The Journal of Systems and Software 86 (2013) 2371–2393

Contents lists available at SciVerse ScienceDirect

The Journal of Systems and Software

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

On evaluating commercial Cloud services: A systematic review

Zheng Li [a], He Zhang [b,*], Liam O'Brien [c], Rainbow Cai [e], Shayne Flint [d]

[a] NICTA, School of Computer Science, Australian National University, Canberra, Australia
[b] State Key Laboratory of Novel Software Technology, Software Institute, Nanjing University, Jiangsu, China
[c] Geoscience Australia, Canberra, Australia
[d] School of Computer Science, Australian National University, Canberra, Australia
[e] Division of Information, Australian National University, Canberra, Australia

ARTICLE INFO

Article history:
Received 30 September 2012
Received in revised form 15 February 2013
Accepted 6 April 2013
Available online 26 April 2013

Keywords:
Cloud Computing
Cloud service evaluation
Systematic literature review

ABSTRACT

Background: Cloud Computing is increasingly booming in industry with many competing providers and services. Accordingly, evaluation of commercial Cloud services is necessary. However, the existing evaluation studies are relatively chaotic. There exists tremendous confusion and gap between practices and theory about Cloud services evaluation.
Aim: To facilitate relieving the aforementioned chaos, this work aims to synthesize the existing evaluation implementations to outline the state-of-the-practice and also identify research opportunities in Cloud services evaluation.
Method: Based on a conceptual evaluation model comprising six steps, the systematic literature review (SLR) method was employed to collect relevant evidence to investigate the Cloud services evaluation step by step.
Results: This SLR identified 82 relevant evaluation studies. The overall data collected from these studies essentially depicts the current practical landscape of implementing Cloud services evaluation, and in turn can be reused to facilitate future evaluation work.
Conclusions: Evaluation of commercial Cloud services has become a world-wide research topic. Some of the findings of this SLR identify several research gaps in the area of Cloud services evaluation (e.g., Elasticity and Security evaluation of commercial Cloud services could be a long-term challenge), while some other findings suggest the trend of applying commercial Cloud services (e.g., compared with PaaS, IaaS seems more suitable for customers and is particularly important in industry). This SLR study itself also confirms some previous experiences and records new evidence-based software engineering (EBSE) lessons.

1. Introduction

By allowing customers to access computing services without owning computing infrastructures, Cloud Computing has emerged as one of the most promising computing paradigms in industry (Buyya et al., 2009). Correspondingly, there are more and more commercial Cloud services supplied by an increasing number of providers available in the market (Prodan and Ostermann, 2009)[LYKZ10].[1] Since different and competitive Cloud services may be offered with different terminologies, definitions, and goals (Prodan and Ostermann, 2009), Cloud services evaluation would be crucial and beneficial for both service customers (e.g., cost–benefit analysis) and providers (e.g., direction of improvement) [LYKZ10].

However, the evaluation of commercial Cloud services is inevitably challenging for two main reasons. Firstly, previous evaluation results may become quickly out of date. Cloud providers may continually upgrade their hardware and software infrastructures, and new commercial Cloud services and technologies may gradually enter the market. For example, at the time of writing, Amazon is still acquiring additional sites for Cloud data centre expansion (Miller, 2011); Google is moving its App Engine service from CPU usage model to instance model (Alesandre, 2011); while IBM just offered a public and commercial Cloud (Harris, 2011). As a result, customers would have to continuously re-design and repeat evaluation for employing commercial Cloud services.

Secondly, the back-ends (e.g., configurations of physical infrastructure) of commercial Cloud services are uncontrollable (often invisible) from the perspective of customers. Unlike consumer-owned computing systems, customers have little knowledge or

* Corresponding author. Tel.: +86 25 83621369; fax: +86 25 83621370.
E-mail addresses: zheng.li@nicta.com.au (Z. Li), dr.hezhang@gmail.com (H. Zhang), liamob99@hotmail.com (L. O'Brien), shayne.flint@anu.edu.au (S. Flint).

[1] We use two types of bibliography formats: the alphabetic format denotes the Cloud service evaluation studies (primary studies) of the SLR, while the name-year format (present in the "References" section) refers to the other references for this article.

0164-1212/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jss.2013.04.021

control over the precise nature of Cloud services even in the "locked down" environment [SSS+08]. Evaluations in the context of public Cloud Computing are then inevitably more challenging than that for systems where the customer is in direct control of all aspects [Sta09]. In fact, it is natural that the evaluation of uncontrollable systems would be more complex than that of controllable ones.

Meanwhile, the existing Cloud services evaluation research is relatively chaotic. On one hand, the Cloud can be viewed from various perspectives (Stokes, 2011), which may result in market hype and also skepticism and confusion (Zhang et al., 2010). As such, it is hard to point out the range of Cloud Computing and a full scope of metrics to evaluate different commercial Cloud services. On the other hand, there exists a tremendous gap between practice and research about Cloud services evaluation. For example, although the traditional benchmarks have been recognized as being insufficient for evaluating commercial Cloud services [BKKL09], they are still predominately used in practice for Cloud services evaluation.

To facilitate relieving the aforementioned research chaos, it is necessary for researchers and practitioners to understand the state-of-the-practice of commercial Cloud services evaluation. For example, the existing evaluation implementations can be viewed as primary evidence for adjusting research directions or summarizing feasible evaluation guidelines. As the main methodology applied for evidence-based software engineering (EBSE) (Dybå et al., 2005), the Systematic Literature Review (SLR) has been widely accepted as a standard and rigorous approach to evidence aggregation for investigating specific research questions (Kitchenham and Charters, 2007; Zhang and Babar, 2011). Naturally, we adopted the SLR method to identify, assess and synthesize the relevant primary studies to investigate Cloud services evaluation. In fact, according to the popular aims of implementing a systematic review (Lisboa et al., 2010), the results of this SLR can help identify gaps in current research and also provide a solid background for future research activities in the field of Cloud services evaluation.

This paper outlines the work involved in conducting this SLR on evaluating commercial Cloud services. Benefitting from this SLR, we confirm the conceptual model of Cloud services evaluation; the state-of-the-practice of the Cloud services evaluation is finally revealed; and several findings are highlighted as suggestions for future Cloud services evaluation work. In addition to the SLR results, the lessons learned from performing this SLR are also reported in the end. By observing the detailed implementation of this SLR, we confirm some suggestions supplied by the previous SLR studies, and also summarize our own experiences that could be helpful in the community of EBSE (Dybå et al., 2005). In particular, to distinguish and elaborate some specific findings, three parts (namely evaluation taxonomy (Li et al., 2012a), metrics (Li et al., 2012c), and factors (Li et al., 2012b)) of the outcome derived from this SLR have been reported separately. To avoid duplication, the previously reported results are only briefly summarized (cf. Section 5.3, 5.4, and 5.6) in this paper.

The remainder of this paper is organized as follows. Section 2 supplements the background of this SLR, which introduces a spatial perspective as prerequisite to investigating Cloud services evaluations. Section 3 elaborates the SLR method and procedure employed in this study. Section 4 briefly describes the SLR results, while Section 5 answers the predefined research questions and highlights the findings. Section 6 discusses our own experiences in using the SLR method, while Section 7 shows some limitations with this study. Conclusions and some future work are discussed in Section 8.

2. Related work and a conceptual model of Cloud services evaluation

Evaluation of commercial Cloud services emerged as soon as those services were published [Gar07b,HLM+10]. In fact, Cloud services evaluation has rapidly and increasingly become a worldwide research topic during recent years. As a result, numerous research results have been published, covering various aspects of Cloud services evaluation. Although it is impossible to enumerate all the existing evaluation-related studies, we can roughly distinguish between different studies according to different evaluation aspects on which they mainly focused. Note that, since we are interested in the practices of Cloud services evaluation, *Experiment-Intensive Studies* are the main review objects in this SLR. Based on the rough differentiation, the general process of Cloud services evaluation can be approximately summarized and profiled using a conceptual model.

2.1. Different studies of Cloud services evaluation

Service feature-emphasized studies:

Since Cloud services are concrete representations of the Cloud Computing paradigm, the Cloud service features to be evaluated have been discussed mainly over Cloud Computing-related introductions, surveys, or research agendas. For example, the characteristics and relationships of Clouds and related technologies were clarified in Buyya et al. (2009), Foster et al. (2008), and Zhang et al. (2010), which hinted the features that commercial Cloud services may generally embrace. The authors portrayed the landscape of Cloud Computing with regard to trust and reputation (Habib et al., 2010). Most of the studies (Armbrust et al., 2010; Buyya et al., 2009; Rimal et al., 2009; Zhang et al., 2010) also summarized and compared detailed features of typical Cloud services in the current market. In particular, the Berkeley view of Cloud Computing (Armbrust et al., 2010) emphasized the economics when employing Cloud services.

Metrics-emphasized studies:

When evaluating Cloud services, a set of suitable measurement criteria or metrics must be chosen. As such, every single evaluation study inevitably mentions particular metrics when reporting the evaluation process and/or result. However, we did not find any systematic discussion about metrics for evaluating Cloud services. Considering that the selection of metrics plays an essential role in evaluation implementations (Obaidat and Boudriga, 2010), we performed a comprehensive investigation into evaluation metrics in the Cloud Computing domain based on this SLR. The investigation result has been published in Li et al. (2012c). To the best of our knowledge, this is the only metrics-intensive study of Cloud services evaluation.

Benchmark-emphasized studies: Although traditional benchmarks have been widely employed for evaluating commercial Cloud services, there are concerns that traditional benchmarks may not be sufficient to meet the idiosyncratic characteristics of Cloud Computing. Correspondingly, the authors theoretically portrayed what an ideal Cloud benchmark should be [BKKL09]. In fact, several new Cloud benchmarks have been developed, for example Yahoo! Cloud Serving Benchmark (YCSB) [CST+10] and CloudStone [SSS+08]. In particular, six types of emerging scale-out workloads were collected to construct a benchmark suite, namely CloudSuite (Ferdman et al., 2012), to represent today's dominant Cloud-based applications, such as Data Serving, MapReduce, Media Streaming, SAT Solver, Web Frontend, and Web Search.

Experiment-emphasized studies:

To reveal the rapidly changing and customer-uncontrollable nature of commercial Cloud services, evaluations have to be implemented through practical experiments. In detail, an evaluation experiment is composed of experimental environment and experimental manipulation. If only focusing on the Cloud side, experimental environment indicates the involved Cloud resources like amount [Sta09] or location [DPhC09] of service instances, while experimental manipulation refers to the necessary

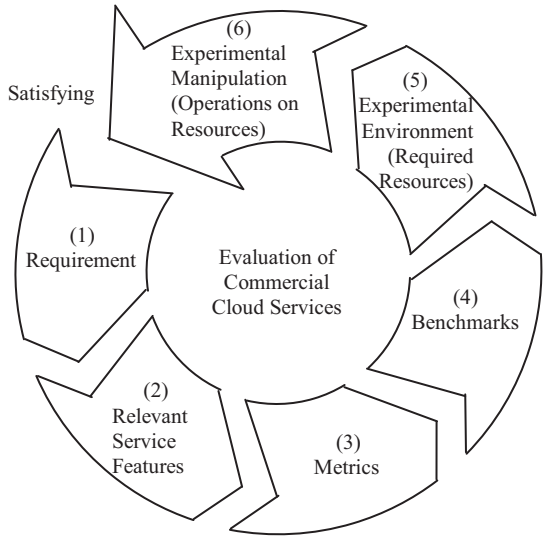

Fig. 1. A conceptual model of the generic process of Cloud services evaluation.

operations on the Cloud resources together with workloads, for example increasing resource amount [BIN10] or varying request frequency [ZLK10]. In fact, given the aforementioned motivation, the existing experiment-intensive studies have been identified and used as the review objects in this SLR.

2.2. A conceptual model of the generic process of Cloud services evaluation

As mentioned previously, Cloud Computing is an emerging computing paradigm (Buyya et al., 2009). When it comes to the evaluation of a computing system (commercial Cloud services in this case), one of the most common issues may be the performance evaluation. Therefore, we decided to borrow the existing lessons from performance evaluation of traditional computing systems to investigate the generic process of Cloud services evaluation. In fact, to avoid possible evaluation mistakes, the steps common to all performance evaluation projects have been summarized ranging from *Stating Goals* to *Presenting Results* (Jain, 1991). By adapting these steps to the above-discussed related work, we decomposed an evaluation implementation process into six common steps and built a conceptual model of Cloud services evaluation, as illustrated in Fig. 1 and specified below.

(1) First of all, the requirement should be specified to clarify the evaluation purpose, which essentially drives the remaining steps of the evaluation implementation.
(2) Based on the evaluation requirement, we can identify the relevant Cloud service features to be evaluated.
(3) To measure the relevant service features, suitable metrics should be determined.
(4) According to the determined metrics, we can employ corresponding benchmarks that may already exist or have to be developed.
(5) Before implementing the evaluation experiment, the experimental environment should be constructed. The environment includes not only the Cloud resources to be evaluated but also resources involved in the experiment.
(6) Given all the aforementioned preparation, the evaluation experiment can be done with human manipulations, which finally satisfies the evaluation requirement.

The conceptual model then played a background and foundation role in the conduction of this SLR. Note that this generic evaluation model can be viewed as an abstract of evaluating any computing paradigm. For Cloud services evaluation, the step adaptation is further explained and discussed as a potential validity threat of this study in Section 7.1.

3. Review method

According to the guidelines for performing SLR (Kitchenham and Charters, 2007), we made minor adjustments and planned our study into a protocol. Following the protocol, we unfold this SLR within three stages.

Planning review:

- Justify the necessity of carrying out this SLR.
- Identify research questions for this SLR.
- Develop SLR protocol by defining search strategy, selection criteria, quality assessment standard, and data extraction schema for conducting review stage.

Conducting review:

- Exhaustively search relevant primary studies in the literature.
- Select relevant primary studies and assess their qualities for answering research questions.
- Extract useful data from the selected primary studies.
- Arrange and synthesize the initial results of our study into review notes.

Reporting review:

- Analyze and interpret the initial results together with review notes into interpretation notes.
- Finalize and polish the previous notes into an SLR report.

3.1. Research questions

Corresponding to the overall aim of this SLR that is to investigate the procedures and experiences of evaluation of commercial Cloud services, six research questions were determined mainly to address the individual steps of the general evaluation process, as listed in Table 1.

In particular, we borrowed the term "scene" from the drama domain for the research question RQ6. In the context of drama, a scene is an individual segment of a plot in a story, and usually settled in a single location. By analogy, here we use "setup scene" to represent an atomic unit for constructing a complete experiment for evaluating commercial Cloud services. Note that, for the convenience of discussion, we broke the investigation of Service features-oriented step into two research questions (RQ2 and RQ3), while we used one research question (RQ6) to cover both Experimental Environment and Experimental Manipulation steps of the evaluation process (cf. Table 1).

3.2. Research scope

We employed three points in advance to constrain the scope of this research. First, this study focused on the commercial Cloud services only to make our effort closer to industry's needs. Second, this study paid attention to Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) without concerning Software as a Service (SaaS). Since SaaS is not used to further build individual business applications [BKKL09], various SaaS implementations may comprise infinite and exclusive functionalities to be evaluated, which

Table 1
Research questions

ID	Research question	Main motivation	Investigated step of the general evaluation process
RQ1	What are the purposes of evaluating commercial Cloud services?	To identify the purposes/requirements of evaluating commercial Cloud services.	Requirement
RQ2	What commercial Cloud services have been evaluated?	To identify the most popular Cloud service and its provider that has attracted the dominant research effort.	Service features
RQ3	What aspects and their properties of commercial Cloud services have been evaluated?	To outline a full scope of aspects and their properties that should be concerned when evaluating Cloud services.	Service features
RQ4	What metrics have been used for evaluation of commercial Cloud services?	To find metrics practically used in the evaluation of commercial Cloud services.	Metrics
RQ5	What benchmarks have been used for evaluation of commercial Cloud services?	To find benchmarks practically used in the evaluation of commercial Cloud services.	Benchmarks
RQ6	What experimental setup scenes have been adopted for evaluating commercial Cloud services?	To identify the components of environment and operations for building evaluation experiments	Experimental environment & Experimental manipulation

could make this SLR out of control even if adopting extremely strict selection/exclusion criteria. Third, following the past SLR experiences (Ali et al., 2010), this study also concentrated on the formal reports in academia rather than the informal evaluation practices in other sources.

3.3. Roles and responsibilities

The members involved in this SLR include a PhD student, a two-people supervisory panel, and a two-people expert panel. The PhD student is new to the Cloud Computing domain, and plans to use this SLR to unfold his research topic. His two supervisors have expertise in the two fields of service computing and evidence-based software engineering respectively, while the expert panel has strong background of computer system evaluation and Cloud Computing. In detail, the expert panel was involved in the discussions about review background, research questions, and data extraction schema when developing the SLR protocol; the specific review process was implemented mainly by the PhD student while under close supervision; the supervisors randomly cross-checked the student's work, for example the selected and excluded publications; regular meetings were held by the supervisory panel with the student to discuss and resolve divergences and confusions over paper selection, data extraction, etc.; unsure issues and data analysis were further discussed by the five members all together.

3.4. Search strategy and process

The rigor of the search process is one of the distinctive characteristics of systematic reviews (Zhang and Ali Babar, 2010). To try to implement an unbiased and strict search, we set a precise publication time span, employed popular literature libraries, alternatively used a set of short search strings, and supplemented a manual search to compensate the automated search for the lack of typical search keywords.

