Digital Communications and Networks 3 (2017) 101–111

HOSTED BY

Contents lists available at ScienceDirect

Digital Communications and Networks

journal homepage: www.elsevier.com/locate/dcan

Cloud service performance evaluation: status, challenges, and opportunities – a survey from the system modeling perspective☆

Qiang Duan*

Information Sciences & Technology Department, The Pennsylvania State University, Abington College, USA

ARTICLE INFO

Keywords:
Cloud computing
Cloud services
Quality of Service
Performance evaluation

ABSTRACT

With rapid advancement of Cloud computing and networking technologies, a wide spectrum of Cloud services have been developed by various providers and utilized by numerous organizations as indispensable ingredients of their information systems. Cloud service performance has a significant impact on performance of the future information infrastructure. Thorough evaluation on Cloud service performance is crucial and beneficial to both service providers and consumers; thus forming an active research area. Some key technologies for Cloud computing, such as virtualization and the Service-Oriented Architecture (SOA), bring in special challenges to service performance evaluation. A tremendous amount of effort has been put by the research community to address these challenges and exciting progress has been made. Among the work on Cloud performance analysis, evaluation approaches developed with a system modeling perspective play an important role. However, related works have been reported in different sections of the literature; thus lacking a big picture that shows the latest status of this area. The objectives of this article is to present a survey that reflects the state of the art of Cloud service performance evaluation from the system modeling perspective. This articles also examines open issues and challenges to the surveyed evaluation approaches and identifies possible opportunities for future research in this important field.

1. Introduction

Cloud computing is a large scale distributed computing paradigm driven by economies of scale, in which a pool of abstracted, virtualized, dynamically scalable computing resources are delivered on demand as services to external customers over networks. Key characteristics of this emerging computing paradigm include the illusion of infinite computing resources; the elimination of an up-front commitment by Cloud users; and the ability to pay for use as needed [1]. Virtualization, which abstracts data center hardware for supporting virtual machines, forms a technical foundation in Cloud computing for realizing these features [4].

The rapid advances in Cloud computing technologies in the past decade have enabled a wide spectrum of Cloud services. Cloud services are services made available to users on demand via networks from data centers operated by Cloud computing providers. A Cloud service can dynamically scale to meet the needs of its users, and because the service provider supplies the hardware and software necessary for the service, there is no need for a user to provision or deploy its own resources or allocate IT staff to manage the service. The Service-Oriented Architecture (SOA) [19], which encapsulates computational resources through abstract interfaces to decouple services from their implementations, plays a key role in Cloud service provisioning. Typical Cloud service models include Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS).

Cloud computing allows multiple deployment models including public, private, and hybrid Cloud infrastructures. In addition to public commercial Cloud services offered by major service providers such as Amazon, Google, and Microsoft, numerous organizations have constructed their own private Clouds or established hybrid Clouds to provide various services to their employees and/or customers. More recently, developments in network technologies have significantly improved data transmission throughput and network control/management functions, which makes the new Cloud federation paradigm possible. Cloud federation allows computing resources in different data centers to be orchestrated to provide composite services to end users, which may greatly enrich Cloud service provisioning.

Recent innovations in the field of networking may further broaden the landscape of Cloud services. Virtualization has been adopted as a key attribute in future networking through the recently proposed Network Function Virtualization (NFV) [20]. The SOA principle has also been applied to networking, thus introducing the Network-as-a-

Peer review under responsibility of Chongqing University of Posts and Telecommunications.
* Corresponding author.
E-mail address: qduan@psu.edu.

http://dx.doi.org/10.1016/j.dcan.2016.12.002
Received 22 July 2015; Received in revised form 14 December 2016; Accepted 15 December 2016
Available online 23 December 2016

Service (NaaS) paradigm. These emerging technologies may bridge the gap between networking and Cloud computing to support convergence of network and Cloud service provisioning [17]. Such a convergence may stimulate innovations in service developments that lead to a wide variety of composite network-Cloud services.

With the rapid developments in Cloud computing and networking technologies, a wide variety of Cloud services have been adopted as dispensable ingredients in the information systems of numerous organizations; therefore Cloud service performance has a significant impact on the performance of the entire future information infrastructure. Cloud service providers need to have deep insights about the relationship between service performance and available resources in order to fully utilize their infrastructures while meeting users' performance requirements. Cloud service consumers also want effective methods for evaluating the performance that can be guaranteed by various Cloud services, especially different offers of the same service function, in order to make decisions on optimal service selection and composition. Evaluation on Cloud service performance is crucial and beneficial for both service providers and consumers.

On the other hand, key Cloud technologies such as virtualization and SOA bring in new challenges to service performance evaluation. Virtualization enables heterogeneous services to be hosted upon shared abstract infrastructures. The SOA decouples Cloud service functions from their implementations, thus allowing the same service to be realized on various host platforms with heterogeneous implementations. In addition, the host platform of a service may change with time or migrate to different locations. The lack of information of service implementation details makes performance analysis difficult. Cloud federation requires more flexible methods for evaluating performance of Cloud services whose implementation span across multiple data centers. Network-Cloud service convergence calls for holistic approaches to evaluating composite services comprising networking as well as computing systems.

Therefore, developing effective methods for evaluating Cloud service performance becomes a very important research problem that has attracted extensive attention from both academia and industry. Researchers have taken different types of approaches to tackle this challenging problem from various perspectives, for example the perspectives of system modeling, system design, and network protocol. Among the work on Cloud performance evaluation, approaches from the system modeling perspective have formed an active area with much progress recently. However, the numerous published works on cloud performance are scattered in different sections of the literature mainly due to the diverse methodologies that they took. A comprehensive survey that provides a big picture of Cloud performance evaluation from the system modeling perspective and reflects the state of the art of related technologies will be very beneficial to both researchers and practitioners.

Although there are a few existing surveys on Cloud performance evaluation, they focus on a certain aspect of related technologies instead of showing the entire landscape of this field. The systematic literature review on performance evaluation for Cloud services given by Li et al. [47] focused on measurement-based practical methods for evaluating performance of commercial Cloud services. Theoretical approaches to cloud performance analysis and technologies for evaluating private or open-source Cloud service performance are excluded from this survey. Sakellari and Loukas presented a survey on available models, simulation tools, and test beds for research in Cloud computing, including Cloud service performance evaluation [51]. However, queueing theory-based and network calculus-based techniques for Cloud performance analysis are not discussed and performance valuations of inter-Cloud services and composite network-Cloud services are also missed in this survey.

The main objectives of this article is to present the state of the art of available approaches for Cloud service performance evaluation from a system modeling perspective, including both measurement-based methods and analytical modeling-based methods. This article also examines the challenges and open issues for cloud performance evaluation and identifies possible opportunities for future research and technical innovations in this important field. The new contributions made in this article include a holistic view of Cloud performance evaluation approaches, including analytical modeling-based theoretical methods as well as measurement-based practical methods, and a discussion on how these two types of methods may complement each other to develop more effective evaluations for cloud service performance. In addition, the survey on theoretical evaluations given in this article not only provides a comprehensive review covering queueing theory-based methods, network calculus-based methods, and other statistical model-based methods, but also gives a comparison among them in terms of their application circumstances. To the author's best knowledge, such a comprehensive review and insights about various methodologies for Cloud performance evaluation are not available in any existing work.

In this article the related works on Cloud service performance evaluation are classified based on their research methodologies into two main categories: measurement-based methods and analytical modeling-based methods. The goal of this article is to present the most representative works selected through careful literature review in order to fully reflect the state of the art of both categories, rather than being exhaustive to include each single work that has ever been published in the literature. The latest status, challenges, and research opportunities of the two types of approaches are presented respectively in Sections 2 and 3. Then a concluding remark is given in Section 4.

2. Measurement-based evaluation on Cloud service performance

2.1. Representative work on measurement-based Cloud performance evaluation

Cloud infrastructures can only become viable alternatives for traditional enterprise information systems if they provide a proper level of performance. Therefore researchers started evaluating performance of Cloud services as soon as such services became available in order to provide a guideline to potential users for making decisions about Cloud service adoption. Many of the evaluations were based on performance measurements conducted on a Cloud infrastructure as a testbed.

