state to prevent data loss [196]. Resource scaling is also at the slope of enlightenment region of the Hype Curve. Provision of different kinds of resources will increase challenges in resource scaling. It will require 3 to 10 years for the uncertainty-aware resource scaling to reach its plateau of productivity.

Data privacy is a quite saturated field. There are highly secured algorithms and still, many occurrences are happening on data-stealing or account hijacking. The expectation from the data privacy is flat at this moment, and therefore it has reached its plateau of productivity.

7 SUMMARY AND CONCLUSIONS

The recent enormous growth of the cloud industry has increased the uncertainty in cloud computing. Numerous uncertainty-aware systems have been developing over the years to meet the

demand. With the traditional approaches to design uncertainty-aware systems, researchers have also started to quantify uncertainties. Although the uncertainty is not efficiently quantifiable for a single event, the effect of a large number of events is quantifiable.

The cloud is connected to the entire world through the Internet. A small influence from one country can affect a user in another country slightly. Therefore, we select five major uncertain parameters: price, availability, traffic, workload, and security. Major uncertain parameters are directly affected by numerous influencing factors. Change in major uncertain parameters affects the QoS.

This work also describes popular uncertainty-aware approaches. A rough hype curve is also drawn to present the situation. The hype curve presents the maturity region and time to reach the plateau for each approach. Therefore, the situation of a trend becomes easily understandable. Dependency on the cloud will increase over time. Moreover, numerous automation approaches will require uncertainty quantification in the near future. This work will direct future researchers and developers of cloud computing to design uncertainty-aware efficient cloud management systems.

REFERENCES

[1] Rajkumar Buyya, Satish Narayana Srirama, Giuliano Casale, Rodrigo Calheiros, Yogesh Simmhan, Blesson Varghese, Erol Gelenbe, Bahman Javadi, Luis Miguel Vaquero, Marco A. S. Netto et al. 2018. A manifesto for future generation cloud computing: Research directions for the next decade. *ACM Comput. Surv.* 51, 5 (2018), 1–38.

[2] Blesson Varghese and Rajkumar Buyya. 2018. Next generation cloud computing: New trends and research directions. *Fut. Gen. Comput. Syst.* 79 (2018), 849–861.

[3] Sukhpal Singh Gill and Rajkumar Buyya. 2018. A taxonomy and future directions for sustainable cloud computing: 360 degree view. *ACM Comput. Surv.* 51, 5 (2018), 1–33.

[4] Qinghua Lu, Xiwei Xu, Liming Zhu, Len Bass, Zhanwen Li, Sherif Sakr, Paul L. Bannerman, and Anna Liu. 2013. Incorporating uncertainty into in-cloud application deployment decisions for availability. In *Proceedings of the IEEE 6th International Conference on Cloud Computing (CLOUD'13)*. IEEE, 454–461.

[5] Rodrigo N. Calheiros, Enayat Masoumi, Rajiv Ranjan, and Rajkumar Buyya. 2014. Workload prediction using ARIMA model and its impact on cloud applications' QoS. *IEEE Trans. Cloud Comput.* 3, 4 (2014), 449–458.

[6] Songpu Ai, Antorweep Chakravorty, and Chunming Rong. 2019. Household power demand prediction using evolutionary ensemble neural network pool with multiple network structures. *Sensors* 19, 3 (2019), 721.

[7] Mohammad Reza Chalak Qazani, Houshyar Asadi, and Saeid Nahavandi. 2019. High-fidelity hexarot simulation-based motion platform using fuzzy incremental controller and model predictive control-based motion cueing algorithm. *IEEE Syst. J.* 14, 4 (2019), 5073–5083.

[8] Seyed Mohammad Jafar Jalali, Parham M. Kebria, Abbas Khosravi, Khaled Saleh, Darius Nahavandi, and Saeid Nahavandi. 2019. Optimal autonomous driving through deep imitation learning and neuroevolution. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC'19)*. IEEE, 1215–1220.

[9] Anastasia Zabaniotou. 2020. A systemic approach to resilience and ecological sustainability during the COVID-19 pandemic: Human, societal, and ecological health as a system-wide emergent property in the Anthropocene. *Global Transit.* 2 (2020), 116–126.

[10] H. M. Dipu Kabir, Abbas Khosravi, M. Anwar Hosen, Saeid Nahavandi, and Rajkumar Buyya. 2019. Probability density for amazon spot instance price. In *Proceedings of the IEEE International Conference on Industrial Technology (ICIT'19)*. IEEE, 887–892.

[11] Éloi Bossé and Basel Solaiman. 2018. Fusion of information and analytics: A discussion on potential methods to cope with uncertainty in complex environments (big data and IoT). *Int. J. Dig. Sig. Smart Syst.* 2, 4 (2018), 279–316.

[12] In Lee and Kyoochun Lee. 2015. The Internet of Things (IoT): Applications, investments, and challenges for enterprises. *Bus. Horiz.* 58, 4 (2015), 431–440.

[13] Shakti Goel and Rahul Bajpai. 2020. Impact of uncertainty in the input variables and model parameters on predictions of a long short term memory (LSTM) based sales forecasting model. *Mach. Learn. Knowl. Extract.* 2, 3 (2020), 256–270.

[14] Siham Tabik, Ricardo F. Alvear-Sandoval, María M. Ruiz, José-Luis Sancho-Gómez, Aníbal R. Figueiras-Vidal, and Francisco Herrera. 2020. MNIST-NET10: A heterogeneous deep networks fusion based on the degree of certainty to reach 0.1% error rate. ensembles overview and proposal. *Inf. Fusion* 62 (2020), 73–80. DOI: https://doi.org/10.1016/j.inffus.2020.04.002

[15] H. M. Kabir, Moloud Abdar, Seyed Mohammad Jafar Jalali, Abbas Khosravi, Amir F. Atiya, Saeid Nahavandi, and Dipti Srinivasan. 2020. SpinalNet: Deep neural network with gradual input. *arXiv preprint arXiv:2007.03347* (2020).

