IaaS Cloud Benchmarking: Approaches, Challenges, and Experience

(Invited Paper)

Alexandru Iosup*, Radu Prodan†, and Dick Epema*
*Parallel and Distributed Systems, Delft University of Technology, Delft, the Netherlands. Contact: A.Iosup@tudelft.nl
†Parallel and Distributed Systems, University of Innsbruck, Innsbruck, Austria. Contact: Radu@dps.uibk.ac.at

***Abstract*—Infrastructure-as-a-Service (IaaS) cloud computing is an emerging commercial infrastructure paradigm under which clients (users) can lease resources when and for how long needed, under a cost model that reflects the actual usage of resources by the client. For IaaS clouds to become mainstream technology and for current cost models to become more client-friendly, benchmarking and comparing the non-functional system properties of various IaaS clouds is important, especially for the cloud users. In this article we focus on the IaaS cloud-specific elements of benchmarking, from a user's perspective. We propose a generic approach for IaaS cloud benchmarking, discuss numerous challenges in developing this approach, and summarize our experience towards benchmarking IaaS clouds. We argue for an experimental approach that requires, among others, new techniques for experiment compression, new benchmarking methods that go beyond blackbox and isolated-user testing, new benchmark designs that are domain-specific, and new metrics for elasticity and variability.**

***Index Terms*—Cloud computing, Benchmarking, Distributed Systems, Distributed applications, Performance evaluation, Metrics/Measurement, Performance measures.**

I. Introduction

Infrastructure-as-a-Service (IaaS) clouds are becoming a rich and active branch of commercial ICT services. Users of IaaS clouds can provision "processing, storage, networks, and other fundamental resources" [1] on-demand, that is, when needed, for as long as needed, and paying only for what is actually consumed. For the past five years, commercial IaaS clouds such as Amazon's EC2 have gained an increasing user base, from small and medium businesses [2] to scientific HPC users [3], [4]. However, the increased adoption of clouds and perhaps even the pricing models depend on the ability of (prospective) cloud users to benchmark and compare commercial cloud services. In this article, we investigate the IaaS cloud-specific elements of benchmarking from the user perspective.

An important characteristic of IaaS clouds is good performance, which needs to be ensured on-demand and sustained when needed over a long period of time. However, as we have witnessed happening with several other new technologies while still in their infancy, notably with grid computing, we believe IaaS clouds may also undergo a period of inconsistent performance management.

Benchmarking is a traditional approach to verify that the performance of a system meets the requirements. When benchmarking results are published, for example through mixed consumer-provider organizations such as SPEC and TPC, the consumers can easily compare products and put pressure on the providers to use best-practices and perhaps lower costs. At the moment, the use of clouds is fragmented across many different application areas, such as hosting applications, media, games, and web sites, E-commerce, On-Demand Workforce and CRM, high-performance computing, search, and raw resources for various usage. Each application area has its own (de facto) performance standards that have to be met by commercial clouds, and some have even developed benchmarks (e.g., BioBench for Bioinformatics and RUBiS for online business).

For IaaS clouds, we conjecture that the probable characteristics of current and near-future workloads can be derived from three major trends emerging from the last decade of grid and large-scale computing. First, individual jobs are now predominantly split into smaller compute or data-intensive tasks (many tasks [5]); there are almost no tightly coupled parallel jobs. Second, the duration of individual tasks is diminishing with every year; few tasks are still running for longer than one hour and a majority require only a few minutes to complete. Third, compute-intensive jobs are split either into bags-of-tasks (BoTs) or DAG-based workflows, but data-intensive jobs may use a variety of programming models, from MapReduce to general dataflow.

Cloud benchmarking is not a straightforward application of older benchmarking techniques. In the past, there have been several large-scale computing environments that have similarities with clouds. Already decades ago, such institutes as CERN and the IBM T.J. Watson Research Center had large numbers of mainframes (using virtualization through the Virtual Machine operating system!) that also used multi-tenancy across their departments. Similarly, some vendors had large-scale installations for paid use by customers through Remote Job Entry facilities. In these environments, benchmarking and capacity planning were performed in close collaboration between owners and customers. A big difference, and advantage, for customers wishing to benchmark their prospective computing environments is that they can simply use access by credit card to deploy and benchmark their applications in the cloud: clouds do not only offer elasticity on demand, they also offer (resources for) capacity planning and benchmarking on demand. The new challenge is that customers will have to gain, through benchmarking, sufficient trust in the performance, the

elasticity, the stability, and the resilience of clouds to rely on them for the operation of their businesses. As a matter of fact, they may want to do this both when migrating to the cloud, and on a continuous basis to assess the operation of their applications in the cloud. Thus, of great importance is the ability of cloud benchmarks to allow users to gain trust without requiring long setups and costly operation.

We argue in this work for a focused, community-based approach to IaaS cloud benchmarking in which the main challenges are jointly identified, and best-practice and experiences can be easily shared. Although we have seen in the past few years numerous approaches to benchmarking and performance evaluation of various systems, there is no unified view of the main challenges facing researchers and practitioners in the field of benchmarking. Our work, which should help with system procurement and performance management, aims at providing this unified view. In this sense, this work follows the earlier efforts on benchmarking middleware [6], [7], on benchmarking databases [8], on the performance evaluation of grid and parallel-system schedulers [9]–[12], and on benchmarking systems in general [13], [14]. Towards this end, our main contribution is threefold:

1) We introduce a generic approach for IaaS cloud benchmarking (Section III).
2) We discuss numerous challenges in developing our and other approaches for cloud benchmarking (Section IV). We focus on methodological, system-, workload-, and metrics-related issues.
3) We summarize our experience towards benchmarking IaaS clouds (Section V). We summarize two initiatives of the SPEC Research Group and its Cloud Working Group, of which some of the authors are members. We also summarize our experience with building models and tools that can become useful building blocks for IaaS cloud benchmarking.