3.4.1. Publication time span

As the term "Cloud Computing" started to gain popularity in 2006 (Zhang et al., 2010), we focused on the literature published from the beginning of 2006. And also considering the possible delay of publishing, we restricted the publication time span between January 1st, 2006 and December 31st, 2011.

3.4.2. Search resources

With reference to the existing SLR protocols and reports for referential experiences, as well as the statistics of the literature search engines (Zhang et al., 2011), we believed that the following five electronic libraries give a broad enough coverage of relevant primary studies:

- ACM Digital Library (http://dl.acm.org/)
- Google Scholar (http://scholar.google.com)
- IEEE Xplore (http://ieeexplore.ieee.org)
- ScienceDirect (http://www.sciencedirect.com)
- SpringerLink (http://www.springer.com)

3.4.3. Proposing search string

We used a three-step approach to proposing search string for this SLR:

(1) Based on the keywords and their synonyms in the research questions, we first extracted potential search terms, such as: "cloud computing", "cloud provider", "cloud service", evaluation, benchmark, metric, etc.
(2) Then, by rationally modifying and combining these search terms, we constructed a set of candidate search strings.
(3) At last, following the Quasi-Gold Standard (QGS) based systematic search approach (Zhang et al., 2011), we performed several pilot manual searches to determine the most suitable search string according to the search performance in terms of sensitivity and precision.

Particularly, the sensitivity and precision of a search string can be calculated as shown in Eqs. (1) and (2) respectively (Zhang et al., 2011).

$$\text{Sensitivity} = \frac{\text{Number of relevant studies retrieved}}{\text{Total number of relevant studies}} 100\% \quad (1)$$

$$\text{Precision} = \frac{\text{Number of relevant studies retrieved}}{\text{Number of studies retrieved}} 100\% \quad (2)$$

In detail, we selected seven Cloud-related conference proceedings (cf. Table 2) to test and contrast sensitivity and precision of different candidate search strings. According to the suggestions of search strategy scales (Zhang et al., 2011), we finally proposed a search string with the *Optimum* strategy, as shown below:

("cloud computing" OR "cloud platform" OR "cloud provider" OR "cloud service" OR "cloud offering") AND (evaluation

Table 2
Sensitivity and precision of the search string with respect to several conference proceedings.

Target proceedings	Sensitivity	Precision
CCGRID 2009	100% (1/1)	100% (1/1)
CCGRID 2010	N/A (0/0)	N/A (0/2)
CCGRID 2011	100% (1/1)	50% (1/2)
CloudCom 2010	100% (3/3)	27.3% (3/11)
CloudCom 2011	100% (2/2)	33.3% (2/6)
CLOUD 2009	N/A (0/0)	N/A (0/0)
CLOUD 2010	N/A (0/0)	N/A (0/6)
CLOUD 2011	66.7% (2/3)	25% (2/8)
GRID 2009	100% (1/1)	50% (1/2)
GRID 2010	100% (1/1)	100% (1/1)
GRID 2011	N/A (0/0)	N/A (0/0)
Total	91.7% (11/12)	28.2% (11/39)

OR evaluating OR evaluate OR evaluated OR experiment OR benchmark OR metric OR simulation) AND (<Cloud provider's name> OR...)

Note that the **(<Cloud provider's name> OR...)** denotes the "OR"-connected names of the top ten Cloud providers (SearchCloudComputing, 2010). The specific sensitivity and precision of this search string with respect to those seven proceedings are listed in Table 2. Given such high sensitivity and more than enough precision (Zhang et al., 2011), although the search string was locally optimized, we have more confidence to expect a globally acceptable search result.

3.4.4. Study identification process

There are three main activities in the study identification process, as listed below: Quickly Scanning based on the automated search, Entirely Reading and Team Meeting for the initially identified studies, and manual Reference Snowballing. The whole process of study identification has been illustrated as a sequence diagram in Fig. 2.

(1) Quickly scanning:

Given the pre-determined search strings, we unfolded automated search in the aforementioned electronic libraries respectively. Relevant primary studies were initially selected by scanning titles, keywords and abstracts.

(2) Entirely reading and team meeting:

The initially identified publications were decided by further reviewing the full-text, while the unsure ones were discussed in the team meeting.

(3) Reference snowballing:

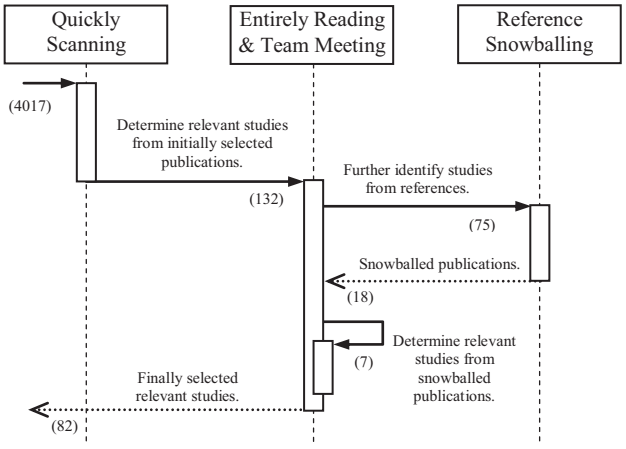

Fig. 2. Study identification process in sequence diagram. The numbers in the brackets denote how many publications were identified/selected at different steps.

To further find possibly missed publications, we also supplemented a manual search by snowballing the references (Kitchenham et al., 2011) of the selected papers found by the automated search. The new papers identified by reference snowballing were also read thoroughly and/or disscussed.

3.5. Inclusion and exclusion criteria

In detail, the inclusion and exclusion criteria can be specified as:

Inclusion criteria:

(1) Publications that describe practical evaluation of commercial Cloud services.
(2) Publications that describe evaluation tool/method/framework for Cloud Computing, and include practical evaluation of commercial Cloud services as a demonstration or case study.
(3) Publications that describe practical evaluation of comparison or collaboration between different computing paradigms involving commercial Cloud services.
(4) Publications that describe case studies of adapting or deploying the existing applications or systems to public Cloud platforms with evaluations. This scenario can be viewed as using real applications to benchmark commercial Cloud services. Note the difference between this criterion and Exclusion Criterion (3).
(5) In particular, above inclusion criteria apply only to regular academic publications (Full journal/conference/workshop papers, technical reports, and book chapters).

Exclusion criteria:

(1) Publications that describe evaluation of non-commercial Cloud services in the private Cloud or open-source Cloud.
(2) Publications that describe only theoretical (non-practical) discussions, like [BKKL09] (cf. Table C.14), about evaluation for adopting Cloud Computing.
(3) Publications that propose new Cloud-based applications or systems, and the aim of the corresponding evaluation is merely to reflect the performance or other features of the proposed application/system. Note the difference between this criterion and Inclusion Criterion (4).
(4) Publications that are previous versions of the later published work.
(5) In addition, short/position papers, demo or industry publications are all excluded.

3.6. Quality assessment criteria

Since a relevant study can be assessed only through its report, and Cloud services evaluation belongs to the field of experimental computer science [Sta09], here we followed the reporting structure of experimental studies (cf. Table 9 in Runeson and Höst, 2009) to assess the reporting quality of one publication. In particular, we divided the reporting structural concerns into two categories: the generic Research reporting quality and the experimental Evaluation reporting quality.

- Research reporting: Is the paper or report well organized and presented following a regular research procedure?
- Evaluation reporting: Is the evaluation implementation work described thoroughly and appropriately?

In detail, we proposed eight criteria as a checklist to examine different reporting concerns in a relevant study:

Criteria of research reporting quality:

Table 3
The data extraction schema.

ID	Data extraction attribute	Data extraction question	Corresponding research question	Investigated step in the general evaluation process
(1)	Author	Who is/are the author(s)?	N/A (Metadata)	N/A (Generic investigation in SLR)
(2)	Affiliation	What is/are the authors' affiliation(s)?		
(3)	Publication title	What is the title of the publication?		
(4)	Publication year	In which year was the evaluation work published?		
(5)	Venue type	What type of the venue does the publication have? (Journal, Conference, Workshop, Book Chapter, or Technical Report)		
(6)	Venue name	Where is the publication's venue? (Acronym of name of journal, conference, workshop, or institute, e.g., ICSE, TSE)		
(7)	Purpose	What is the purpose of the evaluation work in this study?	RQ1	Requirement
(8)	Provider	By which commercial Cloud provider(s) are the evaluated services supplied?	RQ2	Service features
(9)	Service	What commercial Cloud services were evaluated?		
(10)	Service aspect	What aspect(s) of the commercial Cloud services was/were evaluated in this study?	RQ3	Service features
(11)	Aspect property	What properties were concerned for the evaluated aspect(s)?		
(12)	Metric	What evaluation metrics were used in this study?	RQ4	Metrics
(13)	Benchmark	What evaluation benchmark(s) was/were used in this study?	RQ5	Benchmarks
(14)	Environment	What environmental setup scene(s) were concerned in this study?	RQ6	Experimental environment
(15)	Operation	What operational setup scene(s) were concerned in this study?		Experimental manipulation
(16)	Evaluation time	If specified, when was the time or period of the evaluation work?	N/A (Additional data)	N/A (To note evaluation time/period)
(17)	Configuration	What detailed configuration(s) was/were made in this study?	N/A (Additional data)	N/A (To facilitate possible replication of review)

(1) Is the research problem clearly specified?
(2) Are the research aim(s)/objective(s) clearly identified?
(3) Is the related work comprehensively reviewed?
(4) Are findings/results reported?

Criteria of evaluation reporting quality:

(5) Is the period of evaluation work specified?
(6) Is the evaluation environment clearly described?
(7) Is the evaluation approach clearly described?
(8) Is the evaluation result analyzed or discussed?

Each criterion was used to judge one aspect of the quality of a publication, and to assign a quality score for the corresponding aspect of the publication. The quality score can be 1, 0.5, or 0, which represent the quality from excellent to poor as answering Yes, Partial, or No respectively. The overall quality of a publication can then be calculated by summing up all the quality scores received.

3.7. Data extraction and analysis

According to the research questions we previously identified, this SLR used a data extraction schema to collect relevant data from primary studies, as listed in Table 3. The schema covers a set of attributes, and each attribute corresponds to a data extraction question. The relationships between the data extraction questions and predefined research questions are also specified.

In particular, the collected data can be distinguished between the metadata of publications and experimental data of evaluation work. The metadata was mainly used to perform statistical investigation of relevant publications, while the Cloud services evaluation data was analyzed to answer those predefined research questions. Moreover, the data of *evaluation time* collected by question (14) was used in the quality assessment; the data extraction question (15) about detailed *configuration* was to snapshot the evaluation experiments for possible replication of review.

4. Review results

To distinguish the metadata analysis from the evaluation data analysis in this SLR, we first summarize the results of metadata analysis and quality assessment in this section. The findings and answers to those predefined research questions are then discussed in the next section.

Following the search sequence (cf. Fig. 2), 82 relevant primary studies in total were identified. In detail, the proposed search string initially brought 1198, 917, 225, 366 and 1281 results from the ACM Digital Library, Google Scholar, IEEE Xplore, ScienceDirect, and SpringerLink respectively, as listed in the column *Number of Retrieved Papers* of Table 4.

By reading titles and abstracts, and quickly scanning publications in the automated search process, we initially gathered 132 papers. After entirely reading these papers, 75 were selected for this SLR. In particular, 17 undecided papers were finally excluded after our discussion in team meetings; two technical reports and four conference papers were excluded due to the duplication of their latter versions. A set of typical excluded papers (cf. Appendix E) were particularly explained to demonstrate the application of predefined exclusion criteria, as shown in Appendix C. Finally, seven more papers were chosen by reference snowballing in the manual search process. The finally selected 82 primary studies have been listed in Appendix D. The distribution of the identified publications from different electronic databases is listed in Table 4. Note that the four manually identified papers were further located by using Google Scholar.

Table 4
Distribution of relevant studies over electronic libraries.

Electronic library	Number of retrieved papers	Number of relevant papers	Percentage in total relevant papers
ACM Digital Library	1198	21	25.6%
Google Scholar	917	14	17.1%
IEEE Xplore	255	36	43.9%
ScienceDirect	366	0	0%
SpringerLink	1281	11	13.4%
Total	4017	82	100%

Table 5
Distribution of studies over quality.

Type	Score	Number of papers	Percentage
Research reporting quality	2	2	2.44%
	2.5	2	2.44%
	3	22	26.83%
	3.5	3	3.66%
	4	53	64.63%
	Total	82	100%
Evaluation reporting quality	1	1	1.22%
	2	8	9.76%
	2.5	13	15.85%
	3	17	45.12%
	3.5	13	15.85%
	4	10	12.2%
	Total	82	100%

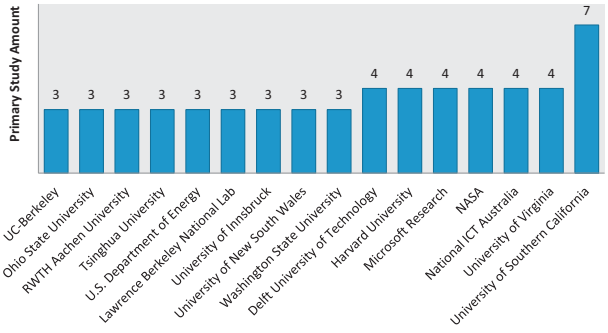

Fig. 4. Universities/organizations with three or more publications.

These 82 primary studies were conducted by 244 authors (co-authors) in total. 40 authors were involved in more than one evaluation works. Interestingly, only four primary studies included co-authors with a direct affiliation with a Cloud services vendor (i.e. Microsoft). On one hand, it may be fairer and more acceptable for third parties' evaluation work to be published. On the other hand, this phenomenon may result from the limitation with our research scope (cf. Section 7.2). To visibly illustrate the distribution of authors' affiliations, we mark their locations on a map, as shown in Fig. 3. Note that the amount of authors' affiliations is more than the total number of the selected primary studies, because some evaluation work could be collaborated between different research organizations or universities. The map shows that, although major research efforts were from USA, the topic of evaluation of commercial Cloud services has been world-widely researched.

Furthermore, we can make those affiliations be accurate to: (1) the background universities of institutes, departments or schools; and (2) the background organizations of individual research laboratories or centers. In this paper, we only focus on the universities/organizations that have published three or more primary studies, as shown in Fig. 4. We believe these universities/organizations may have more potential to provide further and continual work on evaluation for commercial Cloud services in the future.

The distribution of publishing time can be illustrated by grouping the primary studies into years, as shown in Fig. 5. It is clear that the research interests in evaluation of commercial Cloud services have been rapidly increased during the past five years.

In addition, these 82 studies on evaluation of commercial Cloud services scattered in as many as 57 different venues. Such a number of publishing venues are more dispersive than we expected. Although there was not a dense publication zone, in general, those venues could be categorized into five different types: Book Chapter, Technical Report, Journal, Workshop, and Conference, as shown in Fig. 6. Not surprisingly, the publications of evaluation work were relatively concentrated in the Cloud and Distributed Computing related conferences, such as CCGrid, CloudCom, and IPDPS. Moreover, the emerging and Cloud-dedicated books, technical reports, and workshops were also typical publishing venues for Cloud services evaluation work.

As for the quality assessment, instead of listing the detailed quality scores in this paper, here we only show the distribution of the studies over their total reporting quality and total working quality respectively, as listed in Table 5.

Fig. 3. Study distribution over the (co-)author's affiliations.

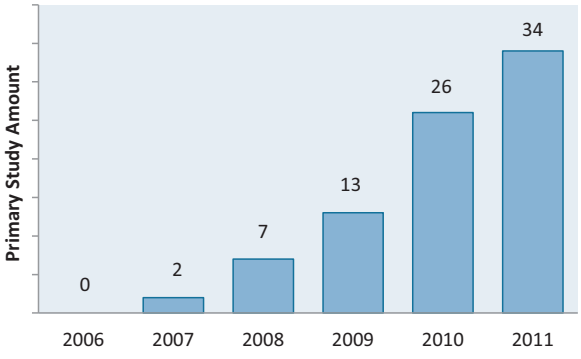

Fig. 5. Study distribution over the publication years.

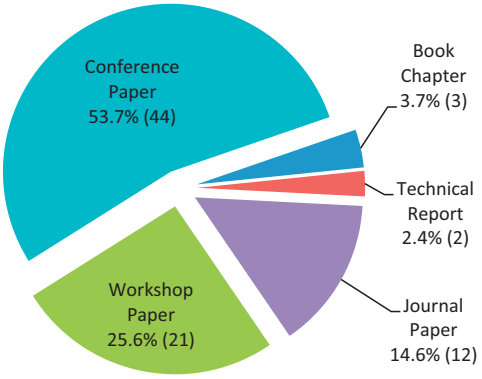

Fig. 6. Study distribution over the publishing venue types.

Table 6
Distribution of studies over evaluation purpose.