[16] Konstantin Posch and Juergen Pilz. 2020. Correlated parameters to accurately measure uncertainty in deep neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 3 (2020), 1037–1051.
[17] Rodrigo N. Calheiros, Rajiv Ranjan, Anton Beloglazov, César A. F. De Rose, and Rajkumar Buyya. 2011. CloudSim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Softw.: Pract. Exper.* 41, 1 (2011), 23–50.
[18] Sanjeevi Pandiyan and Viswanathan Perumal. 2017. A survey on various problems and techniques for optimizing energy efficiency in cloud architecture. *Walailak J. Sci. Technol.* 14, 10 (2017), 749–758.
[19] Manuel Trenz, Jan C. Huntgeburth, and Daniel J. Veit. 2013. The role of uncertainty in cloud computing continuance: Antecedents, mitigators, and consequences. *ECIS 2013 - Proceedings of the 21st European Conference on Information Systems*.
[20] Andrei Tchernykh, Uwe Schwiegelsohn, Vassil Alexandrov, and El-ghazali Talbi. 2015. Towards understanding uncertainty in cloud computing resource provisioning. *Procedia Comput. Sci.* 51 (2015), 1772–1781.
[21] Andrei Tchernykh, Uwe Schwiegelsohn, El-ghazali Talbi, and Mikhail Babenko. 2019. Towards understanding uncertainty in cloud computing with risks of confidentiality, integrity, and availability. *J. Comput. Sci.* 36 (2019), 100581.
[22] Sukhpal Singh and Inderveer Chana. 2016. A survey on resource scheduling in cloud computing: Issues and challenges. *J. Grid Comput.* 14, 2 (2016), 217–264.
[23] Yijie Wang, Xiaoyong Li, Xiaoling Li, and Yuan Wang. 2013. A survey of queries over uncertain data. *Knowl. Inf. Syst.* 37, 3 (2013), 485–530.
[24] Sunilkumar S. Manvi and Gopal Krishna Shyam. 2014. Resource management for Infrastructure as a Service (IaaS) in cloud computing: A survey. *J. Netw. Comput. Applic.* 41 (2014), 424–440.
[25] Muhammad Shafie Abd Latiff et al. 2017. A checkpointed league championship algorithm-based cloud scheduling scheme with secure fault tolerance responsiveness. *Appl. Soft Comput.* 61 (2017), 670–680.
[26] Richard Bradley and Mareile Drechsler. 2014. Types of Uncertainty. *Erkenntnis* 79, 6 (2014), 1225–1248.
[27] H. M. Dipu Kabir, Abbas Khosravi, Mohammad Anwar Hosen, and Saeid Nahavandi. 2018. Neural network-based uncertainty quantification: A survey of methodologies and applications. *IEEE Access* 6 (2018), 36218–36234. DOI: 10.1109/ACCESS.2018.2836917
[28] H. M. Dipu Kabir, Abbas Khosravi, Saeid Nahavandi, and Abdollah Kavousi-Fard. 2019. Partial adversarial training for neural network-based uncertainty quantification. *IEEE Trans. Emerg. Topics Comput. Intell.* (2019). 1–12. DOI: 10.1109/TETCI.2019.2936546
[29] David Hillson. 2002. Extending the risk process to manage opportunities. *International Journal of Project Management* 20, 3 (2002), 235–240.
[30] Abbas Khosravi, Saeid Nahavandi, and Doug Creighton. 2013. Quantifying uncertainties of neural network-based electricity price forecasts. *Appl. Energy* 112 (2013), 120–129.
[31] Wenjie Zhang, Hao Quan, and Dipti Srinivasan. 2018. An improved quantile regression neural network for probabilistic load forecasting. *IEEE Trans. Smart Grid* 10, 4 (2018), 4425–4434.
[32] Mehdi Rafiei, Taher Niknam, and Mohammad-Hassan Khooban. 2017. Probabilistic forecasting of hourly electricity price by generalization of ELM for usage in improved wavelet neural network. *IEEE Trans. Industr. Inform.* 13, 1 (2017), 71–79.
[33] Eric W. Weisstein. 2006. Dice. https://mathworld.wolfram.com/.
[34] Lex Hoogduin. 2018. Decision Making in a Complex and Uncertain World. Retrieved from https://www.futurelearn.com/courses/complexity-and-uncertainty.
[35] Maria De Giorgi, Paolo Congedo, and Maria Malvoni. 2014. Photovoltaic power forecasting using statistical methods: Impact of weather data. *IET Sci., Meas. Technol.* 8, 3 (2014), 90–97.
[36] Maria Malvoni, Maria Grazia De Giorgi, and Paolo Maria Congedo. 2017. Forecasting of PV power generation using weather input data-preprocessing techniques. *Energy Proc.* 126 (2017), 651–658.
[37] Sang-Bing Tsai, Youzhi Xue, Jianyu Zhang, Quan Chen, Yubin Liu, Jie Zhou, and Weiwei Dong. 2017. Models for forecasting growth trends in renewable energy. *Renew. Sustain. Energy Rev.* 77 (2017), 1169–1178.
[38] M. Malvoni, M. C. Fiore, G. Maggiotto, L. Mancarella, R. Quarta, V. Radice, P. M. Congedo, and M. G. De Giorgi. 2016. Improvements in the predictions for the photovoltaic system performance of the Mediterranean regions. *Energy Convers. Manag.* 128 (2016), 191–202.
[39] Maria Alejandra Rodriguez and Rajkumar Buyya. 2017. A taxonomy and survey on scheduling algorithms for scientific workflows in IaaS cloud computing environments. *Concurr. Comput.: Pract. Exper.* 29, 8 (2017), e4041.
[40] Abadhan Saumya Sabyasachi, Hussain Mohammed Dipu Kabir, Ahmed Mohamed Abdelmoniem, and Subrota Kumar Mondal. 2017. A resilient auction framework for deadline-aware jobs in cloud spot market. In *Proceedings of the IEEE 36th Symposium on Reliable Distributed Systems (SRDS'17)*. IEEE, 247–249.
[41] Curtis Marshall, Blake Roberts, and Michael Grenn. 2017. Intelligent control & supervision for autonomous system resilience in uncertain worlds. In *Proceedings of the 3rd International Conference on Control, Automation and Robotics (ICCAR'17)*. IEEE, 438–443.