A general procedure for measurement-based service performance evaluation on a Cloud testbed is comprised of the following steps. First, the researchers need to specify the purpose and scope of the evaluation, and then identify the features/aspects of the Cloud services that are to be evaluated. The next step is to determine the performance metrics that will be analyzed and select appropriate benchmarks applications for testing. Then an experimental environment should be setup and testing experiments can be performed.

Stantchev presented a general approach for evaluating nonfunctional QoS properties of individual Cloud services in [52]. This approach is based on an architectural transparent “black box” methodology and is comprised of the following steps: identifying benchmarks, identifying configurations, running tests, analyzing results, and making recommendations. Various performance metrics may be used for evaluating different features of Cloud services. A list of typical performance metrics for evaluating general Cloud services is given in Table 1. A catalogue of metrics for evaluating performance of commercial Cloud services was published in [45], which categorizes metrics into groups for evaluating the communication, computation, memory, and storage aspects of a Cloud computing platform. Generating an appropriate test workload is also an important aspect of measurement-based evaluation methods. A framework for generating and submitting test workloads to evaluating Cloud infrastructure performance was designed in [60].

Table 1
Typical metrics used for evaluating cloud service performance.

Metrics	Description
Service response time (delay)	The latency time between service request and service completion
Service throughput	The number of jobs that can be processed by the service provider in a time unit
Service availability	The probability that a service request can be accepted by the service provider
System utilization	The percentage of system resources that are busy for service provisioning
System resilience	The stability of system performance over time especially under bursty loads
System scalability	The ability of a system to performance well when it is changed in size or volume
System elasticity	The ability of a system to adapt to changes in its loads

Amazon is the first vendor of a wide variety of public Cloud services that have been widely adopted by numerous users. Therefore, many researchers used Amazon Clouds as a testbed for their evaluations on Cloud service performance. An early representative work was reported in [27] where the authors detailed their working experience and performance testing results about Amazon Cloud services including Elastic Computing Cloud (EC2), Simple Storage Service (S3), and Simple Queueing Service (SQS). The authors found that Amazon services offer a practical alternative for organizations interested in storing large amounts of data or making use of Cloud computing. On the other hand, the authors also expressed their concerns about performance consistency and security issues of the available Amazon Cloud services.

High Performance Computing (HPC) for scientific applications typically requires large amounts of computational and storage resources as well as network bandwidth, thus are particularly challenging to Cloud services. Therefore, evaluating performance of commodity commercial Cloud services for supporting high-performance scientific applications attracted much research attention since Cloud services appeared. In [21], the authors tested Amazon EC2 service performance and found that EC2 may support small sized responsive on-demand HPC applications. A quantitative performance analysis of high performance computing on EC2 infrastructure was conducted in [33] and the obtained results show that EC2 offered reasonable service performance for small-scale HPC applications. However, the evaluation reported in [37] indicates that EC2 is slower than a typical mid-range Linux cluster and much slower than a modern HPC system when supporting realistic super-computing applications.

The above studies primarily focused on tightly-coupled HPC applications, for example, Message-Passing Interface (MPI) programs. Workflows are loosely-coupled parallel applications that consist of computational tasks interconnected through data and control dependencies. Performance of the EC2 Cloud service was evaluated from the perspective of scientific workflows in [39] and compared with a typical HPC system. The authors found that although performance of EC2 is not equivalent to a traditional HPC system, it is reasonable given the resources available. In [35], the authors analyzed the performance of Cloud services for Many-Task Computing (MTC) workloads, which are also loosely coupled applications. Performance of four commercial Cloud services, Amazon EC2, GoGrid, ElasticHosts, and Mosso, were measured. The obtained results indicated that the tested Cloud services need an order of magnitude in performance improvement for supporting high-performance MTC scientific applications.

It has been found in the reported research that the poor network performance caused by virtualization I/O overhead, use of commodity interconnection technologies, and processor sharing were the main factors that limit performance and scalability in public commercial Clouds such as Amazon EC2. To overcome these constraints, Amazon added the Cluster Compute (CC) platform into its EC2 Cloud services. The CC platform is a family of instance types that employ more powerful CPUs, dedicated physical node allocation, and high-speed networks for supporting HPC and other demanding network-bound applications. ExpóSito et al. [23] evaluated the performance of two Amazon EC2 Cluster Compute instances – the quadruple extra-large and eight-extra large instances. They found that the scalability of HPC applications on public Cloud infrastructures relies heavily on the performance of communications, which depends on both the network fabric and its efficient support in the virtualization layer.

Cloud services may achieve quite different levels of performance under various workloads generated by diverse applications. Unlike computation and communication-intensive applications, data-intensive applications typically show strong demands for high-performance I/O and storage access in a Cloud infrastructure. Ghoshal et al. [30] compared the I/O performance of Amazon EC2 confronted with Magellan, a private Cloud platform, and an HPC cluster. The obtained results highlighted the overhead and variability of I/O performance in both public and private Cloud solutions and also indicated that NFS performance of regular EC2 instances is many orders of magnitude worse than that of the parallel file system installed in the HPC cluster.

In order to improve I/O performance to support data-intensive applications, Amazon lunched the storage-optimized instance family that includes High I/O quadruple extra-large (HIl) instance and High Storage eight extra-large (HSl) instance. I/O performance of the Amazon CC platform and the High I/O instance was evaluated in [22]. The obtained results revealed that the HIl instance provided significantly better write performance than any other instance type when writing very large files, although the overall performance is ultimate limited by network throughout. Performance evaluation of the Amazon storage-optimized instance family, including both HIl and HSl instances, was reported in [24]. The main experimental results indicated that the unique configurability advantage offered by public Clouds can significantly benefit data-intensive applications.

The aforementioned work mainly targeted high-performance scientific computing. Performance of public Cloud services for business applications, such as supporting Web servers and e-commerce systems, has also been extensively studied. For example, Jiang et al. investigated performance stability and homogeneity of small instances in Amazon EC2 under workloads of typical service-oriented Web applications [12]. They particularly studied the impact of using virtualization in Clouds on these two performance properties and evaluated performance-aware resource provisioning to service-oriented applications in Clouds. In addition to Amazon Web services, Google Application Engine (GAE) and Microsoft Azure have also been widely adopted as commercial commodity Cloud services. Performance metrics of both Amazon and GAE services, including response time and latency for database queries and service delay for data processing, have been measured in [36]. Performance of Azure service for running Windows applications was evaluated in [49].

Elastic service provisioning is one of the distinguishing features of Cloud computing. With a certain level of QoS guarantee, a Cloud data center automatically allocates more resources when the workload increases beyond a certain threshold (scale-up or scale out), and releases unused resources when the load reduces (scale-down or scale-in). The impact of resource scaling on service performance has also been analyzed as an important aspect of Cloud performance evaluation. A representative work on Cloud performance evaluation

Table 2
Comparison of representative measurement-based evaluation methods for cloud service performance.

	Service time	Service throughput	Service availability	Utilization/scalability	Application circumstances
Li [47]	+	+	+		Public IaaS taxonomy
Leitner [44]	+	+	+		Public IaaS performance predictability
Garfinkel [27]	+	+			Amazon EC2, S3, SQS
Jackson [37]	+	+			HPC apps on Amazon IaaS
Hill [33]	+	+			HPC apps on Amazon EC2
ExpóSito [23]	+	+		+	HPC apps on Amazon EC2 with cluster compute
Juve [39]	+				Workflow apps on Amazon EC2
Iosup [35]	+	+		+	Multi-task apps on IaaS
ExpóSito [22]		+		+	I/O intensive apps on Amazon cluster compute
Ghoshal [30]		+			Data-intensive apps on Amazon and Magellan clouds
ExpóSito [24]		+		+	Data-intensive apps on Amazon EC2
Dejun [12]	+			+	Service-oriented apps on Amazon EC2

with elastic scaling strategies was presented by Hwang et al. [34]. In this paper, the authors tested IaaS Cloud services under scale-out and scale-up workloads through benchmark experiments conducted on Amazon EC2. The obtained results indicated that scaling-out instances have much lower overhead than those experienced in scale-up experiments, while scaling up is more cost-effective under a sustained heavier workload.