II. A Primer on Benchmarking Computer Systems

We review in this section the main reasons for benchmarking and the main elements of the typical benchmarking process, which are basically unchanged since the early 1990s. For more detail, we refer to canonical texts on benchmarking [8] and performance evaluation [13] of computer systems.

A. Why Benchmarking?

Benchmarking computer systems is the process of evaluating their performance and other non-functional characteristics with the purpose of comparing them with other systems or with industry-agreed standards. Traditionally, the main use of benchmarking has been to facilitate the informed procurement of computer systems through the publication of verifiable results by system vendors and third-parties. However, benchmarking has grown as a support process for several other situations, which we review in the following.

Use in system design, tuning, and operation: Benchmarking has been shown to increase pressure on vendors to design better systems, as has been for example the experience of the TPC-D benchmark [8, Ch.3, Sec.IV]. For this benchmark, insisting on the use of SQL has driven the wide acceptance of the ANSI SQL-92; furthermore, the complexity of a majority of the queries has lead to numerous improvements in the design of aggregate functions and support for them. This benchmark also led to a wide adoption of the geometric mean for aggregating normalized results [14]. The tuning of the DAS multi-cluster system has benefited from the benchmarking activity of some of the authors of this article in the mid-2000s [15]; then, our distributed computing benchmarks exposed various (fixable) problems of the in-operation system.

Use in training: One of the important impediments in the adoption of a new technology is the lack of expertise of potential users. Market shortages of qualified personnel in computer science are a major cause of concern for the European Union and the US. Benchmarks, through their open-source nature and representation of industry-accepted standards, can represent best-practices and thus be valuable training material.

On alternatives to benchmarking: Several alternative methods have been used for the purposes described earlier in this section, among them empirical performance evaluation, simulation, and even mathematical analysis. We view benchmarking as an empirical evaluation of performance that follows a set of accepted procedures and best-practices. Thus, the use of empirical performance evaluation is valuable, but perhaps without the representativeness of a (de facto) standard benchmark. We see a role for (statistical) simulation [16]–[18] and mathematical analysis when the behavior of the system is well-understood and for long-running evaluations that would be impractical otherwise. However, simulating new technology, such as cloud computing, requires careful (and time-consuming) validation of assumptions and models.

B. Elements of Benchmarking

Inspired by canonical texts [8], [13], we review here the main elements of a benchmarking process. The main requirements of a benchmark—relevance, portability, scalability, and simplicity—have been discussed extensively in related literature, for example in [8, Ch.1].

The *System Under Test (SUT)* is the system that is being evaluated. A *white box* system exposes its full operation, whereas a *black box* system does not expose operational details and is evaluated only through its outputs.

The *workload* is the operational load to which the SUT is subjected. Starting from the empirical observation that "20% of the code consumes 80% of the resources", simple *microbenchmarks* (*kernel benchmarks* [8, Ch.9]) are simplified or reduced-size codes designed to stress potential system bottlenecks. Using the methodology of Saavedra et al. [19] and later refinements such as Sharkawi et al. [20], the results of microbenchmarks can be combined with application profiles to provide credible performance predictions for any platform. *Synthetic* and even *real-world (complex) applications* are also used for benchmarking purposes, as a response to system improvements that make microbenchmarks run fast but do not affect the performance of much larger codes. For distributed

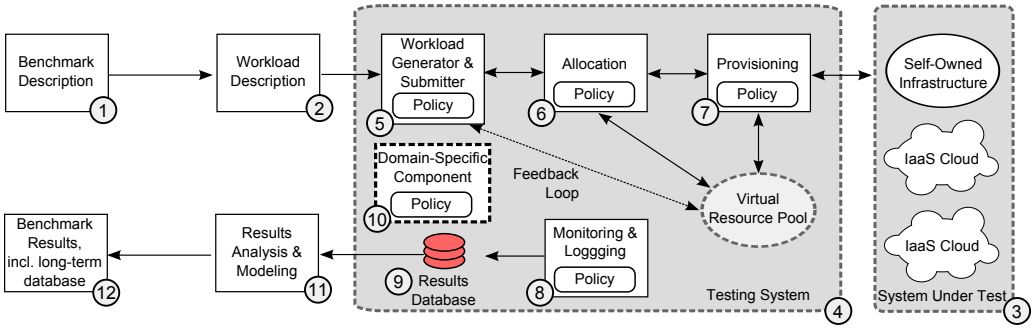

Fig. 1. Overview of our generic architecture for IaaS cloud benchmarking.

and large-scale systems such as IaaS clouds, *simple workloads* comprised of a single application and a (realistic) job arrival process represent better the typical system load and have been used for benchmarking [15]. *Complex workloads*, that is, the combined simple workloads of multiple users, possibly with different applications and job characteristics, have started to be used in the evaluation of distributed systems [15], [21]; we see an important role for them in benchmarking.

III. A Generic Architecture for IaaS Cloud Benchmarking

We propose in this section a generic architecture for IaaS cloud benchmarking. Our architecture focuses on conducting benchmarks as sets of (real-world) experiments that lead to results with high statistical confidence, on considering and evaluating IaaS clouds as evolving black-box systems, on employing complex workloads that represent multi-tenancy scenarios, on domain-specific scenarios, and on a combination of traditional and cloud-specific metrics.