[42] Stephen Thorne. 2018. Tenets of SRE, available in Retrieved from https://medium.com/@jerub/tenets-of-sre-8af6238ae8a8.
[43] Farnad Nasirzadeh, H. M. Dipu Kabir, Mahmood Akbari, Abbas Khosravi, Saeid Nahavandi, and David G. Carmichael. 2020. ANN-based prediction intervals to forecast labour productivity. *Eng., Construct. Archit. Manag.* 27, 9 (2020).
[44] Haithem Mezni, Sabeur Aridhi, and Allel Hadjali. 2018. The uncertain cloud: State of the art and research challenges. *Int. J. Approx. Reas.* 103 (2018), 139–151.
[45] Tjark Vredeveld. 2012. Stochastic online scheduling. *Comput. Sci.-res. Devel.* 27, 3 (2012), 181–187.
[46] Jorge M. Cortés-Mendoza, Ana-Maria Simionovici, Pascal Bouvry, Sergio Nesmachnow, Bernabe Dorronsoro et al. 2015. VoIP service model for multi-objective scheduling in cloud infrastructure. *Int. J. Metaheur.* 4, 2 (2015), 185–203.
[47] I. Bychkov, G. Oparin, A. Tchernykh, A. Feoktistov, V. Bogdanova, Yu Dyadkin, V. Andrukhova, and O. Basharina. 2017. Toolkit for simulation modeling of logistics warehouse in distributed computing environment. In *Proceedings of the 3rd International Conference on Information Technology and Nanotechnology, Science and Engineering*. 1106–1111.
[48] Hamid Mohammadi Fard, Sasko Ristov, and Radu Prodan. 2016. Handling the uncertainty in resource performance for executing workflow applications in clouds. In *Proceedings of the IEEE/ACM 9th International Conference on Utility and Cloud Computing (UCC'16)*. IEEE, 89–98.
[49] Fabio Lopez-Pires, Benjamin Baran, Leonardo Benítez, Saul Zalimben, and Augusto Amarilla. 2018. Virtual machine placement for elastic infrastructures in overbooked cloud computing datacenters under uncertainty. *Fut. Gen. Comput. Syst.* 79 (2018), 830–848.
[50] Roland Mathá, Sasko Ristov, and Radu Prodan. 2017. Simulation of a workflow execution as a real cloud by adding noise. *Simul. Modell. Pract. Theor.* 79 (2017), 37–53.
[51] Marin Aranitasi, Benjamin Byholm, and Mats Neovius. 2017. Quantifying uncertainty for preemptive resource provisioning in the cloud. In *Proceedings of the 28th International Workshop on Database and Expert Systems Applications (DEXA'17)*. IEEE, 127–131.
[52] K. Bhargavi and B. Sathish Babu. 2017. Soft-set based DDQ scheduler for optimal task scheduling under uncertainty in the cloud. In *Proceedings of the 2nd International Conference On Emerging Computation and Information Technologies (ICECIT'17)*. IEEE, 1–6.
[53] Mohamed Abdel-Basset, Mai Mohamed, and Victor Chang. 2018. NMCDA: A framework for evaluating cloud computing services. *Fut. Gen. Comput. Syst.* 86 (2018), 12–29.
[54] Robert C. Hilborn. 2004. Sea gulls, butterflies, and grasshoppers: A brief history of the butterfly effect in nonlinear dynamics. *Amer. J. Phys.* 72, 4 (2004), 425–427.
[55] Saurabh Kumar Garg, Steve Versteeg, and Rajkumar Buyya. 2013. A framework for ranking of cloud computing services. *Fut. Gen. Comput. Syst.* 29, 4 (2013), 1012–1023.
[56] Orna Agmon Ben-Yehuda, Muli Ben-Yehuda, Assaf Schuster, and Dan Tsafrir. 2013. Deconstructing Amazon EC2 spot instance pricing. *ACM Trans. Econ. Comput.* 1, 3 (2013), 16.
[57] Faruk Caglar and Aniruddha Gokhale. 2014. iOverbook: Intelligent resource-overbooking to support soft real-time applications in the cloud. In *Proceedings of the IEEE 7th International Conference on Cloud Computing (CLOUD'14)*. IEEE, 538–545.
[58] Breno G. S. Costa, Marco Antonio Sousa Reis, Aletéia P. F. Araújo, and Priscila Solis. 2018. Performance and cost analysis between on-demand and preemptive virtual machines. In *Proceedings of the International Conference on Cloud Computing and Services Science*. 169–178.
[59] Juan Li, Yanmin Zhu, Jiadi Yu, Chengnian Long, Guangtao Xue, and Shiyou Qian. 2017. Online auction for IaaS clouds: Towards elastic user demands and weighted heterogeneous VMs. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM'17)*. IEEE, 1–9.
[60] Vivek Kumar Singh and Kaushik Dutta. 2015. Dynamic price prediction for Amazon spot instances. In *Proceedings of the 48th Hawaii International Conference on System Sciences (HICSS'15)*. IEEE, 1513–1520.
[61] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica et al. 2010. A view of cloud computing. *Commun. ACM* 53, 4 (2010), 50–58.
[62] Shijimol Ambi Karthikeyan. 2018. Introduction to Azure IaaS. In *Practical Microsoft Azure IaaS*. Springer, 1–38.
[63] Ashish Kumar Mishra, Brajesh Kumar Umrao, and Dharmendra K. Yadav. 2018. A survey on optimal utilization of preemptible VM instances in cloud computing. *J. Supercomput.* 74, 11 (2018), 5980–6032.
[64] Jose Pergentino Araujo Neto, Donald M and Ralha Pianto, and Ghedini Célia. 2019. MULTS: A multi-cloud fault-tolerant architecture to manage transient servers in cloud computing. *Journal of Systems Architecture* 101 (2019), 101651.
[65] Jogesh Muppala, Gianfranco Ciardo, and Kishor S. Trivedi. 1994. Stochastic reward nets for reliability prediction. *Commun. Reliab., Maintainab. Serviceab.* 1, 2 (1994), 9–20.
[66] Daniel Ford, François Labelle, Florentina I. Popovici, Murray Stokely, Van-Anh Truong, Luiz Barroso, Carrie Grimes, and Sean Quinlan. 2010. Availability in globally distributed storage systems. In *Proceedings of the Symposium on Operating Systems Design and Implementation (OSDI'10)*, Vol. 10. 1–7.

[67] Peter Bodik, Armando Fox, Michael J. Franklin, Michael I. Jordan, and David A. Patterson. 2010. Characterizing, modeling, and generating workload spikes for stateful services. In *Proceedings of the 1st ACM Symposium on Cloud Computing*. ACM, 241–252.

[68] Timothy Wood, Emmanuel Cecchet, Kadangode K. Ramakrishnan, Prashant J. Shenoy, Jacobus E. van der Merwe, and Arun Venkataramani. 2010. Disaster recovery as a cloud service: Economic benefits & deployment challenges. *HotCloud* 10 (2010), 8–15.

[69] Surajit Chaudhuri. 2012. What next?: A half-dozen data management research goals for big data and the cloud. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*. ACM, 1–4.