Purpose	Primary studies
Cloud resource exploration	[ADWC10] [BCA11] [BK09] [BL10] [BT11] [CA10] [CBH+11] [CHK+11] [dAadCB10] [GCR11] [GK11] [Gar07b] [HLM+10] [ILFL11] [IYE11] [LYKZ10] [LW09] [PEP11] [RD11] [RTSS09] [SDQR10] [Sta09] [SASA+11] [TYO10] [VDG11] [WN10] [WVX11] [YIEO09] [ZLK10]
Business computing in the Cloud	[BS10] [BFG+08] [CMS11] [CRT+11] [DPhC09] [GBS11] [Gar07a] [GS11] [JMW+11] [KKL10] [LML+11] [LYKZ10] [SSS+08]
Scientific computing in the Cloud	[AM10] [BIN10] [DDJ+10] [DSL+08] [EH08] [GWC+11] [Haz08] [HH09] [HHJ+11] [HZK+10] [INB11] [IOY+11] [JDV+09] [JDV+10] [JD11] [JMR+11] [JRM+10] [EKKJP10] [LYKZ10] [LHvI+10] [LJB10] [LJ10] [LML+11] [LZZ+11] [MF10] [NB09] [OIY+09] [PIRG08] [RSP11] [RVG+10] [SKP+11] [SMW+11] [TCM+11] [VJDR11] [MVML11] [VPB09] [Wal08] [WKF+10] [WWDM09] [ZG11]
Comparison between computing paradigms	[CHS10] [IOY+11] [KJM+09] [ZLZ+11]

According to the quality assessment, in particular, we can highlight two limitations of the existing Cloud services evaluation work. Firstly, less than 16% publications specifically recorded the time of evaluation experiments. As mentioned earlier, since commercial Cloud services are rapidly changing, the lack of exposing experimental time would inevitably spoil reusing evaluation results or tracking past data in the future. Secondly, some primary studies did not thoroughly specify the evaluation environments or experimental procedures. As a result, it would be hard for others to replicate the evaluation experiments or learn from the evaluation experiences reported in those studies, especially when their evaluation results became out of date.

5. Discussion addressing research questions

The discussion in this section is naturally organized following the sequence of answers to the six predefined research questions.

5.1. RQ 1: What are the purposes of evaluating commercial Cloud services?

After reviewing the selected publications, we have found mainly four different motivations behind the evaluations of commercial Cloud services, as illustrated in Fig. 7.

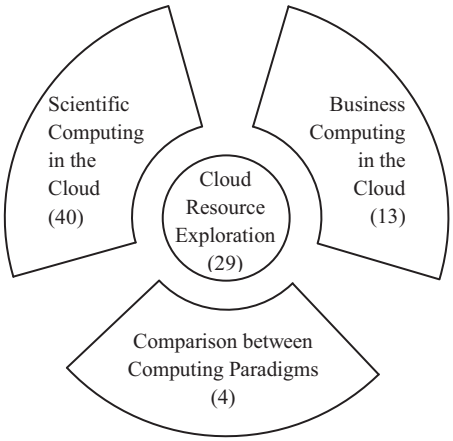

Fig. 7. Purposes of Cloud services evaluation.

The *Cloud Resource Exploration* can be viewed as a root motivation. As the name suggests, it is to investigate the available resources like computation capability supplied by commercial Cloud services. For example, the purpose of study [Sta09] was to purely understand the computation performance of Amazon EC2. The other three research motivations are essentially consistent with the *Cloud Resource Exploration*, while they have specific intentions of applying Cloud resources, i.e., *Scientific/Business Computing in the Cloud* is to investigate applying Cloud Computing to Scientific/Business issues, and *Comparison between Computing Paradigms* is to compare Cloud Computing with other computing paradigms. For example, study [JRM+10] particularly investigated high-performance scientific computing using Amazon Web services; the benchmark Cloudstone [SSS+08] was proposed to evaluate the capability of Cloud for hosting Web 2.0 applications; the study [CHS10] performed a contrast between Cloud Computing and Community Computing with respect to cost effectiveness.

According to these four evaluation purposes, the reviewed primary studies can be differentiated into four categories, as listed in Table 6. Note that one primary study may have more than one evaluation purposes, and we judge evaluation purposes of a study through its described application scenarios. For example, although the detailed evaluation contexts could be broad ranging from Cloud provider selection [LYKZ10] to application feasibility verification [VJDR11, we may generally recognize their purposes as *Scientific Computing in the Cloud* if these studies investigated scientific applications in the Cloud. On the other hand, the studies like "performance evaluation of popular Cloud IaaS providers"

Table 7
Distribution of studies over Cloud service aspects/properties.

Aspect	Property	#Papers	Percentage
Performance	Communication	24	29.27%
	Computation	20	24.39%
	Memory (Cache)	12	14.63%
	Storage	28	34.15%
	Overall performance	48	58.54%
	Total	78	95.12%
Economics	Cost	35	42.68%
	Elasticity	9	10.98%
	Total	40	48.78%
Security	Authentication	1	1.22%
	Data security	4	4.88%
	Infrastructural security	1	1.22%
	Overall security	1	1.22%
	Total	6	7.32%

[SASA+11] only have the motivation *Cloud Resource Exploration* if they did not specify any application scenario.

Apart from the evaluation work motivated by *Cloud Resource Exploration*, we found that there are three times more attention paid to *Scientific Computing in the Cloud* (40 studies) compared to *Business Computing in the Cloud* (13 studies). In fact, the studies aiming at *Comparison between Computing Paradigms* also intended to use Scientific Computing for their discussion and analysis [CHS10, KJM+09]. Given that Cloud Computing emerged as a business model (Zhang et al., 2010), public Cloud services are provided mainly to meet the technological and economic requirements from business enterprises, which does not match the characteristics of scientific computing workloads [HZK+10, OIY+09]. However, the study distribution over purposes (cf. Table 6) suggests that the commercial Cloud Computing is still regarded as a potential and encouraging paradigm to deal with academic issues. We can find a set of reasons for this:

- Since the relevant studies were all identified from academia (cf. Section 7), intuitively, Scientific Computing may seem more academic than Business Computing in the Cloud for researchers.
- Although the public Cloud is deficient for Scientific Computing on the whole due to the relatively poor performance and significant variability [BIN10, JRM+10, OIY+09], smaller scale of computations can particularly benefit from the moderate computing capability of the Cloud [CHS10, HH09, RVG+10].
- The on-demand resource provisioning in the Cloud can satisfy some high-priority or time-sensitive requirements of scientific work when in-house resource capacity is insufficient [CHS10, Haz08, OIY+09, WWDM09].
- It would be more cost effective to carry out temporary jobs on Cloud platforms to avoid the associated long-term overhead of powering and maintaining local computing systems [CHS10, OIY+09].
- Through appropriate optimizations, the current commercial Cloud can be improved for Scientific Computing [EH08, OIY+09].
- Once commercial Cloud vendors pay more attention to Scientific Computing, they can make the current Cloud more academia-friendly by slightly changing their existing infrastructures [HZK+10]. Interestingly, the industry has acknowledged the academic requirements and started offering services for solving complex science/engineering problems (Amazon, 2011).

5.2. RQ 2: What commercial Cloud services have been evaluated?

Evaluations are based on services available from specific Cloud providers. Before discussing the individual Cloud services, we identify the service providers. Nine commercial Cloud providers have been identified in this SLR: Amazon, BlueLock, ElasticHosts, Flexiant, GoGrid, Google, IBM, Microsoft, and Rackspace. Mapping the 82 primary studies to these nine providers, as shown in Fig. 8, we show that the commercial Cloud services attracting most evaluation efforts are provided by Amazon. Note that one primary study may cover more than one Cloud provider. This phenomenon is reasonable because Amazon has been treated as one of the top and key Cloud Computing providers in both industry and academia (Buyya et al., 2009; Zhang et al., 2010).

With different public Cloud providers, we have explored the evaluated Cloud services in the reviewed publications, as listed in Appendix B. Note that the Cloud services are identified according to their commercial definitions instead of functional descriptions. For example, the work [HLM+10] explains Azure Storage Service and Azure Computing Service respectively, whereas we treated them as two different functional resources in the same Windows Azure service. The distribution of reviewed publications over detailed services is illustrated as shown in Fig. 9. Similarly, one primary study

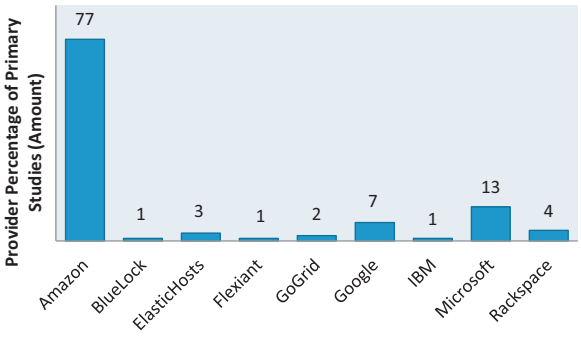

Fig. 8. Distribution of primary studies over Cloud providers.

may perform evaluation of multiple commercial Cloud services. In particular, five services (namely Amazon EBS, EC2 and S3, Google AppEngine, and Microsoft Windows Azure) were the most frequently evaluated services compared with the others. Therefore, they can be viewed as the representative commercial Cloud services, at least in the context of Cloud services evaluation. Note that bias could be involved in the service identification in this work due to the pre-specified providers in the search string, as explained in Section 7.3.

Among these typical commercial Cloud services, Amazon EBS, EC2 and S3 belong to IaaS, Google AppEngine is PaaS, while Microsoft Windows Azure is recognized as a combination of IaaS and PaaS [ZLK10]. IaaS is the on-demand provisioning of infrastructural computing resources, and the most significant advantage is its flexibility [BKKL09]. PaaS refers to the delivery of a platform-level environment including operating system, software development frameworks, and readily available tools, which limits customers' control while taking complete responsibility of maintaining the

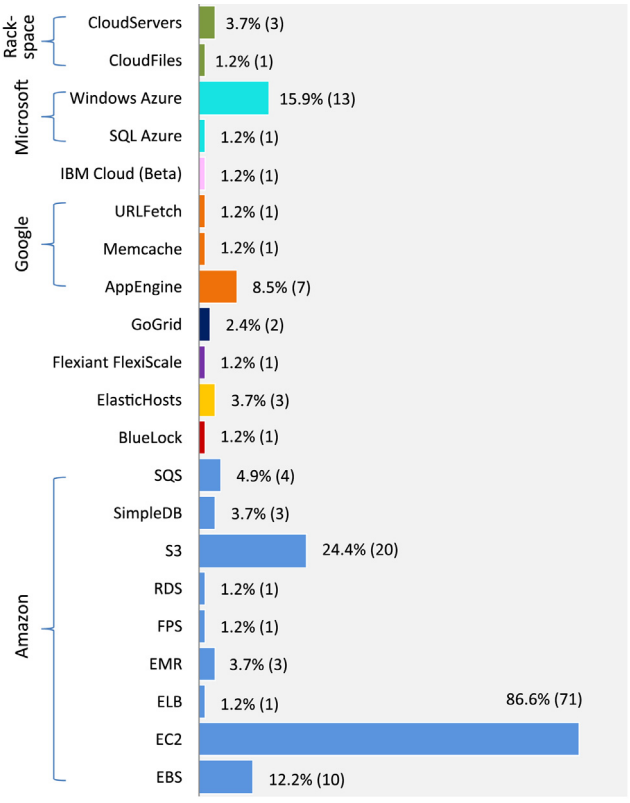

Fig. 9. Distribution of primary studies over Cloud services.

Table 8
Distribution of metrics over Cloud service aspects/properties (based on Li et al., 2012c and updated).

Aspect	Property	#Metrics
Performance	Communication	9
	Computation	7
	Memory (Cache)	7
	Storage	11
	Overall performance	18
Economics	Cost	18
	Elasticity	4
Security	Authentication	1
	Data security	3
	Infrastructural security	1
	Overall security	1

environment on behalf of customers [BKKL09]. The study distribution over services (cf. Fig. 9) indicates that IaaS attracts more attention of evaluation work than PaaS. Such a finding is essentially consistent with the previous discussions when answering RQ1. The flexible IaaS may better fit into the diverse Scientific Computing. In fact, niche PaaS and SaaS are designed to provide additional benefits for their targeting applications, while IaaS is more immediately usable for particular and sophisticated applications [JD11] (Harris, 2012). In other words, given the diversity of requirements in the Cloud market, IaaS and PaaS would serve different types of customers, and they cannot be replaced with each other. This finding can also be confirmed by a recent industry event: the traditional PaaS provider Google just offered a new IaaS – Compute Engine (Google, 2012).

5.3. RQ 3: What aspects and their properties of commercial Cloud services have been evaluated?

The aspects of commercial Cloud services can be initially investigated from general surveys and discussions about Cloud Computing. In brief, from the view of Berkeley (Armbrust et al., 2010), Economics of Cloud Computing should be particularly emphasized in deciding whether to adopt Cloud or not. Therefore, we considered Economics as an aspect when evaluating commercial Cloud services. Meanwhile, although we do not agree with all the parameters identified for selecting Cloud Computing/Provider in Habib et al. (2010), we accepted Performance and Security as two significant aspects of a commercial Cloud service. Such an initial investigation of service aspects has been verified by this SLR. Only Performance, Economics, and Security and their properties have been evaluated in the primary studies.

The detailed properties and the corresponding distribution of primary studies are listed in Table 7. Note that a primary study usually covers multiple Cloud service aspects and/or properties. In particular, we only take into account the physical properties for the Performance aspect in this paper. The capacities of different physical properties and their sophisticated correlations (cf. Fig. 10) have been specified in our previous work (Li et al., 2012a).

Overall, we find that the existing evaluation work overwhelmingly focused on the performance features of commercial Cloud services. Many other theoretical concerns about commercial Cloud Computing, Security in particular, were not well evaluated yet in practice. Given the study distribution over service aspects/properties (cf. Table 7), several research gaps can be revealed or confirmed:

- Since memory/cache could closely work with the computation and storage resources in computing jobs, it is hard to exactly distinguish the effect to performance brought by memory/cache, which may be the main reason why few dedicated Cloud memory/cache evaluation studies were found from the literature. In addition to the memory performance, the memory hierarchy could be another interesting issue to be evaluated [OIY+09].
- Although one major benefit claimed for Cloud Computing is elasticity, it seems difficult for people to know how elastic a Cloud platform is. In fact, evaluating elasticity of a Cloud service is not trivial (Kossmann and Kraska, 2010), and there is little explicit measurement to quantify the amount of elasticity in a Cloud platform (Islam et al., 2012).
- The security of commercial Cloud services has many dimensions and issues people should be concerned with (Armbrust et al., 2010; Zhang et al., 2010). However, not many security evaluations were reflected in the identified primary studies. Similar to the above discussion about elasticity evaluation, the main reason may be that the security is also hard to quantify (Brooks, 2010). Therefore, we conclude that the Elasticity and Security evaluation of commercial Cloud services could be a long-term research challenge.

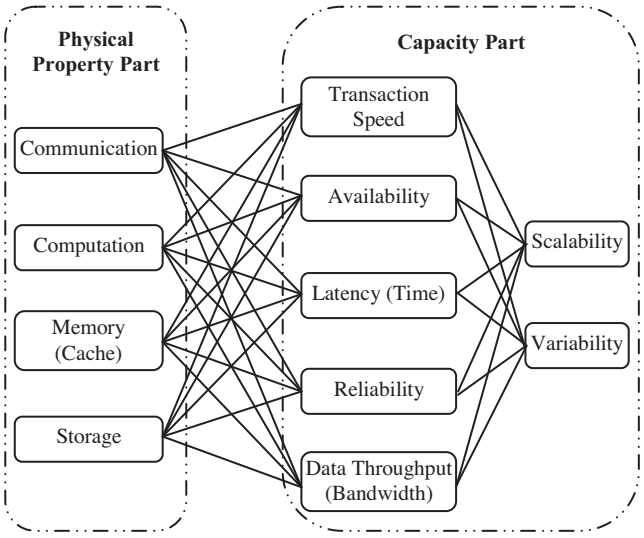

Fig. 10. The properties of the performance aspect (from Li et al., 2012a).

5.4. RQ 4: What metrics have been used for evaluation of commercial Cloud services?

Benefiting from the above investigation of aspects and their properties of commercial Cloud services, we can conveniently identify and organize their corresponding evaluation metrics. In fact, more than 500 metrics including duplications have been isolated from the experiments described in the primary studies. After removing the duplications, we categorized and arranged the metrics naturally following the aforementioned Cloud service aspects/properties. Note that we judged duplicate metrics according to their usage contexts instead of names. Some metrics with different names could be essentially duplicate ones, while some metrics with identical name should be distinguished if they are used for different evaluation objectives. For example, the metric *Upload/Download Data Throughput* has been used for evaluating both Communication [Haz08] and Storage [PIRG08], and therefore it was arranged under both Cloud service properties.

Due to the limit of space, we do not elaborate all the identified metrics in this paper. In fact, we have summarized the existing evaluation metrics into a catalogue to facilitate the future practice and research in the area of Cloud services evaluation (Li et al., 2012c). Here we only give a quick impression of their usage by displaying the distribution of those metrics, as shown in Table 8.

Table 9
The traditional benchmarks used in Cloud services evaluation.