Evaluation of the performance of Cloud computing platform services may need to handle some special issues, including performance metrics and benchmarks appropriate for PaaS. To address such needs, Ataş and Gungor [2] developed a framework for evaluating PaaS performance and proposed a set of benchmark algorithms that help determine the most appropriate PaaS provider based on different resource needs and application requirements. Commercial PaaS services such as Cloud Foundry, Heroku, and OpenShift, were tested in [2] and the obtained results were analyzed by the authors using two evaluation methods: the Analytical Hierarchy Process (AHP) and Logic Scoring of Preference (LSP).

The rapid developments of Cloud computing technologies together with the success of the Cloud business model have enabled a wide spectrum of Cloud services offered by a large number of vendors. These services have been employed by numerous users for supporting various applications, including business and e-commerce applications in addition to scientific computing, which have highly diverse performance requirements. As a result, numerous works on performance evaluations of various commercial Cloud services in different application scenarios have been reported in the literature recently. However, the area of Cloud service provisioning is relatively chaotic. Different Cloud services are offered with different terminologies, definitions, and features. Even the same provider may supply different kinds of services with similar functionalities but for different purposes. Although exciting progress has been made toward a thorough understanding of the performance attributes of the publicly available Cloud services, obtaining a clear picture of this area becomes more challenging.

In order to address this issue, Li et al. employed a systematic literature review method in [47] to outline the state-of-the-practice of evaluating commercial Cloud services by classifying the published works according to the following categories: evaluation purposes, the evaluated Cloud services, aspects and properties of evaluated Cloud services, the metrics measured for evaluation, benchmarks used for testing, and the experimental environment setup. In order to provide a guideline for implementing different types of measurement experiments for evaluating the numerous and diverse Cloud services, Li and his coauthors of [46] developed a taxonomy of IaaS Cloud service performance evaluation and proposed a conceptual model that generalizes the existing measurement-based performance evaluation practices. Inspired by the systematic review work of Li et al. and Leitner et al. [44] conducted a literature review to collect and codify the existing evaluation results of public commercial IaaS Cloud services with a focus on investigating predictability of Cloud service performance; that is, how accurately the performance of an IaaS Cloud service is estimated in advance and how stable the performance will be.

In addition to using commercial Cloud infrastructures as the testbeds for evaluating service performance, some research Cloud testbeds have also been constructed and utilized for Cloud service performance evaluation. Science Cloud [40] is a Cloud testbed infrastructure constructed by the scientific community for service performance evaluations. The infrastructure is configured with the Nimbus toolkit to enable remote releasing of resources in a similar manner as EC2 services. OpenCirrus [3] is a large scale Cloud testbed composed of federated heterogeneous distributed data centers. It enables researchers to exchange data sets and develop standard Cloud computing benchmarks. Virtual machine management in OpenCirrus is done by different services as long as they are compatible with the EC2 interface. Open Cloud [32] is a research Cloud testbed that is designed to support computations that span more than one data centers. Data centers in Open Cloud are interconnected with a dedicated high-speed wide area network. A more detailed review of currently available research Cloud testbeds is found in [51].

A comparison of the representative works on measurement-based Cloud service performance evaluation, including the evaluated performance metrics and typical application circumstances of these methods, is given in Table 2.

2.2. Challenges and opportunities

We now look at measurement-based Cloud service performance evaluations using public commodity Cloud infrastructures as testbeds. The representative works reviewed in the last section covered the performance of the most popular Cloud services, for example Amazon IaaS clouds, therefore offer useful insights and practical guidelines to both service providers and consumers regarding service performance evaluation. On the other hand, testing experiments conducted using public Cloud infrastructures as testbeds are facing some challenges, which also offer opportunities for future research.

Firstly, such evaluations are only performed using the currently available Cloud services with the configuration settings made by service providers. The SOA principle in Cloud provisioning makes internal implementations of Cloud infrastructures transparent to service consumers. Although such transparency greatly enhances the usability and flexibility of Cloud services, it makes measurement-based performance evaluations more difficult. In most of the experiments reported in the literature, researchers had no control of the commercial Cloud configuration they used as their testbed. Therefore, whether the results obtained from these experiments with certain settings are applicable to more general scenarios of Cloud service provisioning needs further investigation.

Secondly, the lack of knowledge and control of infrastructure configuration for service deployments limits the researchers' ability to study the impact of resource management inside Cloud infrastructures

on service performance through measurement-based evaluations. Service users often want to evaluate the achievable performance of new services to support their decision making a service selection. However, it is difficulty for a service customer to use a measurement-based method to obtain insights about performance behaviors of new Cloud services (or service options) until the customer accepts the service offer and starts using the service.

Also, measurement-based evaluation results obtained from commercial Clouds could become invalid and useless soon after the evaluations. Cloud service providers constantly upgrade hardware and software infrastructures to enhance their current Cloud services as well as add new Cloud service offerings. Decoupling of service capabilities and service implementations in Cloud computing makes such upgrades transparent to service consumers, who may use the same services that are hosted on completely different implementations. For example, some earlier performance evaluations on Amazon EC2 found that it was insufficient for typical scientific computing applications. Then Amazon deployed the CC platform and storage-optimized instance family tailored for high-performance computing in EC2, which made the results obtained from previous tests using regular instances invalid.

Existing works on conceptualization of general performance evaluation [46] and predictability of performance variation [44] offer some promising methods that might help researchers to handle the aforementioned challenging issues. Using research Clouds as the testbeds for measurement-based performance evaluations also allows researchers to obtain information about service implementations and acquire control on configuration of service deployments. However, there are still some challenges that such methods face.

It is typically expensive to construct large scale testbeds that represent realistic Cloud service provisioning scenarios. Regular Cloud users cannot afford to build such a testbed in order to evaluate the performance features of the services they are using or consider adopting. They have to rely on the results published by researchers who conduct evaluations on a Cloud testbed. However, the implementation and service configuration on a research Cloud may not be identical to those of commercial Clouds (and the latter are often unknown to both users and external researchers), which limits the applicability of the results obtained from research Cloud testbeds to general Cloud service scenarios. In addition, it might be difficult or time-consuming to make any major modification in the implementation of a research Cloud after its construction is completed. Therefore, it is challenging to keep a Cloud testbed up-to-data to represent the continuous updates that various service providers keep making in their Cloud infrastructures for both new service offerings and new features of current services.

General measurement-based methods, using either commercial Clouds or research Clouds as testbeds, face some common challenges that offer opportunities for future research. Such methods require extensive and often expensive experimentation and measurements, which must be carefully designed in order to obtain useful results. Although the classification of measurement-based evaluations provided in [46] may serve as a guideline for this purpose, multiple factors including heterogeneity in Cloud implementations, variability in application scenarios, and diversity in users' applications and requirements, etc., still make designing appropriate experiments for performance measurement a challenging problem for future research.

Selection of testing benchmarks plays a key role in measurement-based performance evaluation for Cloud services. Although traditional benchmarks have been recognized as insufficient for evaluating Cloud services [6], they are still predominately used in current evaluation research. Therefore, designing dedicated benchmark for Cloud service evaluation, which may generate realistic workloads that sufficiently reflect the demands of typical Cloud applications, is a challenging open problem that offers opportunities for future research.

Providers offer different services with various capabilities and performance guarantees. On the other hand, research works on performance measurements, even for the same Cloud service, may be conducted in different experimental scenarios. Therefore, another challenge to measurement-based evaluation methods that deserve future investigation is to make the reported results comparable. Developing a standard benchmark for Cloud service testing might help to address this issue.