[70] Hariharasudhan Viswanathan, Eun Kyung Lee, Ivan Rodero, and Dario Pompili. 2015. Uncertainty-aware autonomic resource provisioning for mobile cloud computing. *IEEE Trans. Parallel Distrib. Syst.* 26, 8 (2015), 2363–2372.

[71] Joe Long and Dan McCurley. 2018. Parallel cloud computing: Making massive actuarial risk analysis possible. *Predict. Anal. Futur.* 17 (2018), 6.

[72] Bruno Lopes Dalmazo, João P. Vilela, and Marilia Curado. 2013. Predicting traffic in the cloud: A statistical approach. In *Proceedings of the 3rd International Conference on Cloud and Green Computing (CGC'13)*. IEEE, 121–126.

[73] Hitesh Ballani, Paolo Costa, Thomas Karagiannis, and Ant Rowstron. 2011. Towards predictable datacenter networks. *ACM SIGCOMM Comput. Commun. Rev.* 41 (2011). ACM, 242–253.

[74] Theophilus Benson, Aditya Akella, and David A. Maltz. 2010. Network traffic characteristics of data centers in the wild. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*. ACM, 267–280.

[75] Sangho Yi, Derrick Kondo, and Artur Andrzejak. 2010. Reducing costs of spot instances via checkpointing in the amazon elastic compute cloud. In *Proceedings of the IEEE 3rd International Conference on Cloud Computing (CLOUD'10)*. IEEE, 236–243.

[76] Richard Wolski et al. 1997. Forecasting network performance to support dynamic scheduling using the network weather service. In *Proceedings of the ACM International Symposium on High-performance Parallel and Distributed Computing (HPDC'97)*, Vol. 97. 316.

[77] Yang Xinyu, Zeng Ming, Zhao Rui, and Shi Yi. 2004. A novel LMS method for real-time network traffic prediction. In *Proceedings of the International Conference on Computational Science and Its Applications*. Springer, 127–136.

[78] Bruno L. Dalmazo, João P. Vilela, and Marilia Curado. 2016. Online traffic prediction in the cloud. *Int. J. Netw. Manag.* 26, 4 (2016), 269–285.

[79] Deborah Magalhães, Rodrigo N. Calheiros, Rajkumar Buyya, and Danielo G. Gomes. 2015. Workload modeling for resource usage analysis and simulation in cloud computing. *Comput. Electric. Eng.* 47 (2015), 69–81.

[80] Jingqi Yang, Chuanchang Liu, Yanlei Shang, Bo Cheng, Zexiang Mao, Chunhong Liu, Lisha Niu, and Junliang Chen. 2014. A cost-aware auto-scaling approach using the workload prediction in service clouds. *Inf. Syst. Front.* 16, 1 (2014), 7–18.

[81] Quan Liang, Jing Zhang, Yong-hui Zhang, and Jiu-mei Liang. 2014. The placement method of resources and applications based on request prediction in cloud data center. *Inf. Sci.* 279 (2014), 735–745.

[82] Weijia Song, Zhen Xiao, Qi Chen, and Haipeng Luo. 2014. Adaptive resource provisioning for the cloud using online bin packing. *IEEE Trans. Comput.* 63, 11 (2014), 2647–2660.

[83] Yexi Jiang, Chang-Shing Perng, Tao Li, and Rong N. Chang. 2013. Cloud analytics for capacity planning and instant VM provisioning. *IEEE Trans. Netw. Serv. Manag.* 10, 3 (2013), 312–325.

[84] Saurabh Kumar Garg, Adel Nadjaran Toosi, Srinivasa K. Gopalaiyengar, and Rajkumar Buyya. 2014. SLA-based virtual machine management for heterogeneous workloads in a cloud datacenter. *J. Netw. Comput. Applic.* 45 (2014), 108–120.

[85] Jhu-Jyun Jheng, Fan-Hsun Tseng, Han-Chieh Chao, and Li-Der Chou. 2014. A novel VM workload prediction using Grey Forecasting model in cloud data center. In *Proceedings of the International Conference on Information Networking (ICOIN'14)*. IEEE, 40–45.

[86] Rodrigo N. Calheiros, Rajiv Ranjan, and Rajkumar Buyya. 2011. Virtual machine provisioning based on analytical performance and QoS in cloud computing environments. In *Proceedings of the International Conference on Parallel Processing (ICPP'11)*. IEEE, 295–304.

[87] Microsoft Corporation. 2009. Optimal Workloads for the cloud. https://blogs.msdn.microsoft.com/stevecla01/2009/11/26/optimal-workloads-for-the-cloud/.

[88] Gilles Madi Wamba, Yunbo Li, Anne-Cécile Orgerie, Nicolas Beldiceanu, and Jean-Marc Menaud. 2017. Cloud workload prediction and generation models. In *Proceedings of the 29th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD'17)*. IEEE, 89–96.

[89] Rui Cao, Zhaoyang Yu, Trent Marbach, Jing Li, Gang Wang, and Xiaoguang Liu. 2018. Load prediction for data centers based on database service. In *Proceedings of the IEEE 42nd Computer Software and Applications Conference (COMPSAC'18)*. IEEE, 728–737.

[90] Tham Nguyen, Doan Hoang, Diep Nguyen, and Aruna Seneviratne. 2017. Initial trust establishment for personal space IoT systems. In *Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS'17)*. IEEE, 784–789.

[91] Fahed Alkhabbas, Ilir Murturi, Romina Spalazzese, Paul Davidsson, and Schahram Dustdar. 2020. A goal-driven approach for deploying self-adaptive IoT systems. In *Proceedings of the IEEE International Conference on Software Architecture (ICSA'20)*. IEEE, 146–156.

[92] Sarah Maroc and Jian Biao Zhang. 2020. Cloud services security-driven evaluation for multiple tenants. *Cluster Comput.* (2020), 1–19. DOI: 10.1007/s10586-020-03178-z

[93] Ashish Singh and Kakali Chatterjee. 2017. Cloud security issues and challenges: A survey. *J. Netw. Comput. Applic.* 79 (2017), 88–115.

[94] Syed Rizvi, Jungwoo Ryoo, John Kissell, William Aiken, and Yuhong Liu. 2018. A security evaluation framework for cloud security auditing. *J. Supercomput.* 74, 11 (2018), 5774–5796.

[95] Marshall Pease, Robert Shostak, and Leslie Lamport. 1980. Reaching agreement in the presence of faults. *J. ACM* 27, 2 (1980), 228–234.

[96] Leslie Lamport, Robert Shostak, and Marshall Pease. 1982. The Byzantine generals problem. *ACM Trans. Prog. Lang. Syst.* 4, 3 (1982), 382–401.