Benchmark	Type	Applicability	Evaluated Cloud service property (with one study as a sample)				
			Communication	Computation	Memory/Cache	Storage	Overall performance
An Astronomy workflow	Application	1					[VJDR11]
Application/Workflow Suite	Application	3	[JRM+10]	[DPhC09]			[JRM+10]
B+_Tree indexing system	Application	1			[CHK+11]	[CHK+11]	
Badabing Tool	Micro	1	[WN10]				
Betweenness Centrality	Application	1					[RSP11]
BitTorrent	Application	1				[PIRG08]	
BLAST/BLAST+	Application	6					[LJB10]
Bonnie/Bonnie++	Micro	4			[OIY+09]	[OIY+09]	
Broadband	Application	3			[JD11]		[JDV+09]
CacheBench	Micro	2			[OIY+09]		
CAP3	Application	1					[GWC+11]
Classify gene data	Application	1					[VPB09]
Compiling Linux Kernel	Application	1		[BK09]			
CSFV	Application	1					[HZK+10]
Dhrystone	Synthetic	1		[PEP11]			
EnKF-based matching	Application	1					[EKKJP10]
Epigenome	Application	3		[JD11]			[JDV+09]
FEFF84 MPI	Application	1					[RVG+10]
Fibonacci	Micro	1		[IYE11]			
FIO	Micro	1				[SASA+11]	
fMRI brain imaging	Application	1					[VPB09]
GASOLINE	Application	1					[RVG+10]
Grapes	Application	1					[ZLZ+11]
GTM	Application	1					[GWC+11]
Hadoop App	Application	2					[DDJ+10]
hdparm tool	Synthetic	1				[ZLZ+11]	
HPCC: b_eff	Micro	3	[OIY+09]				
HPCC: DGEMM	Micro	5		[JRM+10]	[BIN10]		
HPCC: FFTE	Synthetic	1		[JRM+10]			
HPCC: HPL	Synthetic	8		[OIY+09]	[BIN10]		[AM10]
HPCC: PTRANS	Synthetic	1	[JRM+10]				
HPCC: RandomAccess	Synthetic	3			[JRM+10]		
HPCC: STREAM	Micro	6			OIY+09]		
iperf	Micro	4	[LYKZ10]				
Intel MPI Bench	Micro	3	[HH09]				
IOR	Synthetic	4			[GCR11]	[EH08]	
Isabel	Application	1	[CRT+11]				
KMeans Clustering	Application	1					[BCA11]
Land Elevation Change	Application	1			[CA10]		
Latency Sensitive Website	Application	1	[LYKZ10]				
Livermore Loops	Synthetic	1		[PEP11]			
LMbench	Micro	4		[JMW+11]			[IOY+11]
Lublin99	Synthetic	1					[dAadCB10]
MapReduce App	Application	1					[SDQR10]
MG-RAST +BLAST	Application	1					[WWDM09]
Minion Constraint solver	Application	1					[GK11]
mpptest	Micro	1	[HZK+10]				
MODIS Processing	Application	2					[LHvI+10]
Montage	Application	4				[JD11]	[JDV+09]
NaSt3DGPF	Application	1					[ZG11]
NetPIPE	Micro	1	[JMW+11]				
NPB: BT	Synthetic	2				[AM10]	
NPB: BT-IO	Synthetic	2				[EH08]	
NPB: EP	Micro	1		[AM10]			
NPB: GridNPB: ED	Synthetic	1					[MVML11]
NPB: original	Synth + Micro	4	[ZLZ+11]	[CHS10]			[AM10]
NPB-OMP	Synthetic	2					[Wal08]
NPB-MPI	Synthetic	2	[HZK+10]				[Wal08]
NPB-MZ	Synthetic	1					[HZK+10]
OMB-3.1 with MPI	Micro	1	[EH08]				
Operate/Transfer Data	Micro	19	[BK09]			[LYKZ10]	
PageRank	Application	1					[BCA11]
Passmark CPU Mark	Micro	1		[LML+11]			
PCA	Application	1					[BCA11]
Phoronix Test Suite	Application	1					[LML+11]
ping	Micro	5	[LYKZ10]				
POP	Application	2				[LZZ+11]	[ZLZ+11]
PostMark	Synthetic	1				[WVX11]	
ROIPAC workflow	Application	1					[TCM+11]
RUBBoS + MySQL Cluster	Application	1					[JMW+11]
SAGA BigJob System	Application	1					[LJ10]
Seismic Source Inversion	Application	1					[SMW+11]
Simplex	Micro	1		[SASA+11]			
SNfactory	Application	1	[JMR+11]	[JMR+11]		[JMR+11]	[JMR+11]
Social Website	Application	1					[RD11]

Table 9 (*Continued*)

Benchmark	Type	Applicability	Evaluated Cloud service property (with one study as a sample)				
			Communication	Computation	Memory/Cache	Storage	Overall performance
SPECjvm 2008	Synthetic	1					[LYKZ10]
SPECweb	Synthetic	2	[LW09]	[LW09]			[CBH+11]
Sysbench on MySQL	Application	1					[SSS+08]
Timed Benchmark	Synthetic	1				[GCR11]	
TORCH Benchmark Suite	Synthetic	1					[PEP11]
TPC-E	Synthetic	1					[HLM+10]
TPC-W	Synthetic	4				[LYKZ10]	[KKL10]
Ubench	Micro	1		[SDQR10]	[SDQR10]		
WCD	Application	1					[Haz08]
Whetstone	Synthetic	1		[KJM+09]			
WSTest	Synthetic	1					[Sta09]

Given the distribution together with the catalogue of Cloud services evaluation metrics, we summarize several findings below:

- The existing evaluation work has used a large number of metrics to measure various performance features as well as the cost of commercial Cloud services. This confirms the current fashion of cost evaluation: based on performance evaluation, evaluators analyze and estimate the real expense of using Cloud services [LML+11, ZLZ+11]. We may name this type of evaluated cost as resource cost. In fact, the cost of Cloud Computing may cover a wide range of theoretical concerns, such as migration cost, operation cost, etc. (Armbrust et al., 2010). However, those costs depend on specific systems, technologies, human activities, and even environmental factors. Performing generic cost evaluation could then be a tremendous challenge. A promising solution to this challenge is to replace the cost with other steady factors for evaluation. For example, we may estimate the size of Cloud migration projects instead of directly evaluating the migration cost (Tran et al., 2011).
- There is still a lack of effective metrics for evaluating Cloud elasticity. As mentioned previously, it is not easy to explicitly quantify the amount of elasticity of a Cloud service. To address this research gap, as far as we know, the most recent effort is a sophisticated Penalty Model that measures the imperfections in elasticity of Cloud services for a given workload in monetary units (Islam et al., 2012).
- It seems that there is no suitable metric yet to evaluate security features of Cloud services, which also confirms the previous findings in Section 5.3. Since security is hard to quantify (Brooks, 2010), current security evaluation has been realized mainly by qualitative discussions. A relatively specific suggestion for security evaluation of Cloud services is given in [PIRG08]: the security assessment can start with an evaluation of the involved risks. As such, we can use a pre-identified risk list to discuss the security strategies supplied by Cloud services.

5.5. RQ 5: What benchmarks have been used for evaluation of commercial Cloud services?

This SLR has identified around 90 different benchmarks in the selected studies of Cloud services evaluation. As discussed in the related work (cf. Section 2), there are several emerging and dedicated Cloud benchmarks, such as YCSB [CST+10], CloudStone [SSS+08], and CloudSuite (Ferdman et al., 2012). Traditional benchmarks have still been overwhelmingly used in the existing practices of Cloud services evaluation, as summarized in Table 9. Note that, in Table 9, each benchmark together with a corresponding evaluated service property cites only one relevant study as an instance. In particular, the evaluated Economics and Security properties are not reflected in this table. First, the existing cost evaluation studies were generally based on the corresponding performance evaluation [LML+11, ZLZ+11]. Second, the selected studies did not specify any distinct benchmark for evaluating elasticity and security. Through Table 9 we show that, although the traditional benchmarks were recognized as being insufficient for evaluating commercial Cloud services [BKKL09], traditional benchmarks can still satisfy at least partial requirements of Cloud services evaluation.

Table 10
Popular traditional benchmarks for evaluating different Cloud service properties.

Cloud service property	Popular traditional benchmarks
Communication	iperf, ping, Operate/Transfer Data
Computation	HPCC: DGEMM, HPCC: HPL, LMBench
Memory/Cache	HPCC: STREAM
Storage	Bonnie/Bonnie++, IOR, NPB: BT/BT-IO, Operate/Transfer Data
Overall performance	BLAST, HPCC: HPL, Montage, NPB suite, TPC-W

Moreover, one benchmark may be employed in multiple evaluation practices. The numerous evaluators' experiences can then be used to indicate the applicability of a particular benchmark. Here we define a benchmark's "Applicability" as the number of the related studies. Through the applicability of different traditional benchmarks (cf. Table 9), we list the popular benchmarks as recommendations for Cloud services evaluation, as shown in Table 10.

In addition, following the evolution of benchmarking in the computing area (Lewis and Crews, 1985), we summarized three types of benchmarks used for evaluating commercial Cloud services: Application Benchmark, Synthetic Benchmark, and Micro-Benchmark.

- Application Benchmark refers to the real-world software systems that are deployed to the Cloud and used as potentially true measures of commercial Cloud services.
- Synthetic Benchmark is not a real application, but a well-designed program using representative operations and workload to simulate a typical set of applications.
- Micro-Benchmark is a relatively simple program that attempts to measure a specific component or a basic feature of Cloud services.

Table 11
Distribution of studies over benchmark types.

Benchmark type	#Papers	Percentage
Application Only	27	32.93%
Synthetic Only	11	13.41%
Micro Only	17	20.73%
Application + Synthetic	3	3.66%
Application + Micro	12	14.63%
Synthetic + Micro	6	7.32%
All Three	6	7.32%
Total	82	100%

Table A.12
Detailed score card for the quality assessment of the 82 primary studies.

Study	QA1	QA2	QA3	QA4	Research reporting score	QA5	QA6	QA7	QA8	Evaluation reporting score	Total score
[ADWC10]	0	1	1	1	3	0.5	1	0	1	2.5	5.5
[AM10]	1	1	0	1	3	0	1	0	1	2	5
[BCA11]	1	1	1	1	4	0	1	1	1	3	7
[BFG+08]	1	1	1	1	4	0.5	1	1	1	3.5	7.5
[BIN10]	1	1	1	1	4	0.5	1	1	1	3.5	7.5
[BK09]	1	1	0	1	3	0.5	1	0	0.5	2	5
[BL10]	1	1	0	1	3	0.5	1	0	1	2.5	5.5
[BS10]	1	1	1	1	4	0.5	1	1	1	3.5	7.5
[BT11]	1	1	1	1	4	1	1	1	1	4	8
[CA10]	1	1	1	1	4	0	1	1	1	3	7
[CBH+11]	1	1	0	1	3	0	1	1	1	3	6
[CHS10]	1	1	1	1	4	1	0.5	0.5	1	3	7
[CHK+11]	1	1	1	1	4	0	1	1	1	3	7
[CMS11]	1	1	1	1	4	0	1	1	1	3	7
[CRT+11]	1	1	1	1	4	0	1	1	1	3	7
[dAadCB10]	1	1	1	1	4	0	1	1	1	3	7
[DDJ+10]	1	1	1	1	4	0	1	1	1	3	7
[DPhC09]	1	1	1	1	4	0.5	1	1	1	3.5	7.5
[DSL+08]	1	1	1	1	4	0	1	1	1	3	7
[EH08]	1	1	0	1	3	0	0.5	0.5	1	2	5
[GBS11]	1	1	1	1	4	0	0.5	1	1	2.5	6.5
[GCR11]	1	1	1	1	4	1	1	1	1	4	8
[Gar07a]	0	1	0	1	2	0	0.5	0.5	1	2	4
[GK11]	1	1	1	1	4	0	1	0.5	1	2.5	6.5
[Gar07b]	1	1	1	1	4	1	1	1	1	4	8
[GS11]	1	1	1	1	4	0	1	0.5	1	2.5	6.5
[GWC+11]	0	1	1	1	3	0	1	1	1	3	6
[Haz08]	1	1	0	1	3	0	1	1	1	3	6
[HH09]	1	1	1	1	4	0	1	1	1	3	7
[HHJ+11]	1	1	1	1	4	1	1	1	1	4	8
[HLM+10]	1	1	1	1	4	1	1	1	1	4	8
[HZK+10]	1	1	1	1	4	0	1	0.5	1	2.5	6.5
[ILFL11]	1	1	1	1	4	0	1	1	1	3	7
[[INB11]	1	1	0	1	3	0	1	1	1	3	6
[IOY+11]	1	1	1	1	4	0	1	1	1	3	7
[IYE11]	1	1	1	1	4	1	1	1	1	4	8
[JDV+09]	1	1	0.5	1	3.5	0	1	1	1	3	6.5
[JDV+10]	1	1	1	1	4	0	1	1	1	3	7
[JD11]	1	1	0.5	1	3.5	0	1	1	1	3	6.5
[JMR+11]	1	1	1	1	4	0.5	1	1	1	3	7.5
[JMW+11]	1	1	1	1	4	0	1	1	1	3	7
[JRM+10]	1	1	1	1	4	0	1	1	1	3	7
[KJM+09]	1	1	1	1	4	0	1	0.5	1	2.5	6.5
[EKKJP10]	1	1	1	1	4	0	0.5	0.5	1	2	6
[KKL10]	1	1	1	1	4	0.5	1	1	1	3.5	7.5
[LHvl+10]	1	1	1	1	4	0	1	1	1	3	7
[LJ10]	1	1	0	1	3	0.5	1	0.5	1	3	6
[LJB10]	1	1	1	1	4	0	1	1	1	3	7
[LML+11]	1	1	1	1	4	1	1	1	1	4	8
[LW09]	1	1	0	1	3	0.5	1	1	1	3.5	6.5
[LYKZ10]	1	1	1	1	4	1	0.5	0.5	1	3	7
[LZZ+11]	1	1	1	1	4	0	1	0	1	2	6
[MF10]	0.5	1	0	1	2.5	0	1	1	1	3	5.5
[NB09]	1	1	0.5	1	3.5	0.5	1	1	1	3.5	7
[OIY+09]	1	1	1	1	4	0.5	1	1	1	3.5	7.5
[PEP11]	1	1	1	1	4	0	1	1	1	3	7
[PIRG08]	1	1	0	1	3	1	1	1	1	4	7
[RD11]	1	1	0	1	3	0	0.5	0.5	1	2	5
[RSP11]	1	1	1	1	4	0	1	1	1	3	7
[RTSS09]	1	1	0	1	3	0	1	1	1	3	6
[RVG+10]	1	1	0	1	3	0	1	0.5	1	2.5	5.5
[SASA+11]	1	1	1	1	4	0	1	1	1	3	7
[SDQR10]	1	1	1	1	4	1	1	1	1	4	8
[SKP+11]	0.5	1	0.5	1	3	0.5	1	1	1	3.5	6.5
[SMW+11]	0	1	1	1	3	0	0.5	1	1	2.5	5.5
[SSS+08]	1	1	0	1	3	0	1	0.5	1	2.5	5.5
[Sta09]	1	1	1	1	4	0.5	1	1	1	3.5	7.5
[TCM+11]	1	1	1	1	4	0.5	1	1	1	3.5	7.5
[TYO10]	0	1	0	1	2	0	0.5	0	0.5	1	3
[VDG11]	1	1	0	1	3	0	0.5	1	0.5	2	5
[VJDR11]	1	1	1	1	4	0	1	1	1	3	7
[MVML11]	1	1	1	1	4	0	1	1	1	3	7
[VPB09]	1	1	0	1	3	0	1	0.5	1	2.5	5.5
[Wal08]	1	1	0	1	3	0	1	1	1	3	6

Table A.12 (*Continued*)

Study	QA1	QA2	QA3	QA4	Research reporting score	QA5	QA6	QA7	QA8	Evaluation reporting score	Total score
[WKF+10]	1	1	1	1	4	0	1	0.5	1	2.5	6.5
[WN10]	1	1	1	1	4	0.5	1	1	1	3.5	7.5
[WVX11]	1	1	1	1	4	0	1	1	1	3	7
[WWDM09]	1	1	0	1	3	0.5	1	0.5	1	3	6
[YIEO09]	1	1	1	1	4	0	0.5	1	1	2.5	6.5
[ZG11]	0.5	1	0	1	2.5	0	1	1	1	3	5.5
[ZLK10]	1	1	1	1	4	1	0.5	1	1	3.5	7.5
[ZLZ+11]	1	1	1	1	4	1	1	1	1	4	8
Total	75.5	82	58	82	297.5	22	76	68	80.5	246	544
Average	0.92	1	0.71	1	3.63	0.27	0.93	0.83	0.98	3	6.63

To give a quick impression of what types of benchmarks were adopted in the current Cloud services evaluation work, we list the distribution of primary studies over employed benchmark types, as shown in Table 11.

It can be seen that more than half of the primary studies adopted only one particular type of benchmark to evaluate commercial Cloud services. Given that different types of benchmarks reveal different service natures, it is impossible to use one benchmark to fit all when performing Cloud services evaluation. Thus, a recommendation from this SLR is to employ a suite of mixed types of benchmarks to evaluate Cloud services in the future.

5.6. RQ 6: What experimental setup scenes have been adopted for evaluating commercial Cloud services?

As mentioned in Section 3.1, we used "setup scene" to indicate an atomic unit for constructing complete Cloud services evaluation experiments. Through extracting different data from a primary study for respectively answering the data extraction questions (12) and (13) (cf. Section 3.7), we can distinguish between environmental setup scenes and operational setup scenes. The environmental setup scenes indicate static descriptions used to specify required experimental resources, while the operational setup scenes indicate dynamic operations that usually imply repeating an individual experiment job under different circumstances. For the convenience of analysis, the operational setup scenes were further divided into three groups with respect to experimental Time, Location, and Workload. In detail, ten environmental setup scenes and 15 operational setup scenes have been identified, which can be organized as an experimental setup scene tree, as shown in Fig. 11.

We have developed a taxonomy to clarify and structure these 25 experimental setup scenes in a separate piece of work (Li et al., 2012a). In particular, the rounded rectangle with dashed line (Fig. 11) represents the setup scenes that are either uncontrollable (*Different Physical Locations of Cloud Resource*) or unemployed yet (*Multiple Instance Types*). The physical location of a particular Cloud resource indicates its un-virtualized environment. The un-virtualized difference then refers not only to the difference in underlying hardware like different model of real CPU, but also to the difference between VMs sharing or not sharing underlying hardware. As for the setup scene *Multiple Instance Types*, although it is possible to assign different functional roles to different types of VM instances to finish a single experiment job, we have not found such jobs in the reviewed literature.