3. Analytical modeling-based performance evaluation for cloud services

Analytical modeling and analysis techniques have been applied in performance evaluation for Cloud services. Such techniques offer less expensive approaches to analyzing Cloud service performance because they save the cost of testbed experiments required by measurement-based methods. Also, analytical methods may be able to analyze the impact of a large parameter space on service performance even at planning and design stages of new services, which cannot be done easily with a measurement-based method. Therefore, this type of approach has formed another active research area in Cloud service performance evaluation.

The queueing theory, as a classical approach for computer system modeling and analysis, has been widely employed for evaluating Cloud service performance. Network calculus, which is viewed as an extension of traditional queueing theory with mini-plus algebra, has also offered a promising profile-based approach for addressing the challenges to performance evaluation brought in by some special features of Cloud computing. Stochastic Reward Net (SRN), an augmentation of Stochastic Petri Net (SPN), has also been exploited by researchers for modeling Cloud service provisioning and analyzing Cloud service performance.

3.1. Queueing theory-based analysis for Cloud service performance

In [57] the authors modeled a Cloud service provisioning system as a queuing network consisting of two tandem servers with finite buffer space for each server. The first server represents a Web server and the second server models the Cloud service center. Both the inter-arrival time of service requests to the queuing network and the service time at each server are assumed to have an exponential distribution; therefore each server is modeled as a classical $M/M/1$ queue. The percentile of response time to service requests was evaluated as the performance metric in this paper to study the relationship among the maximal number of customers, the minimal service resources, and the highest-level of service performance. However, the proposed model with two tandem $M/M/1$ queues lacks the ability to represent some special features of Cloud computing, for example virtual machines sharing a physical infrastructure that consists of a large number of interconnected servers.

Virtualization is a key feature of Cloud computing that has a significant impact on service performance. Goswami et al. [31] considered the virtualization feature in the model, they developed for Cloud performance analysis. In this model, applications are modeled as queues and the allocated virtual machines are modeled as servers. Both request inter-arrival time and server service time are assumed to have an exponential distribution. Therefore, the model is essentially an $M/M/m/N$ queue with m servers and a finite buffer of size N. To represent the elastic on-demand feature of Cloud service provisioning, the number of servers (allocated virtual machines) in the model can be dynamically adjusted based on the queue length. A steady queue size distribution state was obtained using a recursive method and the service response time was evaluated as the main performance metric in this paper. Liu et al. [48] particularly considered resource sharing among virtual machines in their modeling and analysis on Cloud service performance. Each service request is assumed to comprise multiple subtasks, which will be assigned to different virtual machines that share and contend for the underlying physical infrastructure.

In order to evaluate Cloud service performance guaranteed to applications with different priorities, Ellens et al. [18] developed an $M/M/m/m$ queueing model for Cloud computing centers with multiple priority classes. There are m servers in the model and the total system capacity is m (i.e., no buffer before the servers). All servers are split into two categories: reserved servers that are assigned to process client requests by following priority scheduling, and shared servers that are used to serve the request from any client based on a FIFO policy. The model assumes that the service request arrival process to be a Poisson process and that the service time is exponentially distributed. The rejection probability of clients in each priority class was studied as the main performance metric in this paper.

All the above works assumed exponentially distributed service time when modeling Cloud service systems. Such an assumption, although simplifies the modeling and analysis, does not precisely represent the realistic service feature of Cloud infrastructures. Considering the heterogeneity in implementation technologies for Cloud service provisioning, general distribution would be more appropriate for modeling the service time of a Cloud server. However, the assumption of a general service time distribution may lead to higher analysis complexity. Queuing theory research has shown that solutions to response time and queue length distributions for the $M/G/m$ model and its variations cannot be obtained directly in a closed form, thus requiring suitable approximations in order to achieve tractable results.

An approximate analytical model for Cloud computing centers was developed in [42]. The authors modeled Cloud server farms as an $M/G/m/m+r$ queuing system with single task arrivals and finite task buffer capacity. The arrival process of the task requests are assumed to be a Poisson process and the service policy is FCFS. The performance of such a queuing system was evaluated using a combination of a transformed-based analytical model and an approximate Markov chain model. Probability distributions of service response time and the number of tasks in the system were obtained in the paper. The authors also discussed the immediate service probability and blocking probability, and determined the buffer size required for keeping the blocking probability below a predefined level.

In order to represent the bursty arrival workloads of Cloud infrastructures in service performance evaluation. Khazaei et al. modeled Cloud computing centers as an $M^{[x]}/G/m/m+r$ queuing system [41]. In such a model, the arrival service request process is assumed to be a sequence of super-tasks, each of which consists of a burst of tasks. The inter-arrival time of super-tasks is exponentially distributed and the service time of each task in a super-task has a general distribution. The system has m servers and a buffer size of r. A super-task will be rejected if there is not sufficient resources for the whole super-task.

Quantifying Cloud service performance requires appropriate models that cover a vast parameter space. A monolithic model may suffer from intractability and poor scalability due to the large number of parameters. An approach to reducing the complexity of Cloud service performance analysis is to divide the system model into sub-models and then obtain the overall solution by iteration over individual sub-model solutions. In [28], end-to-end Cloud service provisioning is considered to have three main steps: resource provisioning decision, VM provisioning, and run-time execution. Compared to a single one-level monolithic model, this analysis method is more tractable and scalable. Performance analysis based on this model is only applicable to service requests with a single task. However, Cloud users may ask for multiple VMs to handle multi-tasks by submitting a single service request. In addition, the effect of virtualization in Cloud infrastructures was not explicitly reflected in the analysis reported in [28].

In [43], the authors employed the idea of a sub-model based analysis approach but particularly considered some important features of Cloud computing centers, including batch arrival of tasks and resource virtualization. The authors developed sub-models for resource allocation and virtual machine provisioning and then implemented the sub-models using an interactive Continuous Time Markov Chain (CTMC). The sub-models are interactive such that output of one sub-model is the input of the other one and vice versa. The overall solutions for performance metrics such as task blocking probability and total waiting time incurred on user requests were obtained by iteration over individual sub-model solutions.

In [56], the authors combined the system details, such as Physical Machine (PM) occupation/release, PM warm-up/cool-down, PM fail/recover, and job rejection, into a general model in order to consider all service provisioning details having simultaneous impacts in determining the overall QoS. This model captures system behaviors in a general state transition model and considers the inter-influence of these behaviors in deciding the final QoS. Expected service completion time and rejection probability are the QoS metrics this paper focused on. However, the complexity of the model, mainly due to the state explosion issue of the Marchov chains employed in the paper, limits the developed techniques only to effectively evaluate small scale Cloud infrastructures. Comparing this combined model against the sub-model methods proposed in [28,43] indicates the need of a tradeoff between accuracy and complexity in developing analytical models for Cloud service performance evaluation.

3.2. Profile-based evaluation of cloud service performance

Queueing techniques for Cloud service modeling and performance analysis are typically developed for a specific system architecture with certain assumptions about the service implementations. However, virtualization as the key technology for Cloud computing makes services transparent to their implementations. The same service, even offered by the same vendor, may have completely different implementations when offered to different users. For example the Amazon EC2 services offered to different users could be hosted in data centers located at different sites, which may be running different types of servers and networking equipment. The implementation of a service may also vary with time, for example, due to a system upgrade or virtual machine migration. In addition, the SOA principle in the Cloud service provisioning allows end users to utilize Cloud services via the IaaS, PaaS, and SaaS paradigms without any knowledge of service implementations. The resource abstraction and service encapsulation enabled by virtualization and SOA in Cloud service provisioning make any assumption on specific Cloud system architecture and implementation technology invalid for performance evaluations from a user's perspective.

Recently a profile-based evaluation approach was proposed to tackle these challenges. The basic idea of this approach is to base its modeling and analysis on the QoS related information described by the Service Level Agreement (SLA) between the service provider and consumer, rather than assuming any specific service implementations and user workloads. The QoS information included in an SLA for a user to receive any level of performance guarantee typically includes the service capacity that should be offered by the service provider and the maximal workload that the user is allowed to submit. If we develop profiles for the minimum amount of service capacity guaranteed by a service provider and the maximal workload generated by the user, then it is possible to derive some bounds for performance metrics, such as the worst case service delay and the maximum service request backlog, based on such profiles.