[97] Dhruba Borthakur. 2007. The Hadoop distributed file system: Architecture and design. *Hadoop Proj. Website* 11 (2007), 21.

[98] Subrota K. Mondal, Abadhan S. Sabyasachi, and Jogesh K. Muppala. 2017. On dependability, cost and security tradeoff in cloud data centers. In *Proceedings of the IEEE 22nd Pacific Rim International Symposium on Dependable Computing (PRDC'17)*. IEEE, 11–19.

[99] Luke M. Leslie, Young Choon Lee, and Albert Y. Zomaya. 2015. RAMP: Reliability-aware elastic instance provisioning for profit maximization. *J. Supercomput.* 71, 12 (2015), 4529–4554.

[100] Wentao Wu, Xi Wu, Hakan Hacigümüş, and Jeffrey F. Naughton. 2014. Uncertainty aware query execution time prediction. *Proc. VLDB Endow.* 7, 14 (2014), 1857–1868.

[101] Guiding Metrics. 2018. *The Cloud Service Industry's 10 Most Critical Metrics*. https://guidingmetrics.com/content/cloud-services-industrys-10-most-critical-metrics/.

[102] Kuljeet Kaur, Tanya Dhand, Neeraj Kumar, and Sherali Zeadally. 2017. Container-as-a-service at the edge: Trade-off between energy efficiency and service availability at fog nano data centers. *IEEE Wirel. Commun.* 24, 3 (2017), 48–56.

[103] Saurabh Kumar Garg, Steve Versteeg, and Rajkumar Buyya. 2011. Smicloud: A framework for comparing and ranking cloud services. In *Proceedings of the 4th IEEE International Conference on Utility and Cloud Computing (UCC'11)*. IEEE, 210–218.

[104] Ahmad Khalil, Nader Mbarek, and Olivier Togni. 2018. Self-configuring IoT service QoS guarantee using QBAIoT. *Computers* 7, 4 (2018), 64.

[105] Mohammed Alodib. 2016. QoS-aware approach to monitor violations of SLAs in the IoT. *J. Innov. Dig. Ecosyst.* 3, 2 (2016), 197–207.

[106] Jinesh Varia. 2011. Best practices in architecting cloud applications in the AWS cloud. *Cloud Comput.: Princ. Parad.* 18 (2011), 459–490.

[107] Flavia C. Delicato, Adnan Al-Anbuky, I. Kevin, and Kai Wang. 2020. Smart cyber-physical systems: Toward pervasive intelligence systems. 107 (2020), 1134–1139. DOI: https://doi.org/10.1016/j.future.2019.06.031

[108] Chenhao Qu, Rodrigo N. Calheiros, and Rajkumar Buyya. 2018. Auto-scaling web applications in clouds: A taxonomy and survey. *ACM Comput. Surv.* 51, 4 (2018), 1–33.

[109] Wei Ai, Kenli Li, Shenglin Lan, Fan Zhang, Jing Mei, Keqin Li, and Rajkumar Buyya. 2016. On elasticity measurement in cloud computing. *Sci. Prog.* 2016 (2016).

[110] Liang Zheng, Carlee Joe-Wong, Chee Wei Tan, Mung Chiang, and Xinyu Wang. 2015. How to bid the cloud. *ACM SIGCOMM Comput. Commun. Rev.* 45 (2015). ACM, 71–84.

[111] Rich Wolski, John Brevik, Ryan Chard, and Kyle Chard. 2017. Probabilistic guarantees of execution duration for Amazon spot instances. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, 18.

[112] H. M. Dipu Kabir, Abadhan S. Sabyasachi, Abbas Khosravi, M. Anwar Hosen, Saeid Nahavandi, and Rajkumar Buyya. 2019. A cloud bidding framework for deadline constrained jobs. In *Proceedings of the IEEE International Conference on Industrial Technology (ICIT'19)*. IEEE, 765–772.

[113] Qihang Sun, Chuan Wu, Zongpeng Li, and Shaolei Ren. 2016. Colocation demand response: Joint online mechanisms for individual utility and social welfare maximization. *IEEE J. Sel. Areas Commun.* 34, 12 (Dec. 2016), 3978–3992. DOI: http://dx.doi.org/10.1109/JSAC.2016.2611918

[114] Liang Zheng, Carlee Joe-Wong, Chee Wei Tan, Mung Chiang, and Xinyu Wang. 2015. How to bid the cloud. *ACM SIGCOMM Comput. Commun. Rev.* 45, 5 (Aug. 2015), 71–84. DOI: http://dx.doi.org/10.1145/2829988.2787473

[115] Linquan Zhang, Zongpeng Li, and Chuan Wu. 2014. Dynamic resource provisioning in cloud computing: A randomized auction approach. *Proceedings of the IEEE IEEE Conference on Computer Communications (INFOCOM'14)*. 433–441. DOI: Retrieved from http://dx.doi.org/10.1109/INFOCOM.2014.6847966

[116] Roshni Pary. 2018. New Amazon EC2 Spot pricing model: Simplified purchasing without bidding and fewer interruptions. Retrieved from https://aws.amazon.com/blogs/compute/new-amazon-ec2-spot-pricing/.

[117] Kimitoshi Sato and Kenichi Nakashima. 2020. Optimal pricing problem for a pay-per-use system based on the Internet of Things with intertemporal demand. *Int. J. Prod. Econ.* 221 (2020), 107477.

[118] Alireza Salehan, Hossein Deldari, and Saeid Abrishami. 2017. An online valuation-based sealed winner-bid auction game for resource allocation and pricing in clouds. *J. Supercomput.* 73, 11 (2017), 4868–4905.

[119] Daniel A. Menasce, Virgilio A. F. Almeida, Lawrence W. Dowdy, and Larry Dowdy. 2004. *Performance by Design: Computer Capacity Planning by Example*. Prentice Hall Professional.

[120] Daniel Gmach, Jerry Rolia, Ludmila Cherkasova, and Alfons Kemper. 2007. Workload analysis and demand prediction of enterprise data center applications. In *Proceedings of the IEEE 10th International Symposium on Workload Characterization (IISWC'07)*. IEEE, 171–180.

[121] Marcus Carvalho, Daniel A. Menascé, and Francisco Brasileiro. 2017. Capacity planning for IaaS cloud providers offering multiple service classes. *Fut. Gen. Comput. Syst.* 77 (2017), 97–111.