Overall, by using the experimental setup scene tree, we can easily locate or enumerate individual environmental and operational setup scenes for Cloud services evaluation studies. As such, the answer to this research question may be employed essentially to facilitate drawing experimental lessons from the existing evaluation reports, and to facilitate the evaluation-related communication among the Cloud Computing community.

6. Experiences of applying the SLR method

This SLR was prepared by a review team and two consultants, implemented primarily by a PhD student under supervision, and discussed and finalized by the whole team. According to our practice of conducting this study, we summarized some experiences to which or against which researchers can refer or debate in future SLR implementations.

First of all, a question-oriented SLR is apparently more efficient than an ad hoc review. For a new comer in a particular research area, it is difficult to measure his/her study progress if he/she is doing an ad hoc literature review. On the contrary, benefiting from the SLR, the progress becomes traceable by following a standardized procedure (Kitchenham and Charters, 2007).

However, it should be noticed that traditional ad hoc reviews cannot be completely replaced with SLRs. Although supervisors can help introduce the background and/or motivation in advance, it is crucial for the student to comprehend enough relevant domain knowledge before starting an SLR. In terms of our experience with this SLR, an ad hoc review still showed its value in obtaining domain knowledge in a short period, which confirms that it is necessary to "thoroughly understand the nature and scale of the task at hand before undertaking a SLR" (Major et al., 2011). When an SLR is supposed to be implemented by PhD students in an unfamiliar area, we should also estimate and consider the additional time on students' traditional review.

Moreover, our study also confirmed that a pilot review is vital for an SLR (Babar and Zhang, 2009). The pilot review of an SLR can be viewed as a bridge between the SLR and the corresponding ad hoc review. On one hand, the pilot review can reinforce or revise the reviewers' comprehension of domain-specific knowledge. On the other hand, the pilot review can help refine research questions, improve search strategy, and verify data extraction schema by trying to answer research questions. Therefore, we suggest that a pilot review can be done together with constructing the SLR protocol.

Additionally, for some research topics, the employment of an SLR is worthy of regular use to keep the relevant data or knowledge current to support those topics. According to Zhang and Babar's survey (Zhang and Babar, 2011), most of existing SLRs in software engineering area seem one-off studies, such as to outline state-of-the-art or to get knowledge within a particular research region. Whereas, for this study, we plan to use the collected data to fill an experience base to support a Cloud services evaluation methodology. Considering the knowledge in an expert system should be updated regularly, it is necessary to always keep the corresponding experience base up to date. In this case, therefore, we will continually collect relevant primary studies, and periodically update this SLR work.

Overall, in this study, the SLR method has been verified suitable and helpful for a first-year PhD student to accumulate knowledge and identify his research opportunities.

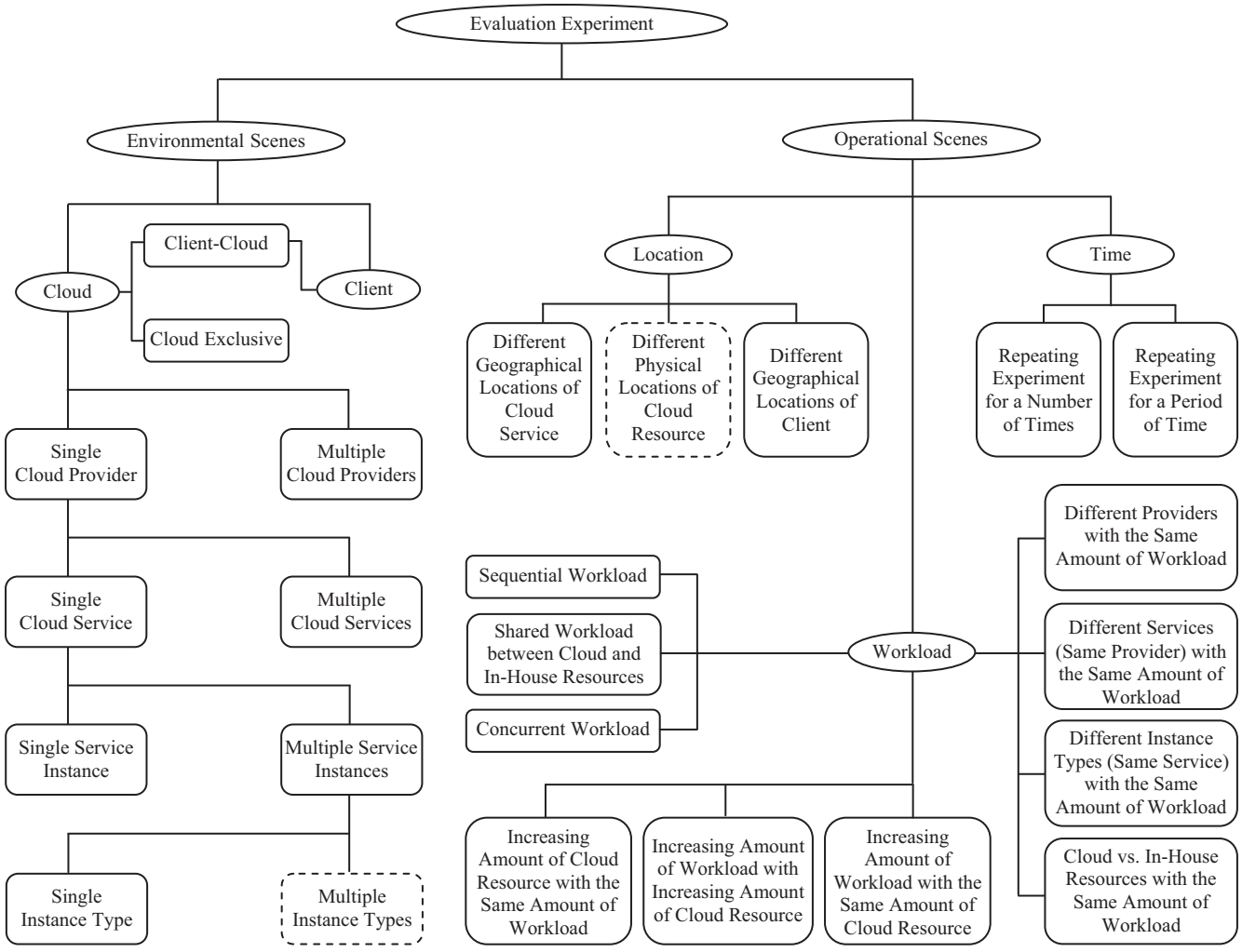

Fig. 11. Experimental setup scene tree of performance evaluation of commercial Cloud services (from Li et al., 2012a).

7. Threats to validity

Although we tried to conduct this SLR study as rigorously as possible, it may have still suffered from several validity threats, as listed below. The future work should take into account these limitations when interpreting or directly using the findings or conclusions in this report.

7.1. Conceptual model of Cloud services evaluation

The construction of this SLR and the following investigation into Cloud services evaluation were based on the proposed conceptual model (cf. Section 2). Therefore, any inaccuracy in the conceptual model of Cloud services evaluation may bring flaws in this study. As previously mentioned, we built this conceptual model by adapting a systematic performance evaluation approach (Jain, 1991). In particular, we deliberately ignored two steps in the general process of evaluation implementation, namely evaluation factor identification and experimental result analysis. The reason for ignoring the former, we found that it was hard to directly extract experimental factors from the primary studies. To the best of our knowledge, although the existing evaluation experiments essentially involved factors, none of the current Cloud evaluation studies specified "experimental factors" (Montgomery, 2009) in advance to design evaluation experiments and analyze the experimental results. In fact, we finally investigated potential factors through a secondary analysis of the answer to RQ6 in this SLR (Li et al., 2012b). The reason for ignoring the latter, as mentioned in the Introduction, we conducted this SLR study to investigate the procedures and experiences of Cloud services evaluation rather than the evaluation results. Overall, although we are not aware of any bias introduced by this conceptual model, other researchers with different interest may have different opinions about the intentionally ignored information.

7.2. Research scope

The practices of Cloud services evaluation are reported in various sources, such as academic publications, technical websites, blogs, etc. In particular, the academic publications are normally formal reports after rigorous peer reviewing. Considering the generally specific and precise documentation of evaluation implementations in formal publications (Ali et al., 2010), we limited this SLR to academic studies only. There is no doubt that informal descriptions of Cloud services evaluation in blogs and technical websites can also provide highly relevant information. However, on the one hand, it is impossible to explore and collect useful data from different study sources all at once. On the other hand, the published evaluation studies can be viewed as typical representatives of the existing ad hoc evaluation practices. By using the SLR method to exhaustively investigate the academic studies, we are still able to rationally show the representative state-of-the-practice

of the evaluation of commercial Cloud services. In fact, we proposed to use the result of this SLR to construct a knowledge base first. The knowledge base can be gradually extended and enriched by including the other informal empirical studies of Cloud services evaluation.

7.3. Completeness

Given the increasing number of studies in this area, we note that we cannot guarantee to have captured all the relevant studies. The possible reasons could be various ranging from the search engines to the search string. Firstly, we did not look into every possible search resource. To balance between the estimated workload and coverage, five electronic libraries were selected based on the existing SLR experiences (cf. Section 3.4.2). In fact, the statistics suggests that these five literature search engines may give a broad enough coverage of relevant studies (Zhang et al., 2011). Secondly, we unfolded automated search through titles, keywords and abstracts instead of full texts. On one hand, using a full text search usually leads to an explosion of search result. On the other hand, the search precision would be reduced quite dramatically by scanning full texts (Dieste et al., 2009). Thirdly, due to the known limitations of the search engines (Brereton et al., 2007), we also noticed and confirmed that the automated search missed important studies. To alleviate this issue, we supplemented a manual search by snowballing the references of the initially selected papers (cf. Section 3.4.4). Fourthly, it is possible that we may have not found the papers using irregular terms to describe Cloud services evaluation. In addition to carefully proposing the search string (cf. Section 3.4.3), similarly, we also resorted to the reference snowballing to further identify the possibly missed publications. Finally, we specified ten Cloud providers in the search string, which may result in bias when identifying the most common services and providers to answer RQ2. However, we had to adopt those search terms as a trade-off for improving the search string's sensitivity of the "commercial Cloud service"-related evaluation studies. Since the top ten Cloud providers were summarized by the third party from the industrial perspective, they can be viewed as weighted popular providers for this study. In fact, other Cloud providers were still able to be identified, such as BlueLock, EasticHosts, and Flexiant (cf. Section 5.2).

7.4. Reviewers reliability

As mentioned in Section 3.3, the detailed review work was implemented mainly by a PhD student to gain understanding of his research topic. Since the student is a new comer in the Cloud Computing domain, his misunderstanding of Cloud services evaluation may incur biased review process and results. To help ensure that the conduction of this SLR was as unbiased as possible, we adopted a supervisory strategy including three points: first, before planning this SLR, the supervisory panel instructed the PhD student to perform an ad hoc review of background knowledge covering Cloud Computing in general and Cloud services evaluation in particular; second, during planning this SLR, the expert panel was involved in helping develop a review protocol prior to conducting the review; third, every step of the conduction of this SLR was under close supervision including regular meetings, and all the unsure issues were further discussed with the expert panel. As such, we have tried our best to reduce the possible bias of the review conduction. However, when it comes to the data analysis, there might still be the possibility of incomplete findings or conclusions due to our personal interest and opinions.

7.5. Data extraction

During the process of data extraction from the reviewed studies, we found that not many papers specified sufficient details about the evaluation background, environment, and procedure, which could be partially reflected by the quality assessment. As a result, sometimes we had to infer certain information through some unclear clues, particularly when we tried to find the purpose or the time of particular evaluation experiments. Therefore, there may be some inaccuracies in the inferred data. However, this point can be considered as a limitation of the current primary studies instead of this SLR. Since the empirical research in Cloud services evaluation falls in the experimental computer science (Feitelson, 2007), we suggest that researchers may employ structural abstract (Budgen et al., 2008) and/or guidelines for conducting and reporting experiments or case studies (Runeson and Höst, 2009) to regulate their future evaluation work.

8. Conclusions and future work

Evaluation of commercial Cloud services has gradually become significant as an increasing number of competing Cloud providers emerge in industry (Prodan and Ostermann, 2009)[LYKZ10]. Given that the Cloud services evaluation is challenging and the existing studies are relatively chaotic, we adopted the SLR method to investigate the existing practices as evidence to outline the scope of Cloud services evaluation. The findings of this SLR lie in three aspects.

(1) The overall data collected in the SLR can lead us to become familiar with the sate-of-the-practice of evaluation of commercial Cloud services. In particular, the answers to those six research questions summarized the key details of the current evaluation implementations. Meanwhile, the summarized data, such as metrics, benchmarks, and experimental setup scenes, were arranged as a dictionary-like fashion for evaluators to facilitate future Cloud services evaluation work.
(2) Some of the findings have identified several research gaps in the area of Cloud services evaluation. First, although Elasticity and Security are significant features of commercial Cloud services, there seems a lack of effective and efficient means of evaluating the elasticity and security of a Cloud service. Our findings also suggest that this could be a long-term research challenge. Second, there is still a gap between practice and research into "real" Cloud evaluation benchmarks. On one hand, theoretical discussions considered that traditional benchmarks were insufficient for evaluating commercial Cloud services [BKKL09]. On the other hand, traditional benchmarks have been overwhelmingly used in the existing Cloud evaluation practices. The findings suggest that those traditional benchmarks will remain in the Cloud services evaluation work unless there is a dedicated Cloud benchmark. Third, the result of a quality assessment of the studies shows that the existing primary studies were not always conducted or reported appropriately. Thus, we suggest that future evaluation work should be regulated following particular guidelines (Budgen et al., 2008; Runeson and Höst, 2009).
(3) Some other findings suggest the trend of applying commercial Cloud services. In general, commercial Cloud Computing has attracted the attention of an increasing number of researchers, which can be confirmed by the world-widely increased research interests in the Cloud services evaluation topic. In addition to satisfying business requirements, commercial Cloud Computing is also regarded as a suitable paradigm to deal with scientific issues. As for specific commercial Cloud services,

although the competitive market changes rapidly, Amazon, Google and Microsoft currently supply the most popular Cloud services. Furthermore, PaaS and IaaS essentially supplement each other to satisfy various requirements in the Cloud market.

We also gained some lessons about conducting SLR from this work. Firstly, our practice has confirmed some previous experiences like the usage of pilot review from other SLR studies (Major et al., 2011; Babar and Zhang, 2009). In particular, future studies should carefully estimate the extra time and effort if considering an ad hoc review as the prerequisite of an SLR conduction. Secondly, our study also revealed new EBSE lesson – continuous collection of evidence for building knowledge base. In other words, for particular research topics, the employment of SLR could be worthy of a regular use to update the data or knowledge to support the research in those topics. In fact, given the initial understanding of Cloud services evaluation in this case, the current stage of this SLR tends to be a systematic mapping study, while the gradual update will accumulate the evaluation outcomes of more primary studies, and then help gain more knowledge.

Our future work will be unfolded in two directions. Firstly, the extracted data in this SLR will be structured and stored into a database for supporting a Cloud services evaluation methodology. Secondly, benefiting from the result of this SLR as a solid s tarting point, we will perform deeper study into Cloud service evaluation, such as developing sophisticated evaluation metrics.

Acknowledgements

We record our sincere thanks for Prof. Barbara Kitchenham's pertinent suggestions and comments that helped us improve the quality of this report.

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

Appendix A. Details of quality rating for primary studies

See Table A.12.

Appendix B. Brief description of the evaluated commercial Cloud services

See Table B.13.

Table B.13
Evaluated commercial Cloud services.

Cloud Provider	Cloud service	Brief description
Amazon	EBS (Elastic Block Store)	Amazon Elastic Block Store (EBS) provides block level storage volumes for use with Amazon EC2 instances.
	EC2 (Elastic Compute Cloud)	Amazon Elastic Compute Cloud (Amazon EC2) provides resizable compute capacity in the cloud.
	ELB (Elastic Load Balancing)	Elastic Load Balancing automatically distributes incoming application traffic across multiple Amazon EC2 instances.
	EMR (Elastic MapReduce)	Amazon Elastic MapReduce enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.
	FPS (Flexible Payment Service)	Amazon FPS is built on top of Amazon's payments infrastructure and provides developers with a convenient way to charge Amazon's tens of millions of customers.
	RDS (Rational Database Service)	Amazon Relational Database Service (Amazon RDS) is used to set up, operate, and scale a relational database in the cloud.
	S3 (Simple Storage Service)	Amazon S3 provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web.
	SimpleDB	Amazon SimpleDB is a non-relational data store that offloads the work of database administration.
	SQS (Simple Queueing System)	Amazon Simple Queue Service (Amazon SQS) offers a hosted queue for storing messages as they travel between computers.
BlueLock	BlueLock	Bluelock Virtual Datacenters are hosted in the public cloud and are based on VMware vCloud technology, which provides full compatibility with any VMware environment.
ElasticHosts	ElasticHosts	ElasticHosts supplies virtual servers running on server farms, located in five fully independent premier-class data centres across two continents.
Flexiant	FlexiScale	Flexible &Scalable Public Cloud Hosting is a pay-as-you-go public cloud platform offering on-demand, scalable hosting services.
GoGrid	GoGrid	GoGrid is a cloud infrastructure service, hosting Linux and Windows virtual machines managed by a multi-server control panel.
Google	AppEngine (Google App Engine)	Google AppEngine is a Cloud Computing platform for developing and hosting web applications in Google-managed data centres.
	Memcache	Memcache is a distributed memory object caching system, primarily intended for fast access to cached results of datastore queries.
	UrlFetch (URL Fetch)	UrlFetch allows scripts to communicate with other applications or access other resources on the web by fetching URLs.
IBM	IBM Cloud (Beta)	The beta version of Cloud Computing platform offered by IBM.
Microsoft	SQL Azure	Microsoft SQL Azure Database is a cloud database service built on SQL Server technologies.
	Windows Azure	Windows Azure is a cloud operating system that serves as a runtime for the applications and provides a set of services that allows development, management and hosting of applications off-premises.
Rackspace	CloudServers	CloudServers is a cloud infrastructure service that allows users to deploy "one to hundreds of cloud servers instantly" and create of "advanced, high availability architectures".
	CloudFiles	CloudFiles is a cloud storage service that provides "unlimited online storage and CDN" for media on a utility computing basis.