The main mathematical tool employed by the profile-based evaluation approach is network calculus [7]. Two key concepts in network calculus are the arrival curve and service curve. Essentially a service curve is a general function of time that gives a lower bound of the service capacity that a server provides to a customer. Similarly an arrival curve in network calculus is a function of time that specifies the maximum amount of workload that a user is allowed to load a server within an arbitrary time interval. The service curve and arrival curve used to respectively model the service capacity offered by a Cloud

service and the workload on the service. Given the arrival and service curves at a server, network calculus provides the techniques for determining the upper bound of delay for any service request and the maximum backlog of service requests at the server.

Network calculus has been applied to analyze performance of various networking systems including service provisioning in the virtualization-based future Internet. Virtualization has been proposed as a key attribute in the future networking to overcome ossification of the current Internet architecture [25]. On the other hand, the SOA principle has also been applied in networking to facilitate flexible service provisioning [50]. Network virtualization together with SOA in networking enables a Network-as-a-Service (NaaS) paradigm. A network calculus-based model was proposed in [13] for analyzing end-to-end network service performance in a network virtualization environment.

Network calculus was first applied in [14] to develop a profile-based model for Cloud service performance analysis. Performance evaluation reported in this paper took a service user's perspective and treated the end-to-end service delivered to a user as the composition of a Cloud service and the network services for accessing the Cloud infrastructure. In this paper, a general service capability profile was developed based on the service curve concept of network calculus. Such a general profile is used for modeling the minimum service capacities guaranteed by both Cloud and network services. A single capability profile for the entire end-to-end service system was obtained through the convolution operation of network calculus. A demand profile was proposed based on the arrival curve concept to describe the arrival workload for the composite network-Cloud service system. Then the maximum end-to-end service delay was determined and evaluated as the main performance metric. The relationship between the delay performance and available network bandwidth and Cloud server capacity was also analyzed in the paper. The obtained results verified the strong impact of available network resources on Cloud service performance.

Network calculus theory was initially developed for analyzing networking systems with a focus on data transmission rather than data processing. Therefore, regular network calculus techniques for analyzing tandem servers assume that the departure process from an upstream server is identical to the arrival process to its immediate downstream server. However, data processing in Cloud infrastructures will transform the data flows, thus violating this assumption. Recent development in network calculus [26] offered a method for dealing with the data transform in service delivery systems consisting of data processing as well as data transmission.

The profile-based model for Cloud service provisioning has been extended first in [15] and then in [16] by employing the scaling function and scaling curve techniques provided in [26]. A single scaling function was added in [15] to model the data transform effect caused by computing process in a Cloud infrastructure. In [16], two scaling functions were added in the model, respectively before and after the Cloud service component, in order to fully represent the data transform effect between the network and Cloud service components. Then the alternative scaled server theorem proved in [26] is employed to switch the scaling curves and network service profiles in the model. Such a switch procedure enables using the convolution theorem of network calculus to obtain the end-to-end service capability profile of the composite network-Cloud service system, thus determining the maximum service delay.

Profile-based evaluation methods employ the service curve concept from network calculus to obtain a general profile of service capability that is agonistic to service implementations. Similarly the arrival curve-based demand profile is able to describe workloads generated by any applications. Therefore, the profile-based performance analysis offers a promising approach to deal with the heterogeneity in service implementations and diversity in applications in Cloud computing environments. In addition, the profile-based analysis method may also naturally support Cloud-federation and network-Cloud composition.

3.3. Other stochastic modeling approaches for Cloud service performance evaluation

In addition to the queueing theory and network calculus theory, some other stochastic modeling approaches have also been exploited by researchers for evaluating Cloud service performance. Among these, the Stochastic Reward Net (SRN) is a typical method that is often applied together with some queueing techniques for evaluating Cloud service performance. Some of the representative related works are summarized in this subsection.

SRNs are essentially augmented Stochastic Petri Nets (SPNs) with the ability of specifying output measures as reward-based functions for evaluating performance of complex systems. In [28], the authors took a sub-model strategy to simplify the complexity of modeling and analysis of large scale IaaS Cloud computing systems. CMTC-based models are developed for the three main steps of service provisioning: resource provisioning, VM provisioning, and run-time execution. Then SRN is employed to develop a monolithic model to represent the interactions among the three steps thus integrating the sub-models together in order to obtain the overall results of Cloud service performance.

Bruneo [8] employed the SRN technique to evaluate Cloud service performance. The system model proposed in this work consists of a queue for arrival jobs, a scheduler for assigning jobs to virtual machines, and a set of physical servers that host virtual machines. Three arrival cases were considered: a homogeneous Poisson process with a rate λ, a Markov Modulated Poisson process, and a bursty arrival process. The service time for a job is assumed to have an exponential distribution. Performance metrics evaluated in this work include system utilization, availability, and response time for service requests. The innovative aspect of the SRN-based model proposed in this work lies in the generic and comprehensive view of a Cloud system. The author attempted to represent Cloud federation in the proposed model by adding an upload queue for redirecting jobs to another Cloud data center. However, the assumption that job redirection occurs only when the local system queue is full limits the model's ability to reflect some realistic Cloud federation scenarios, where job scheduling among Cloud data centers is determined by high-level service orchestration policies as well as low-level resource availability.

SRN-based modeling and analysis methods have also been applied to energy-aware performance analysis, which evaluates Cloud service performance with consideration of the resource allocation and energy efficiency aspects of Cloud computing. For example, an SRN-based model was developed in [9] for evaluating Cloud service performance and energy efficiency under various resource allocation policies. The green IaaS Cloud scenario modeled in [9] consists of a set of physical servers interconnected through a data center network for hosting various VMs for service provisioning. The VMs are consolidated to servers by following some resource allocation policies and the servers may be in either running, idle, or in sleep mode to reduce energy consumption. In order to handle the complexity of the green IaaS Cloud scenario, the authors of [9] took a layered approach to develop the SRN sub-models respectively for the physical layer and the virtual layer. Once the two layers have been represented, then the authors analyze the dependency between the two layers by modeling the resource allocation policies that regulate energy-aware execution of VMs on top of physical servers.

3.4. Evaluation of Cloud service availability

The aforementioned research work about analytical modeling and analysis on cloud service performance focused on delay and throughput-related metrics. That is, the main evaluated performance considered is about how fast a Cloud system may response to a service request and complete the requested service, and how many service requests can be handled by a Cloud system in each second. In addition to such delay/throughput-related performance, service availability is another

Table 3
Comparison of representative analytical modeling-based evaluation methods for cloud service performance.

	Service delay	Service availability	System utilization	System modeling	Application circumstances
Xiong [57]	+			$M/M/1$ queue	Single cloud data center
Goswami [31]	+			$M/M/m/N$ queue	Single cloud data center
Yang [58]	+	+		$M/M/m/m+r$ queue	Single cloud with fault recovery
Liu [48]	+	+		$M/M/1/N$ with server and link failure	single cloud data center
Ellens [18]		+		$M/M/m/m$ queue	single cloud multiple service classes
Khazaei [42]	+	+	+	$M/G/m/m+r$ queue	single cloud data center
Khazaei [41]	+	+	+	$M^{[x]}/G/m/m+r$ queue	single cloud data center
Khazaei [43]	+	+		interactive CTMC	single cloud data center
Xia [56]	+	+		$M/M/1/k$ queue network w/ server failure	single cloud data center
Ghosh [28]	+	+		CTMC sub-models with SRN	single cloud data center
Bruneo [8]	+	+	+	Stochastic Reward Net	single/federated data centers
Bruneo [9]	+	+	+	Stochastic Reward Net	single cloud considering DCN
Yang [59]	+	+		$M/M/m/m+r$ with server fault recovery	single cloud data center
Yang [58]	+	+		$GI^X/M/S/N$ w/ server and link fault recovery	single cloud data center
Bilal [5]		+		multi-layer graph model	cloud data center networks
Duan [14]	+			network calculus	cloud and network services
Duan [16]	+			network calculus w/ scaling functions	cloud and network services

crucial aspect of QoS that must be guaranteed by various Cloud services for meeting user requirements. Therefore, analysis on service availability is also considered as an important part of Cloud performance evaluation.

