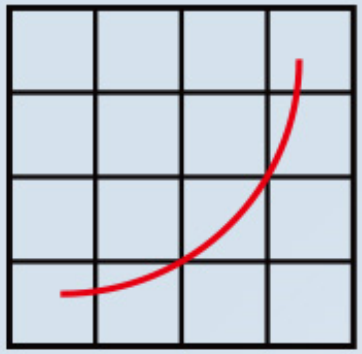


SPEC 2016亚洲峰会
SPEC 2016 ASIA SUMMIT

SPEC Cloud IaaS 2016

Jeremy Arnold (IBM / SPEC Director)
Steve Realmuto (Oracle / OSG Chair)

Topics

What is a cloud?

Cloud characteristics

What is a cloud benchmark?

SPEC Cloud IaaS 2016 benchmark, the first industry standard cloud benchmark

Future considerations

What is a cloud?

NIST SPECIAL PUBLICATION 800-145

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

What is a cloud?

NIST SPECIAL PUBLICATION 800-145

Private cloud

The cloud infrastructure is provisioned for exclusive use by a single organization comprising multiple consumers (e.g., business units). It may be owned, managed, and operated by the organization, a third party, or some combination of them, and it may exist on or off premises.

Public cloud

The cloud infrastructure is provisioned for open use by the general public. It may be owned, managed, and operated by a business, academic, or government organization, or some combination of them. It exists on the premises of the cloud provider.

Hybrid cloud

The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public) that remain unique entities, but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load balancing between clouds).

OSG Cloud Definitions

Whitebox Cloud
The SUT's exact engineering specifications including all hardware and software are known and under the control of the tester. This will typically be the case for **private clouds**.

Blackbox Cloud
A cloud-provider provides a general specification of the system under test (SUT), usually in terms of how the cloud consumer may be billed. The exact hardware details corresponding to these compute units may not be known. This will typically be the case if the entity benchmarking the cloud is different from a cloud provider, e.g., **public clouds** or hosted **private clouds**.

Instance
An abstract execution environment which presents an operating system. This could be implemented using physical machines, virtual machines, or containers.

Application Instance
A group of instances created to run a single workload together. This includes a workload driver instance and a set of instances stressed by the workload driver.

Cloud Characteristics

Elasticity
The degree of which a system is able to adapt to workload changes by provisioning and de-provisioning resources in an automatic manner, such that at each point in time the available resources match the current demand as closely as possible.