Table C.14
Explanation of the typically excluded papers.

Paper	Brief explanation	Corresponding exclusion criteria
[BCK+10]	The evaluation work is for the proposed AppScale Cloud platform.	(3)
[BKKL09]	Theoretical discussion about Cloud services evaluation.	(2)
[BLP11]	The evaluation work is for the proposed modeling approach, and it is in a private virtualized environment.	(1) & (3)
[CST+10]	Mostly theoretical discussion, and evaluation work is in a private environment.	(1) & (2)
[dAadCB09]	This is a previous version of [dAadCB10].	(4)
[EF09]	The evaluation work is done in the open-source Cloud.	(1)
[EKS+11]	Theoratical discussion based on the evaluation work in a private Cloud.	(1) & (2)
[GLMP09]	The evaluation work is for the proposed VBS system.	(3)
[GM11]	The evaluation work is done in the open-source Cloud.	(1)
[GSF11]	The evaluation work is done in the open-source Cloud.	(1)
[GWQF10a]	This is a previous version of [GWC+11].	(4)
[GWQF10b]	The evaluation work is for the proposed AzureMapReduce framework.	(3)
[HM11]	Theoretical discussion about autonomic benchmarking Cloud services.	(2)
[HvQHK11]	The evaluation work is done in a private virtualized environment.	(1)
[IYE10]	This is a previous version of [IYE11].	(4)
[JRRT10]	This is a previous version of [JMR+11].	(4)
[KC11]	This is a poster paper.	(5)
[KMKT11]	The evaluation work is in a private Cloud.	(1)
[LO08]	This work is for the proposed GridBatch with little evaluation.	(3)
[OIY+08]	This is a previous version of [OIY+09].	(4)
[OPF09]	The evaluation work is in a private Cloud.	(1)
[PPDK09]	The evaluation work is for the proposed Swarm framework.	(3)
[RS10]	The evaluation work is done in an academic Cloud: Qloud.	(1)
[Sch09]	The evaluation work is for the proposed MapReduce-based algorithm.	(3)
[Sha10]	The evaluation work is done in the open-source Cloud.	(1)
[SLYP10]	The evaluation work is done in a private virtualized environment.	(1)
[TFN11]	The evaluation work is for the proposed scheduling strategy.	(3)
[TUS11]	The evaluation work is done in a private virtualized environment.	(1)
[VBVB09]	The evaluation work is not on commercial Cloud services.	(1)
[WVX10]	This is a previous version of [WVX11].	(4)
[YTDG01]	Mainly a theoretical discussion about performance evaluation with fault recovery.	(2)

Appendix C. Explanation of the typically excluded papers

See Table C.14. We only show typical publications here instead of listing all the excluded studies. Most of the typically excluded papers were discussed in our group meetings. This appendix may be used as a clue for readers to further identify useful information.

Appendix D. Selected primary studies

[ADWC10] Mohammed Alhamad, Tharam Dillon, Chen Wu, and Elizabeth Chang. Response time for Cloud computing providers. In *Proceedings of the 12th International Conference on Information Integration and Web-based Applications &Services (iiWAS 2010)*, pp. 603–606, Paris, France, November 8–10 2010. ACM Press.

[AM10] Sayaka Akioka and Yoichi Muraoka. HPC benchmarks on Amazon EC2. In *Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications Workshops (WAINA 2010)*, pp. 1029–1034, Perth, Australia, April 20–23 2010. IEEE Computer Society.

[BCA11] Tekin Bicer, David Chiu, and Gagan Agrawal. MATE-EC2: A middleware for processing data with AWS. In *Proceedings of the 4th ACM International Workshop on Many-Task Computing on Grids and Supercomputers (MTAGS 2011)*, pp. 59–68, Seattle, Washington, USA, November 14 2011. ACM Press.

[BFG+08] Matthias Brantner, Daniela Florescu, David Graf, Donald Kossmann, and Tim Kraska. Building a database on S3. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD 2008)*, pp. 251–264, Vancouver, British Columbia, Canada, June 9–12 2008. ACM Press.

[BIN10] Paolo Bientinesi, Roman Iakymchuk, and Jeff Napper. HPC on competitive Cloud resources. In Borko Furht and Armando Escalante, editors, *Handbook of Cloud Computing*, chapter 21, pp. 493–516. Springer-Verlag, New York, NY, 2010.

[BK09] Christian Baun and Marcel Kunze. Performance measurement of a private Cloud in the OpenCirrus™ testbed. In *Proceedings of the 4th Workshop on Virtualization in High-Performance Cloud Computing (VHPC 2009)*, pp. 434–443, Delft, The Netherlands, August 25 2009. Springer-Verlag.

[BL10] Paul Brebner and Anna Liu. Performance and cost assessment of Cloud services. In *Proceedings of the 2010 International Conference on Service-Oriented Computing Workshops (PAASC 2010) in conjunction with 8th International Conference on Service Oriented Computing (ICSOC 2010)*, pp. 39–50, San Francisco, CA, USA, December 7–10 2010. Springer-Verlag.

[BS10] Sean Kenneth Barker and Prashant Shenoy. Empirical evaluation of latency-sensitive application performance in the Cloud. In *Proceedings of the 1st Annual ACM SIGMM Conference on Multimedia Systems (MMSys 2010)*, pp. 35–46, Scottsdale, Arizona, February 22–23 2010. ACM Press.

[BT11] David Bermbach and Stefan Tai. Eventual consistency: How soon is eventual? An evaluation of Amazon S3's consistency behavior. In *Proceedings of the 6th Workshop on Middleware for Service Oriented Computing (MW4SOC 2011)*, pp. 1–6, Lisboa, Portugal, December 12 2011. ACM Press.

[CA10] David Chiu and Gagan Agrawal. Evaluating caching and storage options on the Amazon Web services Cloud. In *Proceedings of the 11th IEEE/ACM International Conference on Grid Computing (GRID 2010)*, pp. 17–24, Brussels, Belgium, October 25–28 2010. IEEE Computer Society.

[CBH+11] Guang Chen, Xiaoying Bai, Xiaofei Huang, Muyang Li, and Lizhu Zhou. Evaluating services on the Cloud using ontology QoS model. In *Proceedings of the 6th IEEE International Symposium on Service Oriented System Engineering (SOSE 2011)*, pp. 312–317, Irving, CA, USA, December 12–14 2011. IEEE Computer Society.

[CHK+11] David Chiu, Travis Hall, Farhana Kabir, Apeksha Shetty, and Gagan Agrawal. Analyzing costs and optimizations for an elastic key-value store on Amazon Web services. *International Journal of Next-Generation Computing*, 1(2):1–21, July 2011.

[CHS10] Adam G. Carlyle, Stephen L. Harrell, and Preston M. Smith. Cost-effective HPC: The community or the Cloud? In *Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2010)*, pp. 169–176, Indianapolis, Indiana, USA, November 30–December 3 2010. IEEE Computer Society.

[CMS11] Matheus Cunha, Nabor Mendonça, and Américo Sampaio. Investigating the impact of deployment configuration and user demand on a social network application in the Amazon EC2 Cloud. In *Proceedings of the 3rd IEEE International Conference on Coud Computing Technology and Science (CloudCom 2011)*, pp. 746–751, Athens, Greece, November 29–December 1 2011. IEEE Computer Society.

[CRT+11] Javier Cerviño, Pedro Rodríguez, Irena Trajkovska, Alberto Mozo, and Joaquín Salvachúa. Testing a Cloud provider network for hybrid P2P and Cloud streaming architectures. In *Proceedings of the 4th International Conference on Cloud Computing (IEEE CLOUD 2011)*, pp. 356–363, Washington, DC, USA, July 4–9 2011. IEEE Computer Society.

[dAadCB10] Marcos Dias de Assunção, Alexandre di Costanzo, and Rajkumar Buyya. A cost–benefit analysis of using Cloud computing to extend the capacity of clusters. *Cluster Computing*, 13(3):335–347, September 2010.

[DDJ+10] Tolga Dalman, Tim Doernemann, Ernst Juhnke, Michael Weitzel, Matthew Smith, Wolfgang Wiechert, Katharina Noh, and Bernd Freisleben. Metabolic flux analysis in the Cloud. In *Proceedings of the 6th IEEE International Conference on e-Science (e-Science 2010)*, pp. 57–64, Brisbane, Australia, December 7–10 2010. IEEE Computer Society.

[DPhC09] Jiang Dejun, Guillaume Pierre, and Chi hung Chi. EC2 performance analysis for resource provisioning of service-oriented applications. In *Proceedings of the 7th International Conference on Service Oriented Computing (ICSOC-ServiceWave 2009)*, pp. 197–207, Stockholm, Sweden, November 23–27 2009. Springer-Verlag.

[DSL+08] Ewa Deelman, Gurmeet Singh, Miron Livny, Bruce Berriman, and John Good. The cost of doing science on the Cloud: The montage example. In *Proceedings of the 2008 International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2008)*, pp. 1–12, Austin, TX, November 15–21 2008. IEEE Computer Society.

[EH08] Constantinos Evangelinos and Chris N. Hill. Cloud computing for parallel scientific HPC applications: Feasibility of running coupled atmosphere-ocean climate models on Amazon' s EC2. In *Proceedings of the 1st Workshop on Cloud Computing and its Applications (CCA 2008)*, pp. 1–6, Chicago, IL, October 22–23 2008.

[EKKJP10] Yaakoub El-Khamra, Hyunjoo Kim, Shantenu Jha, and Manish Parashar. Exploring the performance fluctuations of HPC workloads on Clouds. In *Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2010)*, pp. 383–387, Indianapolis, Indiana, USA, November 30–December 3 2010. IEEE Computer Society.

[Gar07a] Simson L. Garfinkel. Commodity grid computing with Amazon' s S3 and EC2. *Usenix;Login*, 32(1):7–13, February 2007.

[Gar07b] Simson L. Garfinkel. An evaluation of Amazon' s grid computing services: EC2, S3, and SQS. Technical Report TR-08-07, Center for Research on Computation and Society, School for Engineering and Applied Sciences, Harvard University, Cambridge, MA, 2007.

[GBS11] Francis Gropengießer, Stephan Baumann, and Kai-Uwe Sattler. Cloudy transactions: Cooperative XML authoring on Amazon S3. In *Proceedings of the German Database Conference Datenbanksysteme für Business, Technologie und Web (BTW 2011)*, pp. 307–326, Kaiserslautern, Germany, March 2–4 2011. Bonner Köllen Verlag.

[GCR11] Devarshi Ghoshal, R. Shane Canon, and Lavanya Ramakrishnan. I/O performance of virtualized Cloud environments. In *Proceedings of the 2nd International Workshop on Data Intensive Computing in the Clouds (DataCloud-SC 2011)*, pp. 71–80, Seattle, Washington, USA, November 14 2011. ACM Press.

[GK11] Ian P. Gent and Lars Kotthoff. Reliability of computational experiments on virtualised hardware. In *Proceedings of the Workshops at the 25th AAAI Conference on Artificial Intelligence (2011 AAAI Workshop WS-11-08)*, pp. 8–10, San Francisco, California, USA, August 7 2011. AAAI Press.

[GS11] Francis Gropengießer and Kai-Uwe Sattler. Transactions a la carte – implementation and performance evaluation of transactional support on top of Amazon S3. In *Proceedings of the 25th IEEE International Symposium on Parallel and Distributed Processing Workshops and PhD Forum (IPDPSW 2011)*, pp. 1082–1091, Anchorage, Alaska, USA, May 16–20 2011. IEEE Computer Society.

[GWC+11] Thilina Gunarathne, Tak-Lon Wu, Jong Youl Choi, Seung-Hee Bae, and Judy Qiu. Cloud computing paradigms for pleasingly parallel biomedical applications. *Concurrency and Computation: Practice and Experience*, 23(17):2338–2354, December 2011.

[Haz08] Scott Hazelhurst. Scientific computing using virtual high-performance computing: A case study using the Amazon elastic computing Cloud. In *Proceedings of the 2008 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries: Riding the Wave of Technology (SAICSIT 2008)*, pp. 94–103, Wilderness, South Africa, October 6–8 2008. ACM Press.

[HH09] Zach Hill and Marty Humphrey. A quantitative analysis of high performance computing with Amazon' s EC2 infrastructure: The death of the local cluster? In *Proceedings of the 10th IEEEACM International Conference on Grid Computing (GRID 2009)*, pp. 26–33, Banff, Alberta, Canada, October 12–16 2009. IEEE Computer Society.

[HHJ+11] Marty Humphrey, Zach Hill, Keith Jackson, Catharine van Ingen, and Youngryel Ryu. Assessing the value of Cloudbursting: A case study of satellite image processing on Windows Azure. In *Proceedings of the 7th IEEE International Conference on eScience (eScience 2011)*, pp. 126–133, Stockholm, Sweden, December 5–8 2011. IEEE Computer Society.

[HLM+10] Zach Hill, Jie Li, Ming Mao, Arkaitz Ruiz-Alvarez, and Marty Humphrey. Early observations on the performance of Windows Azure. In *Proceedings of the 1st Workshop on Scientific Cloud Computing (ScienceCloud 2010) in conjunction with the 19th ACM International Symposium on High Performance Distributed Computing (HPDC 2010)*, pp. 367–376, ACM Press, June 21 2010. Chicago, Illinois, USA.

[HZK+10] Qiming He, Shujia Zhou, Ben Kobler, Dan Duffy, and Tom Mcglynn. Case study for running HPC applications in public Clouds. In *Proceedings of the 1st Workshop on Scientific Cloud Computing (ScienceCloud 2010) in conjunction with the 19th ACM International Symposium on High Performance Distributed Computing (HPDC 2010)*, pp. 395–401, Chicago, Illinois, USA, June 21 2010. ACM Press.

[ILFL11] Sadeka Islam, Kevin Lee, Alan Fekete, and Anna Liu. How a consumer can measure elasticity for Cloud platforms. Technical Report 680, School of Information Technologies, University of Sydney, Sydney, Australia, August 2011.

[INB11] Roman Iakymchuk, Jeff Napper, and Paolo Bientinesi. Improving high-performance computations on Clouds through resource underutilization. In *Proceedings of the 26th ACM Symposium on Applied Computing (SAC 2011)*, pp. 119–126, Taichung, Taiwan, March 21–25 2011. ACM Press.

[IOY+11] Alexandru Iosup, Simon Ostermann, M. Nezih Yigitbasi, Radu Prodan, Thomas Fahringer, and Dick H.J. Epema. Performance analysis of cloud computing services for many-tasks scientific computing. *IEEE Transactions on Parallel and Distributed Systems*, 22(6):931–945, June 2011.

[IYE11] Alexandru Iosup, Nezih Yigitbasi, and Dick Epema. On the performance variability of production Cloud services. In *Proceedings of the 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2011)*, pp. 104–113, Newport Beach, CA, USA, May 23–26 2011. IEEE Computer Society.

[JD11] Gideon Juve and Ewa Deelman. Scientific workflows in the Cloud. In Massimo Cafaro and Giovanni Aloisio, editors, *Grids, Clouds and Virtualization*, chapter 4, pp. 71–91. Springer-Verlag, London, UK, 2011.

[JDV+09] Gideon Juve, Ewa Deelman, Karan Vahi, Gaurang Mehta, Bruce Berriman, Benjamin P. Berman, and Phil Maechling. Scientific workflow applications on Amazon EC2. In *Proceedings of the 5th IEEE International Conference on E-Science Workshops (ESCIW 2009)*, pp. 59–66, Oxford, UK, December 9–11 2009. IEEE Computer Society.

[JDV+10] Gideon Juve, Ewa Deelman, Karan Vahi, Gaurang Mehta, Bruce Berriman, Benjamin P. Berman, and Phil Maechling. Data sharing options for scientific workflows on Amazon EC2. In *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2010)*, pp. 1–9, New Orleans, LA, November 13–19 2010. IEEE Computer Society.

[JMR+11] Keith R. Jackson, Krishna Muriki, Lavanya Ramakrishnan, Karl J. Runge, and Rollin C. Thomas. Performance and cost analysis of the Supernova factory on the Amazon AWS Cloud. *Scientific Programming – Science-Driven Cloud Computing*, 19(2–3):107–119, April 2011.

[JMW+11] Deepal Jayasinghe, Simon Malkowski, Qingyang Wang, Jack Li, Pengcheng Xiong, and Calton Pu. Variations in performance and scalability when migrating n-tier applications to different Clouds. In *Proceedings of the 4th International Conference on Cloud Computing (IEEE CLOUD 2011)*, pp. 73–80, Washington, DC, USA, July 4–9 2011. IEEE Computer Society.