We introduce in Section IV the main challenges that need to be addressed for our architecture to be realizable. In Section V-B, we discuss a partial implementation of this architecture that has already achieved good results in practice [21].

A. Overview

Our main design principle is to adapt the proven designs for benchmarking to IaaS clouds at scale. Thus, we design an architecture that builds on our GrenchMark framework for grid benchmarking [15], as presented in Figure 1.

The *Benchmarking Process* consists of the set of rules, prior knowledge (invariants), and procedures used to subject the SUT to the benchmark workload, and to collect and report the results. In our architecture, the process begins with the user (e.g., a prospective IaaS cloud user) defining the benchmark configuration, that is, the complex workloads that define the user's preferred scenario (component 1 in Figure 1). The scenario may focus on processing as much of the workload as possible during a fixed test period or on processing a fixed-size workload as quickly or cheaply as possible. The benchmarking system converts (component 2) the scenario into a set of workload descriptions, one per (repeated) execution. The workload may be defined before the benchmarking process, or change (in particular, increase) during the benchmarking process. To increase the statistical confidence in obtained results, subjecting the SUT to a workload may be *repeated* or the workload may be *long-running*.

After the preparation of the workload, the SUT (component 3 in Figure 1) is subjected to the workload through the job and resource management services provided by the testing system (component 4, which includes components 5–10). In our benchmarking architecture, the SUT can be comprised of one or several self-owned infrastructures, and public and private IaaS clouds. The SUT provides resources for the execution of the workload; these resources are grouped into a *Virtual Resource Pool*. The results produced during the operation of the system may be used to provide a *feedback loop* from the Virtual Resource Pool back into the Workload Generator and Submitter (component 5); thus, our architecture can implement open and closed feedback loops [22].

As a last important sequence of process steps, per-experiment results are combined into higher-level aggregates, first aggregates per workload execution (component 11 in Figure 1), then aggregates per benchmark (component 12). The reporting of metrics should try to avoid the common pitfalls of performance evaluation; see for example [14], [23]. For large-scale distributed systems, it is particularly important to report not only the basic statistics, but also some of the outliers, and full distributions or at least the higher percentiles of the distribution (95-th, 99-th, etc.). We also envision the creation of a general database of results collected by the entire community and shared freely. The organization and operation of such a database is beyond the scope of this article.

B. Distinguishing Design Features

We present in the remainder of this section several of the distinguishing features of this architecture.

In comparison with traditional grid environments, commercial IaaS clouds do not provide services for managing the incoming stream of requests (components 5, 6, and 8 in Figure 1) or the resources leased from the cloud (components 7 and 8). Our architecture supports various policies for provisioning and allocation of resources (components 6 and 7, respectively). In contrast to GrenchMark, our generic cloud-benchmarking architecture also includes support for evolving black-box systems (components 9, 11, and 12), complex

workloads and multi-tenancy scenarios (components 1, 2, and 5), domain-specific components (component 10), etc.

Experiments conducted on large-scale infrastructure should be designed to minimize the time spent effectively using resources. The interplay between components 1, 2, and 5 in Figure 1 can play a non-trivial role in resolving this challenge, through automatic selection and refinement of complex test workloads that balance the trade-off between accuracy of results and benchmark cost; the main element in a dynamic tuning of this trade-off is the policy present in component 5. The same interplay enables multi-tenancy benchmarks.

Several of the possible SUTs expose complete or partial operational information, acting as white or partially white boxes. Our architecture allows exploiting this information, combining results from black-box and white-box testing. Moreover, the presence of the increasingly higher-level aggregations (components 11 and 12 in Figure 1) permits both the long-term evaluation of the system, and the combination of short-term and long-term results. The policy for monitoring and logging in component 8 allows the user to customize what information is processed and stored in the results database. We conclude our architecture goes far beyond simple black-box testing.

Supports domain-specific bechmarks is twofold in our architecture. First, components 5–7 support complex workloads and feedback loops, and policy-based resource and job management. Second, we include in our architecture a domain-specific component (component 10) that can be useful in supporting cloud programming models such as the compute-intensive workflows and bags-of-tasks, and the data-intensive MapReduce and Pregel. The policy element in component 10 allows this component to play a dynamic, intelligent role in the benchmarking process.

IV. Open Challenges in IaaS Cloud Benchmarking

We introduce in this section an open list of surmountable challenges in IaaS cloud benchmarking.

A. Methodological

Challenge 1. Experiment compression: Long setup times, for example of over a day, and/or long periods of continuous evaluation, for example of more than a day per result, reduce the usefulness of a benchmark for the general user. This is a general problem with any experimental approach, but for IaaS clouds it has the added disadvantage of greatly and visibly increasing the cost of benchmarking. We argue that research is needed to reduce the setup and operational time of benchmarks for IaaS clouds. This can be achieved through reduced input and application sets, a clever setup of the experiments, and sharing of results across the community. We also envision the use of combined experimental approaches, in which real-world experiments are combined with emulation [24], [25] or simulation. Our vision for experiment compression represents an extension of the concept of statistical simulation [16]–[18], which has been used for computer architecture studies, to real-world experimentation.

Reduced benchmark input and application sets can be obtained by refining input workloads from real complex workloads, using theoretically sound methods (e.g., statistical models and goodness-of-fit tests). Such reduced benchmark inputs will contrast with traditional synthetic benchmarks, which incorporate many human-friendly parameter values (e.g., “10% queries of type A, 90% queries of type B”) and thus may lack theoretical guarantees for representativeness.