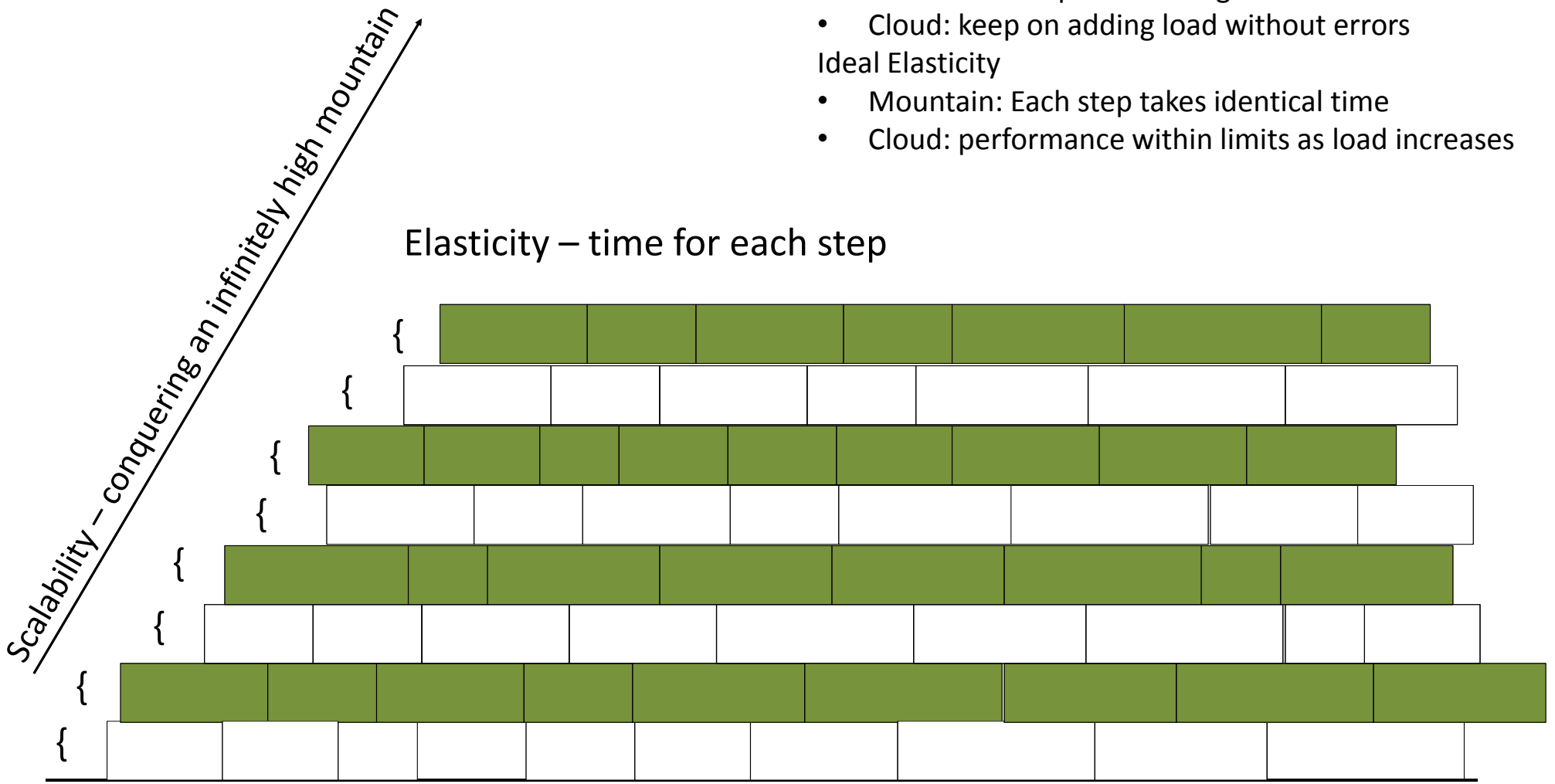
Scalability
The ability of a system to sustain increasing workloads by making use of additional resources, and therefore, in contrast to elasticity, it is not directly related to how well the actual resource demands are matched by the provisioned resources at any point in time.

Source: Ready for Rain? A view from SPEC Research on the future of the Cloud Metrics
https://research.spec.org/fileadmin/user_upload/documents/rg_cloud/endorsed_publications/SPEC-RG-2016-01_CloudMetrics.pdf

Scalability and Elasticity

Ideal Scalability
- Mountain: Keep on climbing
- Cloud: keep on adding load without errors

Ideal Elasticity
- Mountain: Each step takes identical time
- Cloud: performance within limits as load increases



Infinitely Scalable and Elastic?

"Public clouds can have ‘infinite’ scale.”

- In reality
  - Resource limitation: Limited instances can be created within availability zone
  - Budget limitation: for testing
  - Performance variation: due to multiple tenants

“Private clouds can be fully elastic”

- In reality
  - Performance variation: may have better performance under no load, unless configured correctly

A good IaaS cloud benchmark

Measures **meaningful metrics** that are **comparable** are **reproducible**
Measures cloud "elasticity" and "scalability"
Benchmark IaaS clouds, not the workloads!
Measures performance of **public** and **private** infrastructure-as-a-service (IaaS) clouds
Measure **provisioning** and **run-time performance** of a cloud
Uses workloads that resemble "real" workloads

- No micro benchmarks

Places no restriction on how a cloud is configured.
Places no restriction on the type of instances a tester can use (e.g., no restriction on VM/baremetal, CPU, memory, disk, network)

- However, once a tester selects instance types for the test, they cannot be changed
- Minimum machines part of cloud: 4 (with appropriate specs for these machines)

Allows easy addition of new workloads

SPEC Cloud IaaS 2016

The first industry standard benchmark that measures the scalability, elasticity, mean instance provisioning time of clouds, and more

Uses real workloads in multi-instance configuration

SPEC Cloud IaaS 2016 Benchmark

Cloud Vendor: Dell Inc.
Cloud Type / SUT Type: private/whitebox
Hardware Platform: x86_64
Hypervisor: KVM
Cloud Infrastructure: Red Hat Enterprise Linux Openstack Platform 8

Scalability: 29.5 @ 20 Application Instances
Elasticity: 71.9%
Mean Instance Provisioning Time: 135s