[JRM+10] Keith R. Jackson, Lavanya Ramakrishnan, Krishna Muriki, Shane Canon, Shreyas Cholia, John Shalf, Harvey J. Wasserman, and Nicholas J. Wright. Performance analysis of high performance computing applications on the Amazon Web services Cloud. In *Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2010)*, pp. 159–168, Indianapolis, Indiana, USA, November 30–December 3 2010. IEEE Computer Society.

[KJM+09] Derrick Kondo, Bahman Javadi, Paul Malecot, Franck Cappello, and David P. Anderson. Cost–benefit analysis of Cloud computing versus desktop grids. In *Proceedings of the 23rd IEEE International Symposium on Parallel and Distributed Processing (IPDPS 2009)*, pp. 1–12, Rome, Italy, May 23–29 2009. IEEE Computer Society.

[KKL10] Donald Kossmann, Tim Kraska, and Simon Loesing. An evaluation of alternative architectures for transaction processing in the Cloud. In *Proceedings of the 2010 International Conference on Management of Data (SIGMOD 2010)*, pp. 579–590, Indianapolis, Indiana, USA, June 6–11 2010. ACM Press.

[LHvI+10] Jie Li, Marty Humphrey, Catharine van Ingen, Deb Agarwal, Keith Jackson, and Youngryel Ryu. eScience in the Cloud: A MODIS satellite data reprojection and reduction pipeline in the Windows Azure platform. In *Proceedings of the 24th IEEE International Symposium on Parallel and Distributed Processing (IPDPS 2010)*, pp. 1–10, Atlanta, Georgia, USA, April 19–23 2010. IEEE Computer Society.

[LJ10] André Luckow and Shantenu Jha. Abstractions for loosely coupled and ensemble-based simulations on Azure. In *Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2010)*, pp. 550–556, Indianapolis, Indiana, USA, November 30–December 3 2010. IEEE Computer Society.

[LJB10] Wei Lu, Jared Jackson, and Roger Barga. AzureBlast: A case study of developing science applications on the Cloud. In *Proceedings of the 1st Workshop on Scientific Cloud Computing (ScienceCloud 2010) in conjunction with the 19th ACM International Symposium on High Performance Distributed Computing (HPDC 2010)*, pp. 413–420, Chicago, Illinois, USA, June 21 2010. ACM Press.

[LML+11] Alexander Lenk, Michael Menzel, Johannes Lipsky, Stefan Tai, and Philipp Offermann. What are you paying for? Performance benchmarking for Infrastructure-as-a-Service. In *Proceedings of the 4th International Conference on Cloud Computing (IEEE CLOUD 2011)*, pp. 484–491, Washington, DC, USA, July 4–9 2011. IEEE Computer Society.

[LW09] Huan Liu and Sewook Wee. Web server farm in the Cloud: Performance evaluation and dynamic architecture. In *Proceedings of the 1st International Conference on Cloud Computing (CloudCom 2009)*, pp. 369–380, Beijing, China, December 1–4 2009. Springer-Verlag.

[LYKZ10] Ang Li, Xiaowei Yang, Srikanth Kandula, and Ming Zhang. CloudCmp: Comparing public Cloud providers. In *Proceedings of the 10th Annual Conference on Internet Measurement (IMC 2010)*, pp. 1–14, Melbourne, Australia, November 1–3 2010. ACM Press.

[LZZ+11] Mingliang Liu, Jidong Zhai, Yan Zhai, Xiaosong Ma, and Wenguang Chen. One optimized I/O configuration per HPC application: Leveraging the configurability of Cloud. In *Proceedings of the 2nd ACM SIGOPS Asia-Pacific Workshop on Systems (APSys 2011)*, pp. 1–5, Shanghai, China, July 11–12 2011. ACM Press.

[MF10] Raffaele Montella and Ian Foster. Using hybrid Grid/Cloud computing technologies for environmental data elastic storage, processing, and provisioning. In Borko Furht and Armando Escalante, editors, *Handbook of Cloud Computing*, chapter 26, pp. 595–618. Springer-Verlag, New York, NY, 2010.

[MVML11] Rafael Moreno-Vozmediano, Ruben S. Montero, and Ignacio M. Llorente. Multicloud deployment of computing clusters for loosely coupled MTC applications. *IEEE Transactions on Parallel and Distributed Systems*, 22(6):924–930, June 2011.

[NB09] Jeffrey Napper and Paolo Bientinesi. Can Cloud computing reach the top500? In *Proceedings of the Combined Workshops on UnConventional High Performance Computing Workshop plus Memory Access Workshop (UCHPC-MAW 2009)*, pp. 17–20, Ischia, Italy, May 18–20 2009. ACM Press.

[OIY+09] Simon Ostermann, Alexandru Iosup, Nezih Yigitbasi, Radu Prodan, Thomas Fahringer, and Dick Epema. A performance analysis of EC2 Cloud computing services for scientific computing. In *Proceedings of the 1st International Conference on Cloud Computing (CloudComp 2009)*, pp. 115–131, Munich, Germany, October 19–21 2009. Springer-Verlag.

[PEP11] Stephen C. Phillips, Vegard Engen, and Juri Papay. Snow white Clouds and the seven dwarfs. In *Proceedings of the 3rd IEEE International Conference on Coud Computing Technology and Science (CloudCom 2011)*, pp. 738–745, Athens, Greece, November 29–December 1 2011. IEEE Computer Society.

[PIRG08] Mayur R. Palankar, Adriana Iamnitchi, Matei Ripeanu, and Simson Garfinkel. Amazon S3 for science grids: A viable solution? In *Proceedings of the 2008 International Workshop on Data-Aware Distributed Computing (DADC 2008)*, pp. 55–64, Boston, MA, June 23–27 2008. ACM Press.

[RD11] Radhika Ramasahayam and Ralph Deters. Is the Cloud the answer to scalability of ecologies? Using GAE to enable horizontal scalability. In *Proceedings of the 5th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2011)*, pp. 317–323, Daejeon, Korea, May 31–June 3 2011. IEEE Computer Society.

[RSP11] Mark Redekopp, Yogesh Simmhan, and Viktor K. Prasanna. Performance analysis of vertex-centric graph algorithms on the Azure Cloud platform. In *Proceedings of the Workshop on Parallel Algorithms and Software for Analysis of Massive Graphs (ParGraph 2011) in conjunction with the 18th IEEE International Conference on High Performance Computing (HiPC 2011)*, pp. 1–8, Bangalore, India, December 18 2011.

[RTSS09] Thomas Ristenpart, Eran Tromer, Hovav Shacham, and Stefan Savage. Hey, you, get off of my Cloud: Exploring information leakage in third-party compute Clouds. In *Proceedings of the 2009 ACM Conference on Computer and Communications Security (CCS 2009)*, pp. 199–212, Chicago, Illinois, USA, November 9–13 2009. ACM Press.

[RVG+10] J.J. Rehr, F.D. Vila, J.P. Gardner, L. Svec, and M. Prange. Scientific computing in the Cloud. *Computing in Science &Engineering*, 12(3):34–43, May–June 2010.

[SASA+11] K. Salah, M. Al-Saba, M. Akhdhor, O. Shaaban, and M. I. Buhari. Performance evaluation of popular Cloud IaaS providers. In *Proceedings of the 6th International Conference on Internet Technology and Secured Transactions (ICITST 2011)*, pp. 345–349, Abu Dhabi, United Arab Emirates, December 11–14 2011. IEEE Computer Society.

[SDQR10] Jörg Schad, Jens Dittrich, and Jorge-Arnulfo Quiané-Ruiz. Runtime measurements in the Cloud: Observing, analyzing, and reducing variance. *Proceedings of the VLDB Endowment*, 3(1–2):460–471, September 2010.

[SKP+11] Florian Schatz, Sven Koschnicke, Niklas Paulsen, Christoph Starke, and Manfred Schimmler. Mpi performance analysis of Amazon EC2 Cloud services for high performance computing. In *Proceedings of the 1st International Conference on Advances in Computing and Communications (ACC 2011)*, pp. 371–381, Kochi, Kerala, India, July 22–24 2011. Springer-Verlag.

[SMW+11] Vedaprakash Subramanian, Hongyi Ma, Liqiang Wang, En-Jui Lee, and Po Chen. Rapid 3D seismic source inversion using Windows Azure and Amazon EC2. In *Proceedings of the 7th IEEE 2011 World Congress on Services (SERVICES 2011)*, pp. 602–606, Washington, DC, USA, July 4–9 2011. IEEE Computer Society.

[SSS+08] Will Sobel, Shanti Subramanyam, Akara Sucharitakul, Jimmy Nguyen, Hubert Wong, Arthur Klepchukov, Sheetal Patil, Armando Fox, and David Patterson. Cloudstone: Multi-platform, multi-language benchmark and measurement tools for Web 2.0. In *Proceedings of the 1st Workshop on Cloud Computing and its Applications (CCA 2008)*, pp. 1–6, Chicago, IL, USA, October 22–23 2008.

[Sta09] Vladimir Stantchev. Performance evaluation of Cloud computing offerings. In *Proceedings of the 3rd International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP 2009)*, pp. 187–192, Sliema, Malta, October 11–16 2009. IEEE Computer Society.

[TCM+11] John J. Tran, Luca Cinquini, Chris A. Mattmann, Paul A. Zimdars, David T. Cuddy, Kon S. Leung, Oh-Ig Kwoun, Dan Crichton, and Dana Freeborn. Evaluating Cloud computing in the NASA DESDynI ground data system. In *Proceedings of the 2nd International Workshop on Software Engineering for Cloud Computing (SECLOUD 2011)*, pp. 36–42, Waikiki, Honolulu, HI, USA, May 22 2011. ACM Press.

[TYO10] Shiori Toyoshima, Saneyasu Yamaguchi, and Masato Oguchi. Storage access optimization with virtual machine migration and basic performance analysis of Amazon EC2. In *Proceedings of the 24th IEEE International Conference on Advanced Information Networking and Applications Workshops (WAINA 2010)*, pp. 905–910, Perth, Australia, April 20–23 2010. IEEE Computer Society.

[VDG11] Nikos Virvilis, Stelios Dritsas, and Dimitris Gritzalis. Secure Cloud storage: Available infrastructures and architectures review and evaluation. In *Proceedings of the 8th International Conference on Trust, Privacy &Security in Digital Business (TrustBus 2011)*, pp. 74–85, Toulouse, France, August 29–September 2 2011. Sprinter-Verlag.

[VJDR11] Jens-Sönke Vöckler, Gideon Juve, Ewa Deelman, and Mats Rynge. Experiences using Cloud computing for a scientific workflow application. In *Proceedings of the 2nd Workshop on Scientific Cloud Computing (ScienceCloud 2011) in conjunction with the 20th International ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC 2011)*, pp. 15–24, San Jose, California, USA, June 8 2011. ACM Press.

[VPB09] Christian Vecchiola, Suraj Pandey, and Rajkumar Buyya. High-performance Cloud computing: A view of scientific applications. In *Proceedings of the 10th International Symposium on Pervasive Systems, Algorithms, and Networks (I-SPAN 2009)*, pp. 4–16, Kaohsiung, Taiwan, December 14–16 2009. IEEE Computer Society.

[Wal08] Edward Walker. Benchmarking Amazon EC2 for high-performance scientific computing. *Usenix;Login*, 33(5):18–23, October 2008.

[WKF+10] Dennis P. Wall, Parul Kudtarkar, Vincent A. Fusaro, Rimma Pivovarov, Prasad Patil, and Peter J. Tonellato. Cloud computing for comparative genomics. *BMC Bioinformatics*, 11(259):1–12, May 2010.

[WN10] Guohui Wang and T. S. Eugene Ng. The impact of virtualization on network performance of Amazon EC2 data center. In *Proceedings of the 29th Conference on Computer Communications (IEEE INFOCOM 2010)*, pp. 1–9, San Diego, CA, March 14–19 2010. IEEE Communications Society.

[WVX11] Jian-Zong Wang, Peter Varman, and Chang-Sheng Xie. Optimizing storage performance in public Cloud platforms. *Journal of Zhejiang University-SCIENCE C (Computers &Electronics)*, 12(12):951, 964 2011.

[WWDM09] Jared Wilkening, Andreas Wilke, Narayan Desai, and Folker Meyer. Using Clouds for metagenomics: A case study. In *Proceedings of the 2009 IEEE International Conference on Cluster Computing and Workshops (CLUSTER 2009)*, pp. 1–6, New Orleans, Louisiana, USA, August 31–September 4 2009. IEEE Computer Society.

[YIEO09] Nezih Yigitbasi, Alexandru Iosup, Dick Epema, and Simon 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 (CCGRID 2009)*, pp. 472–477, Shanghai, China, May 18–21 2009. IEEE Computer Society.

[ZG11] Peter Zaspel and Michael Griebel. Massively parallel fluid simulations on Amazon' s HPC Cloud. In *Proceedings of the IEEE First International Symposium on Network Cloud Computing and Applications (IEEE NCCA 2011)*, pp. 73–78, Toulouse, France, November 21–23 2011. IEEE Computer Society.

[ZLK10] Liang Zhao, Anna Liu, and Jacky Keung. Evaluating Cloud platform architecture with the CARE framework. In *Proceedings of the 17th Asia Pacific Software Engineering Conference (APSEC 2010)*, pp. 60–69, Sydney, Australia, November 30–December 3 2010. IEEE Computer Society.

[ZLZ+11] Yan Zhai, Mingliang Liu, Jidong Zhai, Xiaosong Ma, and Wenguang Chen. Cloud versus in-house cluster: Evaluating Amazon cluster compute instances for running MPI applications. In *Proceedings of the 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2011)*, pp. 1–10, Seattle, Washington, USA, November 12–18 2011. ACM Press.

Appendix E. Typically excluded primary studies

[BCK+10] Chris Bunch, Navraj Chohan, Chandra Krintz, Jovan Chohan, Jonathan Kupferman, Puneet Lakhina, Yiming Li, and Yoshihide Nomura. An evaluation of distributed datastores using the AppScale Cloud platform. In *Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing (CLOUD '10)*, pp. 305–312, Miami, Florida, USA, July 5–10 2010. IEEE Computer Society.

[BKKL09] Carsten Binnig, Donald Kossmann, Tim Kraska, and Simon Loesing. How is the weather tomorrow?: Towards a benchmark for the Cloud. In *Proceedings of the Second International Workshop on Testing Database Systems (DBTest 2009)*, pp. 1–6, Providence, USA, June 29 2009. ACM Press.

[BLP11] Dario Bruneo, Francesco Longo, and Antonio Puliafito. Evaluating energy consumption in a Cloud infrastructure. In *Proceedings of the 2011 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WOWMOM 2011)*, pp. 1–6, Lucca, Italy, June 20–24 2011. IEEE Computer Society.

[CST+10] Brian F. Cooper, Adam Silberstein, Erwin Tam, Raghu Ramakrishnan, and Russell Sears. Benchmarking Cloud serving systems with YCSB. In *Proceedings of the 1st ACM symposium on Cloud computing (SoCC '10)*, pp. 143–154, Indianapolis, Indiana, USA, June 10–11 2010. ACM Press.

[dAadCB09] M. D. de Assunção, A. di Costanzo, and Rajkumar Buyya. Evaluating the cost–benefit of using Cloud computing to extend the capacity of clusters. In *Proceedings of the 18th ACM International Symposium on High Performance Distributed Computing (HPDC 2009)*, pp. 141–150, Munich, Germany, June 11–13 2009. ACM Press.

[EF09] Jaliya Ekanayake and Geoffrey Fox. High performance parallel computing with Clouds and Cloud technologies. In *Proceedings of the 1st International Conference on Cloud Computing (CloudComp 2009)*, pp. 20–38, Munich, Germany, October 19–21 2009. Springer-Verlag.

[EKS+11] Åke Edlund, Maarten Koopmans, Zeeshan Ali Shah, Ilja Livenson, Frederik Orellana, Niels Bohr, Jukka Kommeri, Miika Tuisku, Pekka Lehtovuori, Klaus Marius Hansen, Helmut Neukirchen, and Ebba Hvannberg. Practical Cloud evaluation from a nordic eScience user perspective. In *Proceedings of the 5th International Workshop on Virtualization Technologies in Distributed Computing (VTDC 2011) in conjunction with the 20th International ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC 2011)*, pp. 29–38, San José, USA, June 8 2011. ACM Press.

[GLMP09] Xiaoming Gao, Mike Lowe, Yu Ma, and Marlon Pierce. Supporting Cloud computing with the virtual block store system. In *Proceedings of the 5th IEEE International Conference on e-Science (e-Science 2009)*, pp. 208–215, Oxford, UK, December 9–11 2009. IEEE Computer Society.

[GM11] Abhishek Gupta and Dejan Milojicic. Evaluation of HPC applications on Cloud. In *Proceedings of the 5th International Event on Open Cirrus Summit (OCS 2011)*, pp. 22–26, Moscow, Russia, June 1–3 2011. IEEE Computer Society.

[GSF11] Pablo Graubner, Matthias Schmidt, and Bernd Freisleben. Energy-efficient management of virtual machines in Eucalyptus. In *Proceedings of the 4th International Conference on Cloud Computing (IEEE CLOUD 2011)*, pp. 243–250, Washington, DC, USA, July 4–9 2011. IEEE Computer Society.