Challenge 2. Beyond black-box testing through testing short-term dynamics and long-term evolution: Similarly to multi-cluster grids, which frequently added clusters or individual nodes to the distributed infrastructure, clouds are continuously extended and tuned by their operators. Moreover, commercial clouds such as Amazon EC2 add frequently new functionality to their systems. Thus, the benchmarking results obtained at any given time may be unrepresentative for the future behavior of the system. We argue that IaaS clouds should not be benchmarked only using traditional black-box and even white-box testing, for which the system under test does not change in size and functionality, but also through new benchmarking methods that evaluate the impact of short-term dynamics and long-term evolution. Specifically, short-term dynamics characterize system changes occurring over short periods (at most hours), and long-term evolution characterizes system changes occurring over long periods (months, years).

A straightforward approach to benchmark both short-term dynamics and long-term evolution is to measure the system under test periodically, with judiciously chosen frequencies [26]. However, this approach increases the pressure of the so-far unresolved Challenge 1.

Challenge 3. Impact of middleware: IaaS clouds are built on several layers of middleware, from the guest operating system of the VM to the data-center resource manager. Each of these layers adds new complexity to testing and possibly also visible or invisible performance bottlenecks. One of the key issues in benchmarking IaaS clouds is to measure the performance of each layer of the middleware in isolation. We argue that a solution for this problem may not be possible under the current assumption of black-box testing, and propose instead to focus on a new methodology that accounts for imprecision in the isolation of root causes of performance.

We believe that good steps towards understanding the performance of various middleware layers can be and have already been taken [7], for example in assessing the impact of virtualization, but that more work is needed to reconcile the results (the situation presented in Challenge 2, where IaaS clouds change over time, may be a source of conflicting experimental results). We have surveyed in our previous work [26], [27] over ten performance studies that use common benchmarks to assess the virtualization overhead on computation (5–15%), I/O (10–30%), and HPC kernels (results vary). We have shown in a recent study of four commercial IaaS clouds [27] that virtualized resources obtained from public clouds can have a much lower performance than the theoretical peak, possibly because of the performance of the middleware layer.

B. System Properties

Challenge 4. Reliability, availability, and related system properties: One of the factors affecting the behavior of large-scale systems is the presence of failures, which are likely inevitable at scale. We have found endemic presence of failures in many popular large-scale systems, from grids [28] to DNS and other distributed services [29]. Benchmarking reliability and related systems properties is difficult, not in the least because of Challenge 2.

Challenge 5. Massive scale, multi-site benchmarking: One of the main product features of IaaS clouds is the promise of seemingly infinite capacity. We argue that benchmarking this promise is difficult, very time-consuming, and very costly. We have seen in our previous work that testing tools can be built to test infrastructures of thousands of cores [15], but performance evaluation tools that work at much larger scale in heterogeneous IaaS clouds have yet to be proven in practice. An important challenge here may be the ability to generate massive-scale workloads.

We have already had experience with companies building *hybrid clouds* [1] out of their own infrastructure and resources leased from IaaS clouds (this process is also referred to as *cloudbursting*, for example by Microsoft). Other cloud deployment models require the use of multiple sites, for reliability and vendor lock-in avoidance. We expect multi-site cloud use to increase, as more companies, with or without existing computational capacity, try out or even decide to use cloud services. We argue that benchmarking across multiple sites raises additional challenges, not in the least the combined availability for testing and scalability of the infrastructure, and the increased cost.

Challenge 6. Performance isolation: The negative effects of the interaction between running jobs in a complex workload have been observed in distributed environments since at least the mid-1990s [30]. Following early work [31], [32], we argue that quantifying the level of isolation provided by an IaaS cloud is a new and important challenge.

Moreover, as IaaS clouds become more international, their ability to isolate performance may suffer most during periods of peak activity. Thus, studying the time patterns of performance isolation is worthwhile.

C. Workload

Challenge 7. Statistical models of workloads or of system performance: Statistical workload modeling is the general technique of producing synthetic models from workload traces collected from real-world systems that are statistically similar to the real-world traces. We argue that building such statistical models raises important challenges, from data collection to trace processing, from finding good models to testing the validity of the models. We also see as an open challenge the derivation of statistical performance models, perhaps through linear regression, from already existing measurements.

We envision that IaaS clouds will also be built for specific, even niche application domains, charging premium rates for the expertise required to run specific classes of applications. This is similar to the appearance of domain-specific grids, such as BioGrid, in the early 2000s; and of domain-specific database-related technology, such as transaction-processing and data warehousing solutions, in the early 1990s [8, Ch.1]. We argue that IaaS cloud benchmarking should begin with domain-specific benchmarks, before transiting to general benchmarks.

Toward building domain-specific benchmarks, we argue for building statistical models of domain-specific or at least programming model-specific workloads. We have conducted in the past extensive research in grid workloads [33], with results in modeling BoTs [34], and in characterizing scientific and engineering workflows [33]. Several studies [35]–[39], including our own study of four large MapReduce clusters [40], have focused on characterizing workloads of MapReduce, which is one of the most popular programming models for data processing in the loud. Open challenges in this context are the formulation of realistic models for workflows, MapReduce, and other programming models for data processing. We also find that the many-task programming model [5] is worthwhile for investigation in this context. We also refer to a recent survey of challenges associated with large-scale log analysis [41].

Challenge 8. Benchmarking performance isolation under different multi-tenancy models: Unlike traditional system benchmarking, where interference of different elements that affect performance—multiple users competing for resources, stressing multiple system resources at the same time—is generally avoided, the expected cloud workload is complex. We argue that for IaaS clouds interference should be expected and benchmarked. Specific focus for this challenge, as an extension of Challenge 8, is to benchmark under a specific multi-tenancy model, from the shared-nothing approach of multi-cluster grids, to shared-hardware and shared-virtualized machine approaches prevalent in today's commercial clouds [42], [43], and possibly others.