[122] Daniel Gmach, Jerry Rolia, Ludmila Cherkasova, and Alfons Kemper. 2007. Capacity management and demand prediction for next generation data centers. In *Proceedings of the IEEE International Conference on Web Services (ICWS'07)*. 43–50. DOI: 10.1109/ICWS.2007.62

[123] Tom Krazit. 2018. How Amazon Web Services uses machine learning to make capacity planning decisions. Retrieved from https://www.geekwire.com/2017/amazon-web-services-uses-machine-learning-make-capacity-planning-decisions/.

[124] Song Yang, Fernando A. Kuipers et al. 2014. Traffic uncertainty models in network planning. *IEEE Commun. Mag.* 52, 2 (2014), 172–177.

[125] David Applegate and Edith Cohen. 2006. Making routing robust to changing traffic demands: Algorithms and evaluation. *IEEE/ACM Trans. Netw.* 14, 6 (2006), 1193–1206.

[126] Murali Kodialam, T. V. Lakshman, James B. Orlin, and Sudipta Sengupta. 2009. Oblivious routing of highly variable traffic in service overlays and IP backbones. *IEEE/ACM Trans. Netw.* 17, 2 (2009), 459–472.

[127] Alexandre Fréchette, F. Bruce Shepherd, Marina K. Thottan, and Peter J. Winzer. 2015. Shortest path versus multihub routing in networks with uncertain demand. *IEEE/ACM Trans. Netw.* 23, 6 (2015), 1931–1943.

[128] Walid Ben-Ameur and Hervé Kerivin. 2005. Routing of uncertain traffic demands. *Optim. Eng.* 6, 3 (2005), 283–313.

[129] Arie M. C. A. Koster, Manuel Kutschka, and Christian Raack. 2013. Robust network design: Formulations, valid inequalities, and computations. *Networks* 61, 2 (2013), 128–149.

[130] Ramon Aparicio-Pardo, Pablo Pavon-Marino, and Biswanath Mukherjee. 2012. Robust upgrade in optical networks under traffic uncertainty. In *Proceedings of the 16th International Conference on Optical Network Design and Modeling (ONDM'12)*. IEEE, 1–6.

[131] Ariel Orda, Raphael Rom, and Moshe Sidi. 1993. Minimum delay routing in stochastic networks. *IEEE/ACM Trans. Netw.* 1, 2 (1993), 187–198.

[132] Elise Miller-Hooks. 2001. Adaptive least-expected time paths in stochastic, time-varying transportation and data networks. *Netw.: Int. J.* 37, 1 (2001), 35–52.

[133] Ying Xiao, Krishnaiyan Thulasiraman, Xi Fang, Dejun Yang, and Guoliang Xue. 2012. Computing a most probable delay constrained path: NP-hardness and approximation schemes. *IEEE Trans. Comput.* 61, 5 (2012), 738–744.

[134] Masoumeh Tajvidi, Michael J. Maher, and Daryl Essam. 2017. Uncertainty-aware optimization of resource provisioning, a cloud end-user perspective. In *Proceedings of the International Conference on Cloud Computing and Services Science (CLOSER'17)*. 293–300.

[135] Pooyan Jamshidi, Claus Pahl, and Nabor C. Mendonça. 2016. Managing uncertainty in autonomic cloud elasticity controllers. *IEEE Cloud Comput.* 3, 3 (2016), 50–60.

[136] Marco L. Della Vedova, Daniele Tessera, and Maria Carla Calzarossa. 2016. Probabilistic provisioning and scheduling in uncertain cloud environments. In *Proceedings of the IEEE Symposium on Computers and Communication (ISCC'16)*. IEEE, 797–803.

[137] Mohamed Amine Ferrag, Leandros A. Maglaras, Helge Janicke, Jianmin Jiang, and Lei Shu. 2017. Authentication protocols for internet of things: A comprehensive survey. *Secur. Commun. Netw.* 2017 (2017).

[138] Jangirala Srinivas, Ashok Kumar Das, Neeraj Kumar, and Joel J. P. C. Rodrigues. 2019. TCALAS: Temporal credential-based anonymous lightweight authentication scheme for Internet of drones environment. *IEEE Trans. Vehic. Technol.* 68, 7 (2019), 6903–6916.

[139] Nima Karimian, Paul A. Wortman, and Fatemeh Tehranipoor. 2016. Evolving authentication design considerations for the internet of biometric things (IoBT). In *Proceedings of the 11th IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*. 1–10.