[GWQF10a] Thilina Gunarathne, Tak-Lon Wu, Judy Qiu, and Geoffrey Fox. Cloud computing paradigms for pleasingly parallel biomedical applications. In *Proceedings of the 1st Workshop on Emerging Computational Methods for the Life Sciences (ECMLS 2010) in conjunction with the 19th ACM International Symposium on High Performance Distributed Computing (HPDC 2010)*, pp. 421–429, Chicago, Illinois, USA, June 21 2010. ACM Press.

[GWQF10b] Thilina Gunarathne, Tak-Lon Wu, Judy Qiu, and Geoffrey Fox. MapReduce in the Clouds for science. In *Proceedings of the 2010 IEEE 2nd International Conference on Cloud Computing Technology and Science (CloudCom '10)*, pp. 565–572, Indianapolis, Indiana, USA, November 30–December 3 2010. IEEE Computer Society.

[HM11] Steffen Haak and Michael Menzel. Autonomic benchmarking for Cloud infrastructures: An economic optimization model. In *Proceedings of the 1st ACM/IEEE Workshop on Autonomic Computing in Economics (ACE 2011) in conjunction with the 8th International Conference on Autonomic Computing (ICAC 2011)*, pp. 27–32, Karlsruhe, Germany, June 14 2011. ACM Press.

[HvQHK11] Nikolaus Huber, Marcel von Quast, Michael Hauck, and Samuel Kounev. Evaluating and modeling virtualization performance overhead for Cloud environments. In *Proceedings of the 1st International Conference on Cloud Computing and Services Science (CLOSER 2011)*, pp. 563–573, Noordwijkerhout, The Netherlands, May 7–9 2011. SciTePress.

[IYE10] Alexandru Iosup, Nezih Yigitbasi, and Dick Epema. On the performance variability of production Cloud services. Parallel and Distributed Systems Report Series PDS-2010-002, Delft University of Technology, Delft, Netherlands, January 2010.

[JRRT10] Keith R. Jackson, Lavanya Ramakrishnan, Karl J. Runge, and Rollin C. Thomas. Seeking supernovae in the Clouds: A performance study. In *Proceedings of the 1st Workshop on Scientific Cloud Computing (ScienceCloud 2010) in conjunction with the 19th ACM International Symposium on High Performance Distributed Computing (HPDC 2010)*, pp. 421–429, Chicago, Illinois, USA, June 21 2010. ACM Press.

[KC11] Pankaj Deep Kaur and Inderveer Chana. Evaluating Cloud platforms – an application perspective. In *Proceedings of the 2011 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM 2011)*, pp. 449–453, Budapest, Hungary, July 3–7 2011. Springer-Verlag.

[KMKT11] Kenji Kobayashi, Shunsuke Mikami, Hiroki Kimura, and Osamu Tatebe. The Gfarm file system on compute Clouds. In *Proceedings of the 25th IEEE International Symposium on Parallel and Distributed Processing (IPDPS 2011)*, pp. 1034–1041, Anchorage, Alaska, USA, May 16–20 2011. IEEE Computer Society.

[LO08] Huan Liu and Dan Orban. GridBatch: Cloud computing for large-scale data-intensive batch applications. In *Proceedings of the 8th IEEE International Symposium on Cluster Computing and the Grid (CCGRID 2008)*, pp. 295–305, Lyon, France, May 19–22 2008. IEEE Computer Society.

[OIY+08] Simon Ostermann, Alexandru Iosup, Nezih Yigitbasi, Radu Prodan, Thomas Fahringer, and Dick Epema. An early performance analysis of Cloud computing services for scientific computing. Parallel and Distributed Systems Report Series PDS-2008-006, Delft University of Technology, Delft, Netherlands, December 2008.

[OPF09] Simon Ostermann, Radu Prodan, and Thomas Fahringer. Extending grids with Cloud resource management for scientific computing. In *Proceedings of the 10th IEEEACM International Conference on Grid Computing (GRID 2009)*, pp. 42–49, Banff, Alberta, Canada, October 12-16 2009. IEEE Computer Society.

[PPDK09] Sangmi Lee Pallickara, Marlon Pierce, Qunfeng Dong, and Chinhua Kong. Enabling large scale scientific computations for expressed sequence tag sequencing over grid and Cloud computing clusters. In *Proceedings of the 8th International Conference on Parallel Processing and Applied Mathematics (PPAM 2009)*, pp. 13–16, Wroclaw, Poland, September 13–16 2009.

[RS10] M. Suhail Rehman and Majd F. Sakr. Initial findings for provisioning variation in Cloud computing. In *Proceedings of the 2nd IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2010)*, pp. 473–479, Indianapolis, Indiana, USA, November 30–December 3 2010. IEEE Computer Society.

[Sch09] Michael C. Schatz. CloudBurst: Highly sensitive read mapping with MapReduce. *Bioinformatics*, 25(11):1363–1369, April 2009.

[Sha10] Jeffrey Shafer. I/O virtualization bottlenecks in Cloud computing today. In *Proceedings of the 2nd Workshop on I/O Virtualization (WIOV 2010) in conjunction with the 15th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2010) and the 2010 ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments (VEE 2010)*, pp. 1–7, Pittsburgh, PA, USA, March 13 2010. USENIX Association.

[SLYP10] Sankaran Sivathanu, Ling Liu, Mei Yiduo, and Xing Pu. Storage management in virtualized Cloud environment. In *Proceedings of the 2010 IEEE 3rd International Conference on Cloud Computing (CLOUD 2010)*, pp. 204–211, Miami, Florida, USA, July 5–10 2010. IEEE Computer Society.

[TFN11] Gabriela Turcu, Ian Foster, and Svetlozar Nestorov. Reshaping text data for efficient processing on Amazon EC2. *Scientific Programming – Science-Driven Cloud Computing*, 19(2–3):133–145, April 2011.

[TUS11] Byung Chul Tak, Bhuvan Urgaonkar, and Anand Sivasubramaniam. To move or not to move: The economics of Cloud computing. In *Proceedings of the 3rd USENIX Conference on Hot Topics in Cloud Computing (HotCloud 2011)*, pp. 1–6, Portland, OR, USA, June 14–15 2011. USENIX Association.

[VBVB09] William Voorsluys, James Brokerg, Srikumar Venugopal, and Rajkumar Buyya. Cost of virtual machine live migration in Clouds: A performance evaluation. In *Proceedings of the 1st International Conference on Cloud Computing (CloudCom '09)*, pp. 254–265, Beijing, China, December 1–4 2009. Springer-Verlag.

[WVX10] Jianzong Wang, Peter Varman, and Changsheng Xie. Avoiding performance fluctuation in Cloud storage. In *Proceedings of the 2010 International Conference on High Performance Computing (HiPC 2010)*, pp. 1–9, Goa, India, December 19–22 2010. IEEE Computer Society.

[YTDG01] Bo Yang, Feng Tan, Yuan-Shun Dai, and Suchang Guo. Performance evaluation of Cloud service considering fault recovery. In *Proceedings of the 1st International Conference on Cloud Computing (CloudCom 2009)*, pp. 571–576, Beijing, China, December 1–4 2001. Springer-Verlag.

References

Alesandre, G.D., 2011. Updated App Engine pricing FAQ!, https://groups.google.com/forum/#!msg/google-appengine/Hluog1_a3n4/uFMhaBWhVi8J

Ali, M.S., Babar, M.A., Chen, L., Stol, K.J., 2010]. A systematic review of comparative evidence of aspect-oriented programming. Information and Software Technology 52, 871–887.

Amazon, 2011. High performance computing (HPC) on AWS, http://aws.amazon.com/hpc-applications/

Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M., 2010]. A view of Cloud computing. Communications of the ACM 53, 50–58.

Babar, M.A., Zhang, H., 2009]. Systematic literature reviews in software engineering: preliminary results from interviews with researchers. In: Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement (ESEM 2009), IEEE Computer Society, Lake Buena Vista, Florida, USA, pp. 346–355.

Brereton, P., Kitchenham, B.A., Budgen, D., Turner, M., Khalil, M., 2007]. Lessons from applying the systematic literature review process within the software engineering domain. Journal of Systems and Software 80, 571–583.

Brooks, C., 2010. Cloud computing benchmarks on the rise, http://searchcloudcomputing.techtarget.com/news/1514547/Cloud-computing-benchmarks-on-the-rise

Budgen, D., Kitchenham, B.A., Charters, S.M., Turner, M., Brereton, P., Linkman, S.G., 2008]. Presenting software engineering results using structured abstracts: a randomised experiment. Empirical Software Engineering 13, 435–468.

Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J., Brandic, I., 2009]. Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems 25, 599–616.

Dieste, O., Grimán, A., Juristo, N., 2009]. Developing search strategies for detecting relevant experiments. Empirical Software Engineering 14, 513–539.

Dybå, T., Kitchenham, B.A., Jørgensen, M., 2005]. Evidence-based software engineering for practitioners. IEEE Software 22, 58–65.

Feitelson, D.G., 2007]. Experimental computer science. Communications of the ACM 50, 24–26.

Ferdman, M., Adileh, A., Kocberber, O., Volos, S., Alisafaee, M., Jevdjic, D., Kaynak, C., Popescu, A.D., Ailamaki, A., Falsafi, B., 2012]. Clearing the Clouds: a study of emerging scale-out workloads on modern hardware. In: Proceedings of the 17th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2012), ACM Press, London, England, UK, pp. 37–48.

Foster, I., Zhao, Y., Raicu, I., Lu, S., 2008]. Cloud computing and Grid computing 360-degree compared. In: Proceedings of the Workshop on Grid Computing Environments (GCE08) in conjunction with the 2008 International Conference for High Performance Computing, Networking, Storage and Analysis (SC 2008), IEEE Computer Society, Austin, TX, pp. 1–10.

Google, 2012. Google Compute Engine, http://cloud.google.com/products/compute-engine.html

Habib, S.M., Ries, S., Mühlhäuser, M., 2010]. Cloud computing landscape and research challenges regarding trust and reputation. In: Proceedings of the 2010 Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing (UIC-ATC 2010), IEEE Computer Society, Xi'an, China, pp. 410–415.

Harris, D., 2011. Watch out, world: IBM finally offers a real Cloud, http://gigaom.com/cloud/watch-out-world-ibm-finally-offers-a-real-cloud/

Harris, D., 2012. What google compute engine means for cloud computing, http://gigaom.com/cloud/what-google-compute-engine-means-for-cloud-computing/

Islam, S., Lee, K., Fekete, A., Liu, A., 2012]. How a consumer can measure elasticity for Cloud platforms. In: Proceedings of the 3rd joint WOSP/SIPEW International Conference on Performance Engineering (ICPE 2012), ACM Press, Boston, USA, pp. 85–96.

Jain, R.K., 1991]. The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling. Wiley Computer Publishing, John Wiley & Sons, Inc., New York, NY.
Kitchenham, B.A., Charters, S., 2007]. Guidelines for Performing Systematic Literature Reviews in Software Engineering. In: Technical Report EBSE 2007-001. Keele University and Durham University Joint Report.
Kitchenham, B.A., Li, Z., Burn, A., 2011]. Validating search processes in systematic literature reviews. In: Proceedings of the 1st International Workshop on Evidential Assessment of Software Technologies (EAST 2011) in conjunction with ENASE 2011, SciTePress, Beijing, China, pp. 3–9.
Kossmann, D., Kraska, T., 2010]. Data management in the Cloud: promises, state-of-the-art, and open questions. Datenbank Spektr 10, 121–129.
Lewis, B.C., Crews, A.E., 1985]. The evolution of benchmarking as a computer performance evaluation technique. MIS Quarterly 9, 7–16.
Li, Z., O'Brien, L., Cai, R., Zhang, H., 2012a]. Towards a taxonomy of performance evaluation of commercial Cloud services. In: Proceedings of the 5th IEEE International Conference on Cloud Computing (CLOUD 2012), IEEE Computer Society, Honolulu, Hawaii, USA, pp. 344–351.
Li, Z., O'Brien, L., Zhang, H., Cai, R., 2012b]. A factor framework for experimental design for performance evaluation of commercial cloud services. In: Proceedings of the 4th IEEE International Conference on Cloud Computing Technology and Science (CloudCom 2012), IEEE Computer Society, Taipei, Taiwan, pp. 169–176.
Li, Z., O'Brien, L., Zhang, H., Cai, R., 2012c]. On a catalogue of metrics for evaluating commercial Cloud services. In: Proceedings of the 13th ACM/IEEE International Conference on Grid Computing (GRID 2012), IEEE Computer Society, Beijing, China, pp. 164–173.
Lisboa, L.B., Garcia, V.C., Lucrédio, D., de Almeida, E.S., de Lemos Meira, S.R., de Mattos Fortes, R.P., 2010]. A systematic review of domain analysis tools. Information and Software Technology 52, 1–13.
Major, L., Kyriacou, T., Brereton, O.P., 2011]. Systematic literature review: teaching novices programming using robots. In: Proceedings of the 15th Annual Conference on Evaluation and Assessment in Software Engineering (EASE 2011), IEEE Computer Society, Durham, UK, pp. 21–30.
Miller, R., 2011. A look inside Amazon's data centers, http://www.datacenterknowledge.com/archives/2011/06/09/a-look-inside-amazons-data-centers/
Montgomery, D.C., 2009]. Design and Analysis of Experiments, 7th ed. John Wiley & Sons, Inc., Hoboken, NJ.
Obaidat, M.S., Boudriga, N.A., 2010]. Fundamentals of Performance Evaluation of Computer and Telecommjnication Systems. John Wiley & Sons, Inc., Hoboken, New Jersey.
Prodan, R., Ostermann, S., 2009]. A survey and taxonomy of Infrastructure as a Service and Web hosting Cloud providers. In: Proceedings of the 10th IEEE/ACM International Conference on Grid Computing (GRID 2009), IEEE Computer Society, Banff, Alberta, Canada, pp. 17–25.
Rimal, B.P., Choi, E., Lumb, I., 2009]. A texonomy and survey of Cloud computing systems. In: Proceedings of the 5th International Joint Conference on INC, IMS and IDC (NCM 2009), IEEE Computer Society, Seoul, Korea, pp. 44–51.
Runeson, P., Höst, M., 2009]. Guidelines for conducting and reporting case study research in software engineering. Empirical Software Engineering 14, 131–164.
SearchCloudComputing, 2010. Top 10 Cloud computing providers of 2010. http://searchcloudcomputing.techtarget.com/feature/Top-10-cloud-computing-providers
Stokes, J., 2011. The PC is order, the Cloud is chaos, http://www.wired.com/insights/2011/12/the-pc-is-order/
Tran, V.T.K., Lee, K., Fekete, A., Liu, A., Keung, J., 2011]. Size estimation of Cloud migration projects with Cloud Migration Point (CMP). In: Proceedings of the 5th International Symposium on Empirical Software Engineering and Measurement (ESEM 2011), IEEE Computer Society, Banff, Canada, pp. 265–274.
Zhang, H., Ali Babar, M., 2010]. On searching relevant studies in software engineering. In: 14th International Conference on Evaluation and Assessment in Software Engineering (EASE'10), BCS, Keele, England.
Zhang, H., Babar, M.A., 2011]. An empirical investigation of systematic reviews in software engineering. In: Proceedings of the 5th International Symposium on Empirical Software Engineering and Measurement (ESEM 2011), IEEE Computer Society, Banff, Canada, pp. 1–10.
Zhang, H., Babar, M.A., Tell, P., 2011]. Identifying relevant studies in software engineering. Information and Software Technology 53, 625–637.
Zhang, Q., Cheng, L., Boutaba, R., 2010]. Cloud computing: State-of-the-art and research challenges. Journal of Internet Services and Applications 1, 7–18.

Zheng Li received his M.E. by research from the University of New South Wales. He is now a PhD candidate in the School of Computer Science at Australian National University, and a graduate researcher with the Software Systems Research Group at NICTA. He is the author of more than 15 journal and conference publications. His research interests include empirical software engineering, software cost/effort estimation, machine learning, web service composition, and Cloud computing.

He Zhang is a professor of software engineering in the Software Institute at Nanjing University, China. He joined academia after 7 years in industry, developing software systems in the areas of aerospace and complex data management. He has published 70+ peer-reviewed research papers in international journals, conferences, and workshops. His current research areas include software & systems process modeling and simulation, process enactment analysis and process improvement, software engineering for embedded systems, cloud computing, evidence-based and empirical software engineering. Dr. Zhang received his PhD in computer science from the University of New South Wales.

Liam O'Brien is solution architect at Geoscience Australia. He is involved in architecting solutions for several of Geoscience Australia's major projects including the Information Platform for Bio-regional Assessments and the CO_2 Infrastructure Assessment Project. Before joining Geoscience Australia, he was the Chief Software Architect for CSIRO's eResearch Program. He is also the Vice-President of the Service Science Society. His research interests also include software and service oriented architecture, reengineering, business transformation, enterprise architectures and cloud computing. He holds a Ph.D in computer science and a B.Sc from the University of Limerick, Ireland and is a member of the IEEE and IEEE Computer Society.

Rainbow Cai received her PhD in software engineering from the University of Auckland in 2009. Her main research interests include: Cloud platform performance evaluation, automated software engineering, software architecture modelling and performance evaluation, and model driven engineering. She is currently leading the metadata system development to unify the research data, people data, and other enterprise data of the Australian National University.

Shayne Flint is a senior lecturer in the Research School of Computer Science at Australian National University. After a 17 year career as a practicing engineer, he completed a PhD at the ANU in 2006. He now develops methodologies and tools for generating radical improvements in software development productivity, quality and satisfaction of stakeholder needs. He takes an inter-disciplinary approach to his research and works closely with industry, government and scientific communities.