D. Metrics

Challenge 9. Beyond traditional performance: Traditional performance metrics—such as utilization, throughput, and makespan—have been defined for statically-sized, homogeneous systems. We have raised in our previous work [12] the challenge of adapting these metrics for distributed on-demand systems, such as the contemporary multi-cluster grids and commercial IaaS clouds. IaaS clouds raise new challenges in defining cloud-related metrics, such as elasticity [44], [45].

We also argue for revisiting the analysis of results and their refinement into metrics. For example, due to their change over time and imperfect performance isolation, IaaS clouds may require revisiting the concept of variability, way beyond the traditional mean (or median) and standard deviation. Our preliminary work [26] on the variability of performance in IaaS and other types of clouds indicates that variability can be high and may vary with time.

Traditionally, system warm-up is excluded from performance evaluation, leaving only the steady-state period of the system for study. However, especially for hybrid and other

multi-site cloud architectures, we argue for the need to also measure the transitional period that occurs when a significant fraction of the system resources are in the process of being leased or released.

Challenge 10. The cost issue: Although cost models were discussed in benchmarking and performance evaluation of both databases and grids, a variety of issues have not been addressed. Specifically, the sub-leasing cost model used in today's commercial IaaS clouds (e.g., Amazon's "spot" instances) provides a new focus. It is also unclear how to define costs for a hybrid cloud infrastructure, especially when the performance of the cloud does not match the expectation [27], [46]. Last but not least, it is unclear how to define the source of budgets, for example either infrastructural or operational funds, a situation which affects a variety of economic metrics. Early approaches exist [3], [4].

V. Experience towards IaaS Cloud Benchmarking

A. Methodology: the SPEC Cloud Working Group

The SPEC Research Group[1] (RG) is a new group within the Standard Performance Evaluation Corporation (SPEC). Among other activities, the SPEC RG facilitates the interaction between academia and industry by co-organizing the Joint ACM/SPEC International Conference on Performance Engineering (ICPE). The Cloud Working Group[2] (CWG) is a branch of the SPEC RG that aims to develop the methodological aspects of cloud benchmarking (**Challenges 1–3** in Section IV). In this section we summarize two initiatives of the SPEC RG and CWG.

Beyond traditional performance: Traditional performance metrics such as utilization and normalized schedule length [47] have been defined for statically sized systems. Redefining these metrics for dynamic systems, especially in the context of black-box resources leased from clouds, is a topic of interest for the CWG (**Challenges 5 and 6**). Beyond performance, the CWG is also interested in other non-functional metrics, such as elasticity, utility, performance isolation, and dependability (**Challenges 4, 9, and 15**).

Reproducibility of experiments: (orthogonal to our challenges) Being able to reproduce experimental results is critical for the validity and lifetime of obtained results. However, this goal is difficult to achieve when the system under test is complex, dynamic, or large-scale; IaaS clouds have all these characteristics. A recent initiative of the RG is to build a repository[3] that can be used to share experimental results, setups, and other meta-data. Moreover, the call for papers issued by ICPE 2013 includes a focus on reproducibility of experiments.

B. SkyMark: A Framework for IaaS Cloud Benchmarking

We have recently implemented a part of the architecture described in Section III as our SkyMark tool for IaaS cloud benchmarking [48]. SkyMark already implements two of the distinguishing features of our architecture (see Section III-B). First, SkyMark provide services for managing the incoming stream of requests (jobs) and the resources leased from the cloud [21]. For the former, SkyMark provides single or multiple job queues, depending on the configuration of the experiment, and each queue supports a variety of simple scheduling policies (e.g., FCFS). For the latter, SkyMark supports several static and dynamic resource provisioning policies.

Second, SkyMark supports complex workloads (**Challenge 7**). Workloads are split into units. Each unit is defined by the characteristic resource to be stressed (e.g., through CPU-intensive jobs), the job arrival pattern (one of uniform, increasing, and bursty), and the job durations. SkyMark is able, for a given target configuration, to generate workloads that lead to a user-specified average utilization in the absence of system overheads.

Using SkyMark, we were able [21] to benchmark three IaaS clouds, including Amazon EC2. We have used in out benchmarks six provisioning policies and three allocation policies, with provisioning and allocation policies considered either independently or together. We were also able [48] to evaluate, for our OpenNebula private clouds, the interference occurring in various multi-tenancy scenarios (**Challenge 8**).

C. Real-World Evaluation of IaaS Cloud Performance

Several of the challenges we formulated in Section IV are the outcome of our previous research conducted from the past three years in benchmarking and understanding the performance of several cloud infrastructures. We summarize in the following some of our main results that motivated this classification.

Challenge 2: We have observed the long-term evolution in performance of clouds since 2007. Then, the acquisition of one EC2 cloud resource took an average time of 50 seconds, and constantly increased to 64 seconds in 2008 and 78 seconds in 2009. The EU S3 service shows pronounced daily patterns with lower transfer rates during night hours (7PM to 2AM), while the US S3 service exhibits a yearly pattern with lowest mean performance during the months January, September, and October. Other services have occasional decreases in performance, such as SDB in March 2009, which later steadily recovered until December [26]. Finally, EC2 spot prices typically follow a long-term step function [49].