[140] Changhee Hahn, Jongkil Kim, Hyunsoo Kwon, and Junbeom Hur. 2020. Efficient IoT management with resilience to unauthorized access to cloud storage. *IEEE Trans. Cloud Comput.* (2020). DOI: 10.1109/TCC.2020.2985046
[141] Ping Zhang, Mimoza Durresi, and Arjan Durresi. 2019. Multi-access edge computing aided mobility for privacy protection in internet of things. *Computing* 101, 7 (2019), 729–742.
[142] Mouna Jouini and Latifa Ben Arfa Rabai. 2019. A security framework for secure cloud computing environments. In *Cloud Security: Concepts, Methodologies, Tools, and Applications*. IGI Global, 249–263.
[143] Pandi Vijayakumar, Victor Chang, L. Jegatha Deborah, Balamurugan Balusamy, and P. G. Shynu. 2018. Computationally efficient privacy preserving anonymous mutual and batch authentication schemes for vehicular ad hoc networks. *Fut. Gen. Comput. Syst.* 78 (2018), 943–955.
[144] Kavous-Fard Abdollah, Wencong Su, and Tao Jin. 2020. A machine learning based cyber attack detection model for wireless sensor networks in microgrids. *IEEE Trans. Industr. Inform.* 17, 1 (2020), 650–658.
[145] Mohammad Ghiasi, Moslem Dehghani, Taher Niknam, and Abdollah Kavousi-Fard. 2020. Investigating overall structure of cyber-attacks on smart-grid control systems to improve cyber resilience in power system. *Network* 1, 1 (2020).
[146] Xuanxia Yao, Zhi Chen, and Ye Tian. 2015. A lightweight attribute-based encryption scheme for the Internet of Things. *Fut. Gen. Comput. Syst.* 49 (2015), 104–112.
[147] T. P. Sharma et al. 2020. Lightweight encryption algorithms, technologies, and architectures in Internet of Things: A survey. In *Innovations in Computer Science and Engineering*. Springer, 341–351.
[148] Christos Stergiou, Kostas E. Psannis, Byung-Gyu Kim, and Brij Gupta. 2018. Secure integration of IoT and cloud computing. *Fut. Gen. Comput. Syst.* 78 (2018), 964–975.
[149] Algirdas Avizienis, J.-C. Laprie, Brian Randell, and Carl Landwehr. 2004. Basic concepts and taxonomy of dependable and secure computing. *IEEE Trans. Depend. Sec. Comput.* 1, 1 (2004), 11–33.
[150] E. Bauer and R. Adams. 2012. *Reliability and Availability of Cloud Computing*. Wiley-IEEE Press.
[151] K. S. Trivedi, D. Wang, and J. Hunt. 2010. Computing the number of calls dropped due to failures. In *Proceedings of the IEEE 21st International Symposium on Software Reliability Engineering (ISSRE'10)*. IEEE, 11–20.
[152] Subrota K. Mondal, Xiaoyan Yin, Jogesh K. Muppala, Javier Alonso Lopez, and Kishor S. Trivedi. 2015. Defects per million computation in service-oriented environments. *IEEE Trans. Serv. Comput.* 8, 1 (2015), 32–46.
[153] Subrota K. Mondal, Fumio Machida, and Jogesh K. Muppala. 2016. Service reliability enhancement in cloud by checkpointing and replication. In *Principles of Performance and Reliability Modeling and Evaluation*. Springer, 425–448.
[154] Zia-ur Rehman, Omar Khadeer Hussain, and Farookh Khadeer Hussain. 2015. User-side cloud service management: State-of-the-art and future directions. *J. Netw. Comput. Applic.* 55 (2015), 108–122.
[155] Sukhpal Singh and Inderveer Chana. 2015. QoS-aware autonomic resource management in cloud computing: A systematic review. *ACM Comput. Surv.* 48, 3 (2015), 1–46.
[156] Zibin Zheng, Xinmiao Wu, Yilei Zhang, Michael R. Lyu, and Jianmin Wang. 2013. QoS ranking prediction for cloud services. *IEEE Trans. Parallel Distrib. Syst.* 24, 6 (2013), 1213–1222.
[157] Zia ur Rehman, Omar Khadeer Hussain, and Farookh Khadeer Hussain. 2014. Parallel cloud service selection and ranking based on QoS history. *Int. J. Parallel Prog.* 42, 5 (2014), 820–852.
[158] Hangwei Qian, Hualong Zu, Chenghua Cao, and Qixin Wang. 2013. CSS: Facilitate the cloud service selection in IaaS platforms. In *Proceedings of the International Conference on Collaboration Technologies and Systems (CTS'13)*. IEEE, 347–354.
[159] Karim Benouaret, Dimitris Sacharidis, Djamal Benslimane, and Allel Hadjali. 2018. Selecting services for multiple users: Let's be democratic. *IEEE Trans. Serv. Comput.* (2018). DOI: 10.1109/TSC.2018.2875691
[160] Sangho Yi, Artur Andrzejak, and Derrick Kondo. 2012. Monetary cost-aware checkpointing and migration on Amazon cloud spot instances. *IEEE Trans. Serv. Comput.* 5, 4 (2012), 512–524.
[161] Zhiquan Sui, Matthew Malensek, Neil Harvey, and Shrideep Pallickara. 2015. Autonomous orchestration of distributed discrete event simulations in the presence of resource uncertainty. *ACM Trans. Autonom. Adapt. Syst.* 10, 3 (2015), 18.
[162] Zaeem Hussain, Taieb Znati, and Rami Melhem. 2018. Partial redundancy in HPC systems with non-uniform node reliabilities. In *Partial Redundancy in HPC Systems with Non-uniform Node Reliabilities*. IEEE.
[163] Walayat Hussain, Farookh Khadeer Hussain, Morteza Saberi, Omar Khadeer Hussain, and Elizabeth Chang. 2018. Comparing time series with machine learning-based prediction approaches for violation management in cloud SLAs. *Fut. Gen. Comput. Syst.* 89 (2018), 464–477.
[164] Parham M. Kebria, Abbas Khosravi, Syed Moshfeq Salaken, Ibrahim Hossain, H. M. Dipu Kabir, Afsaneh Koohestani, Roohallah Alizadehsani, and Saeid Nahavandi. 2018. Deep imitation learning: The impact of depth on policy performance. In *Proceedings of the International Conference on Neural Information Processing*. Springer, 172–181.
[165] Shouping Guan and Zhouying Cui. 2020. Modeling uncertain processes with interval random vector functional-link networks. *J. Proc. Contr.* 93 (2020), 43–52.

[166] Seyed Mohammad Jafar Jalali, Sajad Ahmadian, Abbas Khosravi, Seyedali Mirjalili, Mohammad Reza Mahmoudi, and Saeid Nahavandi. 2020. Neuroevolution-based Autonomous Robot Navigation: A Comparative Study. *Cog. Syst. Res.* 62 (2020), 35–43.

[167] Matjaž Perc, Mahmut Ozer, and Janja Hojnik. 2019. Social and juristic challenges of artificial intelligence. *Palgrave Commun.* 5, 1 (2019), 1–7.

[168] Kinza Shafique, Bilal A. Khawaja, Farah Sabir, Sameer Qazi, and Muhammad Mustaqim. 2020. Internet of things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT scenarios. *IEEE Access* 8 (2020), 23022–23040.

[169] Abu Sufian, Dharm Singh Jat, and Anuradha Banerjee. 2020. Insights of artificial intelligence to stop spread of Covid-19. In *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach*. Springer, 177–190.

[170] Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, and Marimuthu Palaniswami. 2013. Internet of Things (IoT): A vision, architectural elements, and future directions. *Fut. Gen. Comput. Syst.* 29, 7 (2013), 1645–1660.

[171] Sukhpal Singh Gill, Shreshth Tuli, Minxian Xu, Inderpreet Singh, Karan Vijay Singh, Dominic Lindsay, Shikhar Tuli, Daria Smirnova, Manmeet Singh, Udit Jain et al. 2019. Transformative effects of IoT, Blockchain and Artificial Intelligence on cloud computing: Evolution, vision, trends and open challenges. *Internet Things J.* 8 (2019), 100118.

[172] Mamdooh Al-Saud, Ali M. Eltamaly, Mohamed A. Mohamed, and Abdollah Kavousi-Fard. 2019. An intelligent data-driven model to secure intravehicle communications based on machine learning. *IEEE Trans. Industr. Electron.* 67, 6 (2019), 5112–5119.

[173] Mingxi Cheng, Ji Li, and Shahin Nazarian. 2018. DRL-cloud: Deep reinforcement learning-based resource provisioning and task scheduling for cloud service providers. In *Proceedings of the 23rd Asia and South Pacific Design Automation Conference (ASP-DAC'18)*. IEEE, 129–134.