Availability of Cloud services may be quantified using the rejection probability for a service request, that is, the probability that the request for a service cannot be accepted by the Cloud service provider. The rejection probability is related to multiple factors such as the total system capacities (including both server capacity, buffer space, network bandwidth etc.), Cloud service management mechanisms (e.g., the job scheduling policies employed by the Cloud data center), and characteristics of traffic load arrival at the Cloud system.

A typical approach to evaluating service availability is to develop a queueing model and then determine the rejection probability for given system setting and work load. The relationship between service request rejection probability and various system parameters has been analyzed by researchers in their works on Cloud performance evaluation using various queueing models. For example, Ellens et al. determined the rejection probabilities for multiple priority queues in [18]. Khazaei et al. analyzed the rejection probability with different amounts of buffer space available in Cloud servers [42] and evaluated task blocking probability based on the CTMC model they developed in [43]. Service availability may also be evaluated by using other analytical models. For example, Bruneo conducted resilience analysis using an SRN-based model developed in [8] to determine the steady-state probability that a Cloud system is able to accept a service request.

In addition to the limit of system resources, another key impact factor on Cloud service availability is caused by system failures and errors. Fault tolerance designs are often employed for Cloud computing infrastructures, which on one hand enhances Cloud service availability, on the other hand may introduce extra delay and degrade system throughput. Therefore, the impact of high availability designs with fault tolerance mechanisms on the performance of Cloud services becomes an important research topic.

Yang et al. [59] conducted research on Cloud performance evaluation considering fault tolerance and particularly studied the impact of fault recovery on Cloud service performance. An $M/M/m/m+r$ queueing system was proposed in [59] to model a Cloud data center with fault recovery. Both inter-arrival and service times are assumed to be exponentially distributed. The system has m servers and a finite buffer of size r, thus having a total capacity $m+r$. The distribution of response time was obtained as the main performance metric in this paper through dividing response time into waiting, service, execution periods, and assuming independence among them.

The work reported in [59] has been extended in [58] in order to fully reflect practical and realistic operations in typical Cloud data centers. In [58], a $GI^X/M/S/N$ model, i.e., a multi-server queueing system with general inter-arrival time distribution, exponential service time, finite server and buffer capacities, and batch arrival, was developed for evaluating Cloud service performance considering fault recovery. Cloud service performance is quantified by the service response time, whose probability distribution is derived considering fault recovery on nodes that process subtasks and on communication links.

Ghosh et al. proposed a comprehensive model for Cloud service provisioning in [29] to capture various details of Cloud computing, including fault/error recovery and job rejection as well as servers occupation/release and machine warm up/cool down. In order to handle the complexity introduced by the various aspects of details for Cloud service provisioning, the model was decomposed into multiple sub-models and QoS results obtained from the sub-models are integrated to achieve the overall results. In order to study the simultaneous impacts of different aspects of service provisioning, including fault/error recovery, on service performance, Chen et al. [10] proposed a general queueing network-based model that may determine the expected task completion time and service rejection probability considering fault recovery as well as other factors such as system capacity and workload.

Data Center Network (DCN) is an important ingredient of a Cloud computing infrastructure; therefore, its resilience may greatly influence Cloud service availability. A representative investigation on DCN resilience was reported by Bilal et al. in [5]. The authors of [5] developed multilayered graph models for analyzing robustness of typical DCNs architectures and found that classical network robustness metrics are inadequate to appropriately evaluate DCN robustness; then they proposed new procedures to quantify the DCN robustness for addressing this issue.

3.5. Comparison of analytical modeling-based methods for Cloud service performance evaluation

A comparison of the above reviewed analytical modeling-based evaluation methods, including the evaluated performance metrics, the employed system models, and the typical application circumstances of the methods, is given in Table 3.

3.6. Challenges and opportunities

Special features of the Cloud computing paradigm bring in new challenges to system modeling and service performance analysis. In order to accurately represent a Cloud system, an analytical model needs to be scalable to deal with the large amount of resources in Cloud

infrastructures and be flexible to handle different implementations and configurations of Cloud services. Diversity in Cloud service functions, heterogeneity in Cloud implementations, and resource virtualization in Cloud service provisioning are some particular challenging issues that must be fully considered when developing analytical approaches to evaluating Cloud service performance.

Due to the highly diverse Cloud services and applications, precisely characterizing Cloud workloads is important but very challenging step in order to analyze Cloud service performance. A majority of the research on analytical evaluation methods reported in the literature assumed the service request arrival process to be a Poisson process having exponentially distributed inter-arrival time. However, it has been found that large scale distributed systems with a large number of users, such as Cloud data centers, could exhibit self-similarity and long-range dependence with respect to the arrival process [54]. In addition, the wide variety of applications using different Cloud services may generate workloads with different patterns. Therefore, developing appropriate and flexible models for Cloud service workloads that precisely represent the traffic generated from a wide variety of applications is an important open problem for future research.

Another important element of analytical methods for Cloud service performance evaluation is to model the service time of Cloud systems. Exponential distribution has been assumed for service time in many works for simplifying the analysis. However, such an assumption may not precisely represent the actual system behaviors due to the diverse implementations and configurations of Cloud infrastructures. Other researchers assumed the service time to be a random variable with an arbitrary distribution in order to obtain a general model for servers in Clouds. However, such an assumption increases analysis complexity and limits the tractability and scalability of the developed analysis techniques.

In addition to modeling the service time, constructing an appropriate queueing network model to reflect the realistic systems in Cloud infrastructures for service provisioning is also a challenging problem that needs further investigation. Most of the published works modeled Cloud systems as a set of servers that represent physical and/or virtual machines. In a real Cloud infrastructure, a large number of servers are interconnected through a high-speed data center network. Research results have indicated that networking systems in Cloud infrastructures have a significant impact on service performance and may form a bottleneck for supporting high-performance applications [37,55]. However, few of the currently available models fully considered the impact of networks in data centers on Cloud service performance.

All Cloud services are delivered to their end users through networks. Public commercial Cloud services are typically accessed by customers via the Internet and private/hybrid Cloud services need to be accessed through enterprise networks. Therefore, what end users actually perceive is the performance of an end-to-end service offered by both the Cloud infrastructure and the network that provides service access. Therefore, performance analysis from an end user's perspective should consider the combined performance of the Cloud service and the network through which the user accesses the Cloud service. In addition, with the rapid development of Cloud federation, the end-to-end service provisioned to an end user will become a composite service comprising multiple Cloud services and multiple network services. The currently available analytical evaluation methods for Cloud service performance still lack the ability to fully analyze such composite service provisioning scenarios; therefore offering opportunities for future research.

Although profile-based modeling and performance analysis offer a flexible approach to evaluating various Cloud services, there are some challenging technical problems related to this type of methods that need thorough study in future research.

Obtaining a precise service capability profile for the studied Cloud infrastructure and a demand profile that fully represents Cloud workload is a challenging issue. Current work employed the Latency-Rate (LR) server model [53] for Cloud infrastructures and the leaky-bucket arrival curve [11] for arrival processes to Cloud services. Although both LR model for service profiles and leaky-bucket model for demand profiles enable tractable analysis for typical service scenarios, it is desirable to have more precise profiles that are able to characterize Cloud service capacities and workloads in more detail. Some of the measurement-based performance evaluation methods reviewed in Section 2 are used to obtain testing data of service capacities and workloads of Cloud infrastructures, based on which one could develop the service profiles and demand profiles for Cloud services. In this sense, profile-based evaluation could be thought of as a combination of experimental measurement and analytical analysis. However, there is a tradeoff between the complexity of analysis technique and the precision of service and demand profiles. Piece-wise linear service and demand profiles seem to provide a good balance between these two aspects; thus offering an interesting topic that deserves further investigation.