Challenge 3: Depending on the provider and its middleware abstraction, several cloud overheads and performance metrics can have different interpretation and meaning. In IaaS clouds, resource acquisition is typically the sum of the installation time and boot times, and for Amazon EC2 has a stable value in the order of minutes [27]. Other IaaS providers, such as GoGrid, behave similarly to grids and offer highly variable resource acquisition times, i.e., one order magnitude higher than EC2. In contrast, the Google App Engine (GAE), which offers a higher-level PaaS abstraction, defines the acquisition overhead as the time between the issue of a HTTP request until the HTTP response is returned; the overhead of GAE

[1] http://research.spec.org/

[2] http://research.spec.org/working-groups/rg-cloud-working-group.html

[3] ICPE Organizers, Reproducibility repository approved, http://icpe2013.ipd.kit.edu/news/single_view/article/reproducibility-repository-approved/.

is in the order of seconds [50], an order of magnitude lower than for EC2. The performance interpretations and differences can have similarly high variations depending on the middleware. The black-box execution approach in IaaS clouds of externally-compiled software encapsulated in VMs generates high degradations from the expected peak performance, up to six to eight times lower than the theoretical maximum of Amazon's "Elastic Compute Unit" (ECU, 4.4 GOPS) [27]. Parallel computing-wise, the performance of today's IaaS is below the theoretical peak of today's dedicated parallel supercomputers even for demanding conveniently parallel applications by 60-70%. Furthermore, benchmarking the sustained performance of other infrastructures such as GAE is almost prohibited by the sandboxed environment that completely hides the underlying hardware on which the instance is started with no user control, raising the need for **Challenge 6** [50].

Challenge 4: With respect to reliability, the payment models and compensations in case of resource failures make clouds a more promising platform than traditional distributed systems, especially grids. Interesting from the reliability point of view are the EC2 spot instances that allow customers to bid on unused capacity and run those instances for as long as their bid exceeds the current spot price. Our analysis on this risk-reward problem between January 2011 and February 2012 demonstrates that spot instances may represent a cheaper but still reliable solution offering up to 99% availability provided that users make slightly generous bids, such as $0.35 for `m1.large` instances [49].

Challenge 9: Regarding the importance of system warmup, an interesting case is the modern just-in-time (JIT) compilations of Java application running on GAE infrastructure which can boost the performance of interpreted Java byte code by a factor of four in a predictable manner (from the third request onwards in case of GAE) [50].

Challenge 10: The variety of cost models combined with performance variability makes the cloud provider selection a difficult problem for the cloud user. For example, our analysis in [50] shows that computing costs are lower on GAE than in EC2 for very short jobs, mostly due to the cycle-based payment granularity, as opposed to the hourly billing intervals of EC2. The cost model may also vary within one provider. For example, the EC2 reserved instances are cheaper than standard instances if their usage is of about 50% for for one year reservations, and of about 30% for three year reservations [49]. In contrast, spot instances on EC2 may represent a 60% cheaper but equally reliable alternative to standard instances provided that a correct user bet is made [49].

D. Statistical Workload Models

Challenge 7: In our previous work, starting from multi-cluster grid traces, we have proposed statistical models of BoTs [34], and characterized BoTs [33], [34] and workflows [33]. We found, notably, that BoTs are the dominant programming model for compute-intensive workloads in grids—they account for 80-90% of both number of tasks and resource consumption. We have recently characterized and modeled statistically MapReduce workloads, starting from four traces of large clusters, including Google's [40].

E. Open Data: Several Useful Archives

Challenge 7: Workload and operational trace archives are an important tool in developing benchmarks. Although IaaS clouds are new, several online archives could already provide interesting data.

General workload traces for parallel systems and multi-cluster grid are provided by the Parallel Workloads Archive [51] and the Grid Workloads Archive [52], respectively. For an example of domain-specific workload traces, the Game Trace Archive [53] publishes data representative for online gaming.

For operational traces, the Failure Trace Archive [29] and the P2P Trace Archive [54] provide operational information about general and domain-specific (peer-to-peer) distributed systems.

VI. Conclusion

The importance of IaaS cloud benchmarking has grown proportionally to the increased adoption of this technology, from small and medium businesses to scientific HPC users. In contrast to the fragmented field of today, we argue in this work for a more focused approach to IaaS benchmarking, in which the community can join into identifying the main challenges, and then share best-practices and experiences. Such an approach would greatly benefit (prospective) cloud users with system procurement and performance management.

We propose a generic approach for IaaS cloud benchmarking, in which resource and job management can be provided by the testing infrastructure, there is support for black-box systems that change rapidly and can evolve over time, where tests are conducted with complex workloads, and where various multi-tenancy scenarios can be investigated.

We also discuss four classes of challenges in developing this approach: methodological, system property-related, workload-related, and metric-related. Last, we summarize our experience towards benchmarking IaaS clouds.

Acknowledgment

Supported by the STW/NWO Veni grant @larGe (11881). Austrian Science Fund project TRP 72-N23 and the Standortagentur Tirol project RainCloud funded this research.

References

[1] P. Mell and T. Grance, "The NIST definition of cloud computing," National Institute of Standards and Technology (NIST) Special Publication 800-145, Sep 2011, [Online] Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf. Last accessed Oct 2012.

[2] Amazon Web Services, "Case studies," Amazon web site, Oct 2012, [Online] Available: http://aws.amazon.com/solutions/case-studies/. Last accessed Oct 2012.

[3] E. Deelman, G. Singh, M. Livny, J. B. Berriman, and J. Good, "The cost of doing science on the cloud: the Montage example," in *SC*. IEEE/ACM, 2008, p. 50.