[174] Sukhpal Singh Gill, Inderveer Chana, Maninder Singh, and Rajkumar Buyya. 2018. CHOPPER: An intelligent QoS-aware autonomic resource management approach for cloud computing. *Cluster Comput.* 21, 2 (2018), 1203–1241.

[175] Domenico Talia. 2011. Cloud computing and software agents: Towards cloud intelligent services. In *Proceedings of the International Workshop on Optimization and Applications*, Vol. 11. Citeseer, 2–6.

[176] Sukhpal Singh and Inderveer Chana. 2016. Resource provisioning and scheduling in clouds: QoS perspective. *J. Supercomput.* 72, 3 (2016), 926–960.

[177] Abdollah Kavousi Fard and Mohammad-Reza Akbari-Zadeh. 2014. A hybrid method based on wavelet, ANN and ARIMA model for short-term load forecasting. *J. Exper. Theoret. Artif. Intell.* 26, 2 (2014), 167–182.

[178] Abdollah Kavousi-Fard, Taher Niknam, Hoda Taherpoor, and Alireza Abbasi. 2014. Multi-objective probabilistic reconfiguration considering uncertainty and multi-level load model. *IET Sci., Meas. Technol.* 9, 1 (2014), 44–55.

[179] Eric M. Dashofy. 2019. Software engineering in the cloud. In *Handbook of Software Engineering*. Springer, 491–516.

[180] Suyash S. Ghuge, Nishant Kumar, S. Savitha, and V. Suraj. 2020. Multilayer technique to secure data transfer in private cloud for SaaS applications. In *Proceedings of the 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA'02)*. IEEE, 646–651.

[181] Kaiyuan Guo, Song Han, Song Yao, Yu Wang, Yuan Xie, and Huazhong Yang. 2017. Software-hardware codesign for efficient neural network acceleration. *IEEE Micro* 37, 2 (2017), 18–25.

[182] Jeff Dean, David Patterson, and Cliff Young. 2018. A new golden age in computer architecture: Empowering the machine-learning revolution. *IEEE Micro* 38, 2 (2018), 21–29.

[183] H. M. Kabir, Abbas Khosravi, Abdollah Kavousi-Fard, Saeid Nahavandi, and Dipti Srinivasan. 2019. Optimal uncertainty-guided neural network training. *arXiv preprint arXiv:1912.12761* (2019).

[184] Holger R. Maier, Joseph H. A. Guillaume, Hedwig van Delden, Graeme A. Riddell, Marjolijn Haasnoot, and Jan H. Kwakkel. 2016. An uncertain future, deep uncertainty, scenarios, robustness and adaptation: How do they fit together? *Envir. Model. Softw.* 81 (2016), 154–164.

[185] Shashikant Ilager, Rajeev Muralidhar, Kotagiri Rammohanrao, and Rajkumar Buyya. 2020. A data-driven frequency scaling approach for deadline-aware energy efficient scheduling on graphics processing units (GPUs). *arXiv preprint arXiv:2004.08177* (2020).

[186] Maryam Khodayari and Alireza Aslani. 2018. Analysis of the energy storage technology using Hype Cycle approach. *Sustain. Energy Technol. Assess.* 25 (2018), 60–74.

[187] Vladimir Hahanov, Wajeb Gharibi, Ka Lok Man, Igor Iemelianov, Mykhailo Liubarskyi, Vugar Abdullayev, Eugenia Litvinova, and Svetlana Chumachenko. 2018. Cyber-physical technologies: Hype cycle 2017. In *Cyber Physical Computing for IoT-driven Services*. Springer, 259–272.

[188] Eric Minick. 2018. Machine Learning: Essential in Cloud Service Management. Retrieved from https://devops.com/machine-learning-essential-in-cloud-service-management/.

[189] Chandrashekar Jatoth, G. R. Gangadharan, Ugo Fiore, and Rajkumar Buyya. 2018. SELCLOUD: A hybrid multi-criteria decision-making model for selection of cloud services. *Soft Comput.* 23, 13 (2019), 4701–4715.

[190] Sajib Mistry, Athman Bouguettaya, and Hai Dong. 2018. Service providers' long-term QoS prediction model. In *Economic Models for Managing Cloud Services*. Springer, 111–122.
[191] Wenrui Li, Pengcheng Zhang, Hareton Leung, and Shunhui Ji. 2018. A novel QoS prediction approach for cloud services using Bayesian network model. *IEEE Access* 6 (2018), 1391–1406.
[192] Shiva Prakash et al. 2019. Review of quality of service based techniques in cloud computing. In *Data Science and Big Data Analytics*. Springer, 255–265.
[193] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*. 1050–1059.
[194] Aditi D. Joshi and Surendra M. Gupta. 2019. Evaluation of design alternatives of end-of-life products using internet of things. *Int. J. Prod. Econ.* 208 (2019), 281–293.
[195] M. A. López-Medina, Macarena Espinilla, Ian Cleland, C. Nugent, and Javier Medina. 2020. Fuzzy cloud-fog computing approach application for human activity recognition in smart homes. *J. Intell. Fuzzy Syst.* 38, 1 (2020), 709–721.
[196] Amazon.com, Inc. 2018. Amazon EC2 Spot Lets you Pause and Resume Your Workloads. https://aws.amazon.com/about-aws/whats-new/2017/11/amazon-ec2-spot-lets-you-pause-and-resume-your-workloads/.
[197] Agmon Ben-Yehuda, Orna, et al. 2013. Deconstructing Amazon EC2 spot instance pricing. *ACM Transactions on Economics and Computation (TEAC)* 1, 3 (2013), 1–20.
[198] Ma, Junming, et al. 2020. PrTaurus: An Availability-Enhanced EMR Service on preemptible cloud instances. In *2020 IEEE International Conference on Web Services (ICWS'20)*. IEEE.
[199] Cardellini, Valeria, Valerio Di Valerio, and Francesco Lo Presti. 2016. Game-theoretic resource pricing and provisioning strategies in cloud systems. *IEEE Transactions on Services Computing* 13, 1 (2016), 86–98.
[200] Zheng, Liang, et al. 2015. How to bid the cloud. *ACM SIGCOMM Computer Communication Review* 45, 4 (2015), 71–84.
[201] Zheng, Liang, et al. 2016. On the viability of a cloud virtual service provider. *ACM SIGMETRICS Performance Evaluation Review* 44, 1 (2016), 235–248.

Received July 2020; revised November 2020; accepted January 2021