The current profile-based Cloud service modeling and analysis methods are based on deterministic network calculus, which gives the worst-case performance metrics, for example the maximum service delay time. When applied to guide resource management for Cloud service provisioning, the relationship between service capacity and performance obtained from such a deterministic analysis may result in low resource utilization. Actually many applications only expect statistical performance guarantee from Cloud service providers; that is, the applications work well as long as the probability of missing performance expectation is limited at a certain level. Therefore, applying statistic network calculus [38] into profile-based Cloud service performance evaluation would be an interesting and important topic for future research.

With rapid development of the Cloud federation and the emergence of network-Cloud service convergence, end-to-end service provisioning in the future Cloud environments will often traverse multiple heterogeneous networking and computing domains. The currently available profile-based models for Cloud service performance analysis, although considered network and Cloud service composition, are still limited to scenarios with only a single Cloud infrastructure. Further development of the profile-based evaluation approach in order to fully support the Cloud federation offers an interesting topic for future research.

4. Conclusion

Cloud services have become indispensable ingredients of the future information infrastructures. Evaluation of Cloud service performance is crucial and beneficial to both service providers and service consumers. Numerous related works have been published in different sections of the literature. This article gives a comprehensive survey on performance evaluation of Cloud services from the system modeling perspective in order to reflect the latest status of this important area. In this survey, the currently available approaches to evaluating Cloud service performance are classified based on their research methodologies into two categories: measurement-based approaches and analytical modeling-based approaches; and the latter one comprises queueing theory-based, network calculus-based, and other stochastic models such as SRN-based methods. For both categories, the state of the art of related technologies is first reviewed, and then the open issues, challenges, and opportunities for future research are discussed. Each type of method has its own advantages and disadvantages when applied to evaluating Cloud service performance. Table 4 gives a brief summary of such advantages and disadvantages. The survey presented in this article indicates that although exciting progress has been made toward thorough understanding about the performance behaviors of various Cloud services, there are still a wide spectrum of open problems in each category that need further investigation; thus offering interesting topics for future research.

Table 4
Advantages and disadvantages of typical types of approaches for evaluating cloud service performance.

	Advantages	Disadvantages
Measurement-Based methods	Provide performance insights about available Cloud services, reflect service performance in realistic operation scenarios, indicate service performance for practical applications	Limited to available Cloud testbeds and their settings, expensive for conducting testing experiments, cannot predict performance for new services and/or settings
Queueing theory-based methods (may be applied with SRN)	No experiment cost; can predict performance of new services and settings, may evaluate impacts of a large set of parameters	Hard to model heterogeneous Cloud infrastructures, hard to model traffic loads of diverse applications, may not reflect performance of realistic service scenarios
Profile-based methods	Agnostic to heterogenous Cloud service implementations, applicable to diverse application loads, reflect virtualization and abstraction features	effectiveness relies on precision of service and demand profiles, current analysis only for worst-case performance

References

[1] M. Armbrust, O. Fox, R. Griffith, A.D. Joseph, Y. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, Above the Clouds: A Berkeley View of Cloud Computing, Tecnical Report - University of California Berkeley, 2009.

[2] G. Ataş, V.C. Gungor, Performance evaluation of cloud computing platforms using statistical methods, Comput. Electr. Eng. 40 (5) (2014) 1636–1649.

[3] A.I. Avetisyan, R. Campbell, K. Lai, M. Lyons, D.S. Milojicic, H.Y. Lee, Y.C. Soh, N.K. Ming, J.-Y. Luke, H. Namgoong, et al., Open cirrus: a global cloud computing testbed, IEEE Comput. 43 (4) (2010) 35–43.

[4] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, A. Warfield, Xen and the art of virtualization, ACM SIGOPS Oper. Syst. Rev. 37 (5) (2003) 164–177.

[5] K. Bilal, M. Manzano, S.U. Khan, E. Calle, K. Li, A.Y. Zomaya, On the characterization of the structural robustness of data center networks, IEEE Trans. Cloud Comput. 1 (1) (2013) 64–771.

[6] C. Binnig, D. Kossmann, T. Kraska, S. Loesing, How is the weather tomorrow?: towards a benchmark for the cloud, in: Proceedings of the Second ACM International Workshop on Testing Database Systems, 2009, p. 9.

[7] J.L. Boudec, P. Thiran, Network Calculus: A Theory of Deterministic Queueing Systems for the Internet, Springer Verlag, Berlin, 2001.

[8] D. Bruneo, A stochastic model to investigate data center performance and QoS in IaaS Cloud computing systems, IEEE Trans. Parallel Distrib. Syst. 25 (3) (2014) 560–569.

[9] D. Bruneo, A. Lhoas, F. Longo, A. Puliafito, Modeling and evaluation of energy policies in green clouds, IEEE Trans. Parallel Distrib. Syst. 26 (11) (2015) 3052–3065.

[10] P. Chen, Y. Xia, S. Pang, J. Li, A probabilistic model for performance analysis of cloud infrastructures, Concurr. Comput.: Pract. Exp. 27 (17) (2015) 4784–4796.

[11] R.L. Cruz, A calculus for network delay. I. Network elements in isolation, IEEE Trans. Inf. Theory 37 (1) (1991) 114–131.

[12] J. Dejun, G. Pierre, C.-H. Chi, EC2 performance analysis for resource provisioning of service-oriented applications, in: Proceedings of the 11th International Conference on Service Oriented Computing, 2009, pp. 191–207.

[13] Q. Duan, Modeling and analysis of end-to-end quality of service provisioning in virtualization-based future Internet, in: Proceedings of the 19th IEEE International Conference on Computer Communications and Networks (ICCCN2010), 2010, pp. 1–6.

[14] Q. Duan, Modeling and performance analysis on network virtualization for composite network-Cloud service provisioning, in: Proceedings of the 2011 IEEE World Congress on Services (SERVICES2011), 2011, pp. 548–555.

[15] Q. Duan, Modeling and delay analysis for converged network-cloud service provisioning systems, in: Proceedings of the IEEE 2013 International Conference on Computing, Networking and Communications, 2013, pp. 66–70.

[16] Q. Duan, Modeling and performance analysis on composite network-compute service provisioning in software-defined cloud environments, Elsevier Digit. Commun. Netw. J. 1 (3) (2015) 181–190.

[17] Q. Duan, Y. Yan, A.V. Vasilakos, A survey on service-oriented network virtualization toward a convergence of networking and Cloud computing, IEEE Trans. Netw. Serv. Manag. 9 (4) (2012) 373–392.

[18] W. Ellens, M. Zivkovic, J. Akkerboom, R. Litijens, H. Berg, Performance of cloud computing centers with multiple priority classes, in: Proceedings of the 2012 IEEE International Conference on Cloud Computing, 2012, pp. 245–252.

[19] T. Erl, Service-Oriented Architecture – Concepts, Technology, and Design, Prentice Hall, Boston, 2005.

[20] ETSI, Network Functions Virtualization – Introductory White Paper, October 2012.

[21] C. Evangelinos, C. Hill, Cloud computing for parallel scientific hpc applications: feasibility of running coupled atmosphere-ocean climate models on Amazons EC2, in: Proceedings of the 2008 Cloud Computing and its Applications Conference (CCA-08), 2008.

[22] R.R. Expósito, G.L. Taboada, S. Ramos, J. González-Domínguez, J. Touriño, R. Doallo, Analysis of I/O performance on an Amazon EC2 cluster compute and high I/O platform, J. Grid Comput. 11 (4) (2013) 613–631.

[23] R.R. ExpóSito, G.L. Taboada, S. Ramos, J. TouriñO, R. Doallo, Performance analysis of HPC applications in the Cloud, Future Gener. Comput. Syst. 29 (1) (2013) 218–229.