[4] K. R. Jackson, K. Muriki, L. Ramakrishnan, K. J. Runge, and R. C. Thomas, "Performance and cost analysis of the supernova factory on the amazon aws cloud," *Scientific Programming*, vol. 19, no. 2-3, pp. 107–119, 2011.

[5] I. Raicu, Z. Zhang, M. Wilde, I. T. Foster, P. H. Beckman, K. Iskra, and B. Clifford, "Toward loosely coupled programming on petascale systems," in *SC*. ACM, 2008, p. 22.

[6] A. Buble, L. Bulej, and P. Tuma, "Corba benchmarking: A course with hidden obstacles," in *IPDPS*, 2003, p. 279.

[7] P. Brebner, E. Cecchet, J. Marguerite, P. Tuma, O. Ciuhandu, B. Dufour, L. Eeckhout, S. Frénot, A. S. Krishna, J. Murphy, and C. Verbrugge, "Middleware benchmarking: approaches, results, experiences," *Concurrency and Computation: Practice and Experience*, vol. 17, no. 15, pp. 1799–1805, 2005.

[8] J. Gray, Ed., *The Benchmark Handbook for Database and Transasction Systems*, 2nd ed. Mergan Kaufmann, 1993.

[9] A. B. Downey and D. G. Feitelson, "The elusive goal of workload characterization," *SIGMETRICS Performance Evaluation Review*, vol. 26, no. 4, pp. 14–29, 1999.

[10] S. J. Chapin, W. Cirne, D. G. Feitelson, J. P. Jones, S. T. Leutenegger, U. Schwiegelshohn, W. Smith, and D. Talby, "Benchmarks and standards for the evaluation of parallel job schedulers," in *JSSPP*, 1999, pp. 67–90.

[11] E. Frachtenberg and D. G. Feitelson, "Pitfalls in parallel job scheduling evaluation," in *JSSPP*, 2005, pp. 257–282.

[12] A. Iosup, D. H. J. Epema, C. Franke, A. Papaspyrou, L. Schley, B. Song, and R. Yahyapour, "On grid performance evaluation using synthetic workloads," in *JSSPP*, 2006, pp. 232–255.

[13] R. Jain, Ed., *The Art of Computer Systems Performance Analysis*. John Wiley and Sons Inc., 1991.

[14] J. N. Amaral, "How did this get published? Pitfalls in experimental evaluation of computing systems," LTES talk, 2012, [Online] Available: http://webdocs.cs.ualberta.ca/~amaral/Amaral-LCTES2012.pptx. Last accessed Oct 2012.

[15] A. Iosup and D. H. J. Epema, "GrenchMark: A framework for analyzing, testing, and comparing grids," in *CCGrid*, 2006, pp. 313–320.

[16] M. Oskin, F. T. Chong, and M. K. Farrens, "Hls: combining statistical and symbolic simulation to guide microprocessor designs," in *ISCA*, 2000, pp. 71–82.

[17] L. Eeckhout, S. Nussbaum, J. E. Smith, and K. D. Bosschere, "Statistical simulation: Adding efficiency to the computer designer's toolbox," *IEEE Micro*, vol. 23, no. 5, pp. 26–38, 2003.

[18] D. Genbrugge and L. Eeckhout, "Chip multiprocessor design space exploration through statistical simulation," *IEEE Trans. Computers*, vol. 58, no. 12, pp. 1668–1681, 2009.

[19] R. H. Saavedra and A. J. Smith, "Analysis of benchmark characteristics and benchmark performance prediction," *ACM Trans. Comput. Syst.*, vol. 14, no. 4, pp. 344–384, 1996.

[20] S. Sharkawi, D. DeSota, R. Panda, R. Indukuru, S. Stevens, V. E. Taylor, and X. Wu, "Performance projection of hpc applications using spec cfp2006 benchmarks," in *IPDPS*, 2009, pp. 1–12.

[21] D. Villegas, A. Antoniou, S. M. Sadjadi, and A. Iosup, "An analysis of provisioning and allocation policies for infrastructure-as-a-service clouds," in *CCGrid*, 2012, pp. 612–619.

[22] B. Schroeder, A. Wierman, and M. Harchol-Balter, "Open versus closed: A cautionary tale," in *NSDI*, 2006.

[23] A. Georges, D. Buytaert, and L. Eeckhout, "Statistically rigorous java performance evaluation," in *OOPSLA*, 2007, pp. 57–76.

[24] A. Vahdat, K. Yocum, K. Walsh, P. Mahadevan, D. Kostic, J. S. Chase, and D. Becker, "Scalability and accuracy in a large-scale network emulator," in *OSDI*, 2002.

[25] K. V. Vishwanath, A. Vahdat, K. Yocum, and D. Gupta, "Modelnet: Towards a datacenter emulation environment," in *Peer-to-Peer Computing*, 2009, pp. 81–82.

[26] A. Iosup, N. Yigitbasi, and D. H. J. Epema, "On the performance variability of production cloud services," in *CCGRID*, 2011, pp. 104–113.

[27] A. Iosup, S. Ostermann, N. Yigitbasi, R. Prodan, T. Fahringer, and D. H. J. Epema, "Performance analysis of cloud computing services for many-tasks scientific computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 6, pp. 931–945, 2011.

[28] A. Iosup, M. Jan, O. O. Sonmez, and D. H. J. Epema, "On the dynamic resource availability in grids," in *GRID*. IEEE, 2007, pp. 26–33.

[29] D. Kondo, B. Javadi, A. Iosup, and D. H. J. Epema, "The failure trace archive: Enabling comparative analysis of failures in diverse distributed systems," in *CCGrid*, 2010, pp. 398–407.

