

SPEC CloudTM IaaS 2016 Benchmark

Salman Baset
IBM Research
sabaset@us.ibm.com

Marcio Silva
IBM Research
marcios@us.ibm.com

Nicholas Wakou
Dell
nicholas_wakou@dell.com

1. INTRODUCTION

The SPEC CloudTM IaaS 2016 benchmark [1] is the Standard Performance Evaluation Corporation's (SPEC) first benchmark suite to measure cloud performance. The benchmark suite's use is targeted at cloud providers, cloud consumers, hardware vendors, virtualization software vendors, application software vendors, and academic researchers.

The SPEC CloudTM IaaS 2016 Benchmark addresses the performance of infrastructure-as-a-service (IaaS) cloud platforms. IaaS cloud platforms can either be public or private.

The benchmark is designed to stress provisioning as well as runtime aspects of a cloud using I/O and CPU intensive cloud computing workloads. SPEC selected the social media NoSQL database transaction and K-Means clustering using map/reduce as two significant and representative workload types within cloud computing.

Each workload runs in multiple instances, referred to as an application instance. The benchmark instantiates multiple application instances during a run. The application instances and the load they generate stress the provisioning as well as run-time aspects of a cloud. The run-time aspects include CPU, memory, disk I/O, and network I/O of these instances running in a cloud. The benchmark runs the workloads until quality of service (QoS) conditions are reached. The tester can also limit the maximum number of application instances that are instantiated during a run.

The key benchmark metrics are:

- Scalability measures the total amount of work performed by application instances running in a cloud. The aggregate work performed by one or more application instances should linearly scale in an ideal cloud. Scalability is reported for the number of compliant application instances (AIs) completed and is an aggregate of workloads metrics for those AIs normalized against a set of reference metrics.
- Elasticity measures whether the work performed by application instances scales linearly in a cloud when compared to the performance of application instances during baseline phase. Elasticity is expressed as a percentage.
- Mean Instance Provisioning Time measures the time interval between the instance provisioning request and connectivity to port 22 on the instance. This metric is an average across all instances in valid application instances.

2. ACKNOWLEDGEMENTS

The authors were actively involved in the development of SPEC CloudTM IaaS 2016 benchmark and acknowledge the work of all participants.

3. REFERENCES

- [1] SPEC CloudTM IaaS 2016 Benchmark.
https://www.spec.org/cloud_iaas2016/, May 2003.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICPE'17, April 22-26, 2017, L'Aquila, Italy

© 2017 ACM. ISBN 978-1-4503-4404-3/17/04...\$15.00

DOI: <http://dx.doi.org/10.1145/3030207.3053675>