[24] R.R. Expósito, G.L. Taboada, S. Ramos, J. Touriño, R. Doallo, Performance evaluation of data-intensive computing applications on a public IaaS cloud, Comput. J., 2014.

[25] N. Feamster, L. Gao, J. Rexford, How to lease the Internet in your spare time, ACM SIGCOMM Comput. Commun. Rev. 37 (1) (2007) 61–64.

[26] M. Fidler, J.B. Schmitt, On the way to a distributed systems calculus: an end-to-end network calculus with data scaling, in: ACM SIGMETRICS Performance Evaluation Review, vol. 34, 2006, pp. 287–298.

[27] S.L. Garfinkel, An evaluation of Amazons Grid computing services: EC2, S3, and SQS, Tech. rep., Center for Research on Computation and Society, Harvard University, 2007.

[28] R. Ghosh, F. Longo, V.K. Naik, K.S. Trivedi, Modeling and performance analysis of large scale IaaS Clouds, Future Gener. Comput. Syst. 29 (5) (2013) 1216–1234.

[29] R. Ghosh, K.S. Trivedi, V.K. Naik, D.S. Kim, End-to-end performability analysis for infrastructure-as-a-service cloud: an interacting stochastic models approach, in: Proceedings of the 2010 IEEE 16th Pacific Rim International Symposium on Dependable Computing (PRDC2010), 2010, pp. 125–132.

[30] D. Ghoshal, R.S. Canon, L. Ramakrishnan, I/O performance of virtualized cloud environments, in: Proceedings of the 2nd ACM International Workshop on Data Intensive Computing in the Clouds, 2011, pp. 71–80.

[31] V. Goswami, S.S. Patra, G.B. Mund, Performance analysis of cloud with queue-dependent virtual machines, in: Proceedings of the 1st International Conference on Recent Advances in Information Technology, 2012.

[32] R. Grossman, Y. Gu, M. Sabala, C. Bennet, J. Seidman, J. Mambratti, The open Cloud testbed: A Wide Area Testbed for Cloud Computing Utilizing High Performance Network Services, arXiv preprint arXiv:0907.4810, 2009.

[33] Z. Hill, M. Humphrey, A quantitative analysis of high performance computing with Amazon's EC2 infrastructure: the death of the local cluster?, in: Proceedings of the 10th IEEE/ACM International Conference on Grid Computing, 2009, pp. 26–33.

[34] K. Hwang, X. Bai, Y. Shi, M. Li, W.-G. Chen, Y. Wu, Cloud performance modeling with benchmark evaluation of elastic scaling strategies, IEEE Trans. Parallel Distrib. Syst. 27 (1) (2016) 130–143.

[35] A. Iosup, S. Ostemann, M. Yigitbasi, R. Prodan, T. Fahringer, D. Epema, Performance analysis of Cloud computing services for many-tasks scientific computing, IEEE Trans. Parallel Distrib. Syst. 22 (6) (2011) 931–945.

[36] A. Iosup, N. Yigitbasi, D. Epema, On the performance variability of production cloud services, in: Proceedings of the 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, May 2011, pp. 104–113.

[37] K.R. Jackson, K. Muriki, S. Canon, S. Cholia, J. Shalf, Performance analysis of high performance computing applications on the Amazon Web services cloud, in: Proceedings of the 2010 IEEE International Conference on Cloud Computing Technology and Science, 2010, pp. 159–168.

[38] Y. Jiang, Y. Liu, Stochastic Network Calculus 1, Springer, London, 2008.

[39] G. Juve, E. Deelman, K. Vahi, G. Mehta, b.P. Berman, B. Berriman, P. Maechling, Scientific workflow applications on Amazon EC2, in: Proceedings of the 5th IEEE International Conference on e-Science, 2009.

[40] K. Keahey, R. Figueiredo, J. Fortes, T. Freeman, M. Tsugawa, Science clouds: early experiences in cloud computing for scientific applications, Cloud Computing and Applications, 2008, 2008, pp. 825–830.

[41] H. Khazaei, J. Misic, v.B. Misic, Performance analysis of cloud computing centers under burst arrivals and total reject policy, in: Proceedings of the 2011 IEEE Global Communications Conference, 2011.

[42] H. Khazaei, J. Misic, v.B. Misic, Performance analysis of Cloud computing centers using M/G/m/m+r queuing systems, IEEE Trans. Parallel Distrib. Syst. 23 (5) (2012) 936–943.

[43] H. Khazaei, J. Misic, v.B. Misic, A fine-grained performance model of cloud computing centers, IEEE Trans. Parallel Distrib. Syst. 24 (11) (2013) 2138–2147.

[44] P. Leitner, J. Cito, Patterns in the chaosa study of performance variation and predictability in public iaas clouds, ACM Trans. Internet Technol. (TOIT) 16 (3) (2016) 15.

[45] Z. Li, L. O'Brien, H. Zhang, R. Cai, On a catalogue of metrics for evaluating commercial cloud services, in: Proceedings of the 2012 ACM/IEEE 13th International Conference on Grid Computing, 2012, pp. 164–173.

[46] Z. Li, L. O'Brien, H. Zhang, R. Cai, On the conceptualization of performance evaluation of IaaS services, IEEE Trans. Serv. Comput. 7 (4) (2014) 628–641.

[47] Z. Li, H. Zhang, L. O'Brien, R. Cai, S. Flint, On evaluating commercial Cloud services: a systematic review, J. Syst. Softw. 86 (9) (2013) 2371–2393.

[48] X. Liu, W. Tong, X. Zhi, F. ZhiRen, L. WenZhao, Performance analysis of cloud computing services considering resources sharing among virtual machines, J.

Supercomput. 69 (2014) 357–374.
[49] W. Lu, J. Jackson, R. Barga, Azureblast: a case study of developing science applications on the cloud, in: Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, June 2010, pp. 413–420.
[50] T. Magedanz, N. Blum, S. Dutkowski, Evolution of SOA concepts in telecommunications, IEEE Comput. Mag. 40 (11) (2007) 46–50.
[51] G. Sakellari, G. Loukas, A survey of mathematical models, simulation approaches and testbeds used for research in Cloud computing, Simul. Model. Pract. Theory 39 (2013) 92–103.
[52] V. Stantchev, Performance evaluation of cloud computing offerings, in: Proceedings of the 3rd IEEE International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP'09), 2009, pp. 187–192.
[53] D. Stiliadis, A. Varma, Latency-rate servers: a general model for analysis of traffic scheduling algorithms, IEEE/ACM Trans. Netw. (ToN) 6 (5) (1998) 611–624.
[54] A. Verma, G. Dasgupta, T.K. Nayak, P. De, R. Kothari, Server workload analysis for power minimization using consolidation, in: Proceedings of the 2009 conference on USENIX Annual technical conference, USENIX Association, 2009, pp. 28–28.
[55] G. Wang, T.S.E. Ng, The impact of virtualization on network performance of Amazon EC2 data center, in: Proceedings of the IEEE INFOCOM 2010, 2010, pp. 1–9.
[56] Y. Xia, M. Zhou, X. Luo, Q. Zhu, J. Li, Y. Huang, Stochastic modeling and quality evaluation of infrastructure-as-a-service Clouds, IEEE Trans. Autom. Sci. Eng. 12 (1) (2015) 162–170.
[57] K. Xiong, H. Perros, Service performance and analysis in Cloud computing, in: Proceedings of the IEEE 2009 World Congress on Services, 2009, pp. 693–700.
[58] B. Yang, F. Tan, Y.-S. Dai, Performance evaluation of Cloud service considering fault recovery, J. Supercomput. 65 (1) (2013) 426–444.
[59] B. Yang, F. Tan, Y.-S. Dai, S. Guo, Performance evaluation of cloud service considering fault recovery, in: Proceedings of the 2009 IEEE International Conference on Cloud Computing, 2009, pp. 571–576.
[60] N. Yigitbasi, A. Iosup, D. Epema, S. Ostermann, C-meter: a framework for performance analysis of computing Clouds, in: Proceedings of the 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, 2009, pp. 472–477.