[30] R. H. Arpaci-Dusseau, A. C. Arpaci-Dusseau, A. Vahdat, L. T. Liu, T. E. Anderson, and D. A. Patterson, "The interaction of parallel and sequential workloads on a network of workstations," in *SIGMETRICS*, 1995, pp. 267–278.

[31] R. Krebs, C. Momm, and S. Kounev, "Metrics and techniques for quantifying performance isolation in cloud environments," in *Int'l. ACM SIGSOFT conference Quality of Software Architectures (QoSA)*, 2012, pp. 91–100.

[32] N. Huber, M. von Quast, M. Hauck, and S. Kounev, "Evaluating and modeling virtualization performance overhead for cloud environments," in *CLOSER*, 2011, pp. 563–573.

[33] A. Iosup and D. H. J. Epema, "Grid computing workloads," *IEEE Internet Computing*, vol. 15, no. 2, pp. 19–26, 2011.

[34] A. Iosup, O. O. Sonmez, S. Anoep, and D. H. J. Epema, "The performance of bags-of-tasks in large-scale distributed systems," in *HPDC*. ACM, 2008, pp. 97–108.

[35] K. Kim, K. Jeon, H. Han, S. G. Kim, H. Jung, and H. Y. Yeom, "Mrbench: A benchmark for mapreduce framework," in *ICPADS*, 2008, pp. 11–18.

[36] G. Wang, A. R. Butt, P. Pandey, and K. Gupta, "Using realistic simulation for performance analysis of MapReduce setups," in *HPDC Workshops*, 2009, pp. 19–26.

[37] M. Zaharia, D. Borthakur, J. S. Sarma, K. Elmeleegy, S. Shenker, and I. Stoica, "Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling," in *EuroSys*, 2010, pp. 265–278.

[38] A. Ganapathi, Y. Chen, A. Fox, R. H. Katz, and D. A. Patterson, "Statistics-driven workload modeling for the cloud," in *ICDE Workshops*, 2010, pp. 87–92.

[39] Y. Chen, A. Ganapathi, R. Griffith, and R. H. Katz, "The case for evaluating mapreduce performance using workload suites," in *MASCOTS*, 2011, pp. 390–399.

[40] T. A. De Ruiter and A. Iosup, "A workload model for MapReduce," TU Delft MSc thesis, Jun 2012, [Online] Available: http://repository.tudelft.nl/view/ir/uuid:1647e1cb-84fd-46ca-b1e1-21aaf38ef30b/. Last accessed Oct 2012.

[41] A. J. Oliner, A. Ganapathi, and W. Xu, "Advances and challenges in log analysis," *Commun. ACM*, vol. 55, no. 2, pp. 55–61, 2012.

[42] C. Momm and R. Krebs, "A qualitative discussion of different approaches for implementing multi-tenant saas offerings," in *Software Engineering (Workshops)*, 2011, pp. 139–150.

[43] R. Krebs, C. Momm, and S. Kounev, "Architectural concerns in multi-tenant saas applications," in *CLOSER*, 2012, pp. 426–431.

[44] P. Brebner, "Is your cloud elastic enough?: performance modelling the elasticity of infrastructure as a service (iaas) cloud applications," in *ICPE*, 2012, pp. 263–266.

[45] S. Islam, K. Lee, A. Fekete, and A. Liu, "How a consumer can measure elasticity for cloud platforms," in *ICPE*, 2012, pp. 85–96.

[46] E. Walker, "The real cost of a cpu hour," *IEEE Computer*, vol. 42, no. 4, pp. 35–41, 2009.

[47] Y.-K. Kwok and I. Ahmad, "Benchmarking and comparison of the task graph scheduling algorithms," *J. Parallel Distrib. Comput.*, vol. 59, no. 3, pp. 381–422, 1999.

[48] A. Antoniou and A. Iosup, "Performance evaluation of cloud infrastructure using complex workloads," TU Delft MSc thesis, Mar 2012, [Online] Available: http://repository.tudelft.nl/view/ir/uuid:d8eda846-7e93-4340-834a-de3e4aa93f8b/. Last accessed Oct 2012.

[49] S. Ostermann and R. Prodan, "Impact of variable priced Cloud resources on scientific workflow scheduling," in *Euro-Par 2012 – Parallel Processing*, ser. Lecture Notes in Computer Science, C. Kaklamanis, T. Papatheodorou, and P. G. Spirakis, Eds., vol. 7484. Springer, 2012, pp. 350–362. [Online]. Available: http://www.springerlink.com/content/v4q338161171r42v/fulltext.pdf

[50] R. Prodan, M. Sperk, and S. Ostermann, "Evaluating high-performance computing on google app engine," *IEEE Software*, vol. 29, no. 2, pp. 52–58, 2012.

[51] D. Feitelson, "Parallel Workloads Archive," http://www.cs.huji.ac.il/labs/parallel/workload/.

[52] A. Iosup, H. Li, M. Jan, S. Anoep, C. Dumitrescu, L. Wolters, and D. H. J. Epema, "The grid workloads archive," *Future Gener. Comput. Syst.*, vol. 24, no. 7, pp. 672–686, 2008.

[53] Y. Guo and A. Iosup, "The Game Trace Archive," in *NETGAMES*, 2012, pp. 1–6.

[54] B. Zhang, A. Iosup, J. Pouwelse, and D. Epema, "The peer-to-peer trace archive: design and comparative trace analysis," in *ACM CoNEXT Student Workshop*. ACM, 2010, pp. 21:1–2. [Online]. Available: http://doi.acm.org/10.1145/1921206.1921229