Tested by: Dell Revolutionary Cloud and Big Data	SPEC Licence Number : 999	Test Date : Jun-2016

Performance Sections	SUT Configuration Sections		
Performance Summary Performance Details Validation, Errors, and Issues Glossary of Terms	Instance Configuration Cloud Configuration Network Configuration Storage Configuration	**Elasticity + Scalability Phase Date/Time and Test Region** **Elasticity + Scalability Start Time: 2016-06-08_22:08:14_UTC** **Elasticity + Scalability End Time: 2016-06-08_23:33:38_UTC** **Test Region: US Central Time Zone**	**Cloud Informational Metrics** **AI Provisioning Success: 86.96%** **AI Run Success: 100.00** **Total Instances: 131**

Workloads (1/2)

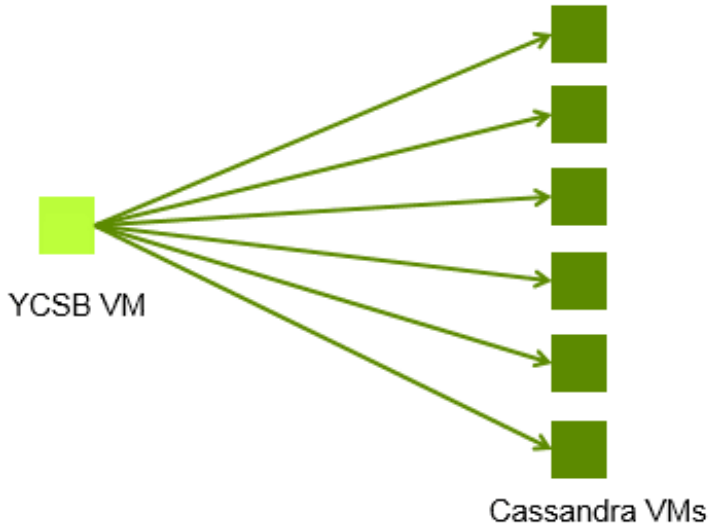
SPEC OSG Cloud subcommittee identified multiple workload classifications already used in current cloud computing services.

SPEC OSG Cloud subcommittee chose I/O and CPU intensive workloads for SPEC Cloud IaaS 2016 benchmark.

- ***NoSQL database transaction*** workload (YCSB / Cassandra)
- ***K-Means clustering using map/reduce*** (Hadoop)

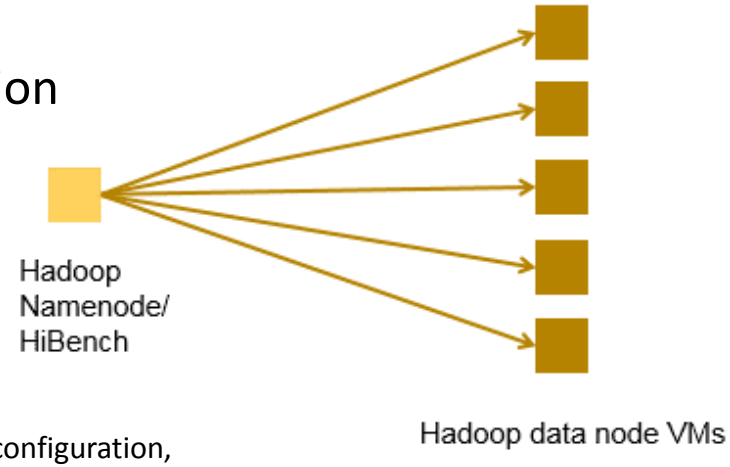
Workloads (2/2)

YCSB Application Instance comprising 7 instances



YCSB / Cassandra
Framework used by a common set of workloads for evaluating performance of different key-value and cloud serving stores.

KMeans Application Instance comprising 6 instances



KMeans
- Hadoop-based CPU intensive workload
- Chose Intel HiBench implementation

No restriction on instance configuration, i.e., cpu, mem, disk, net

Measure Scalability / Elasticity

Run each application YCSB and Kmeans instance once (baseline phase)

Create multiple YCSB and Kmeans application instances over time (elasticity + scalability phase) with statistically similar work

- The more application instances are created and perform work, the higher the scalability
- The low variability in performance during elasticity + scalability phase relative to baseline phase, the higher the elasticity

Benchmark phases

Baseline phase

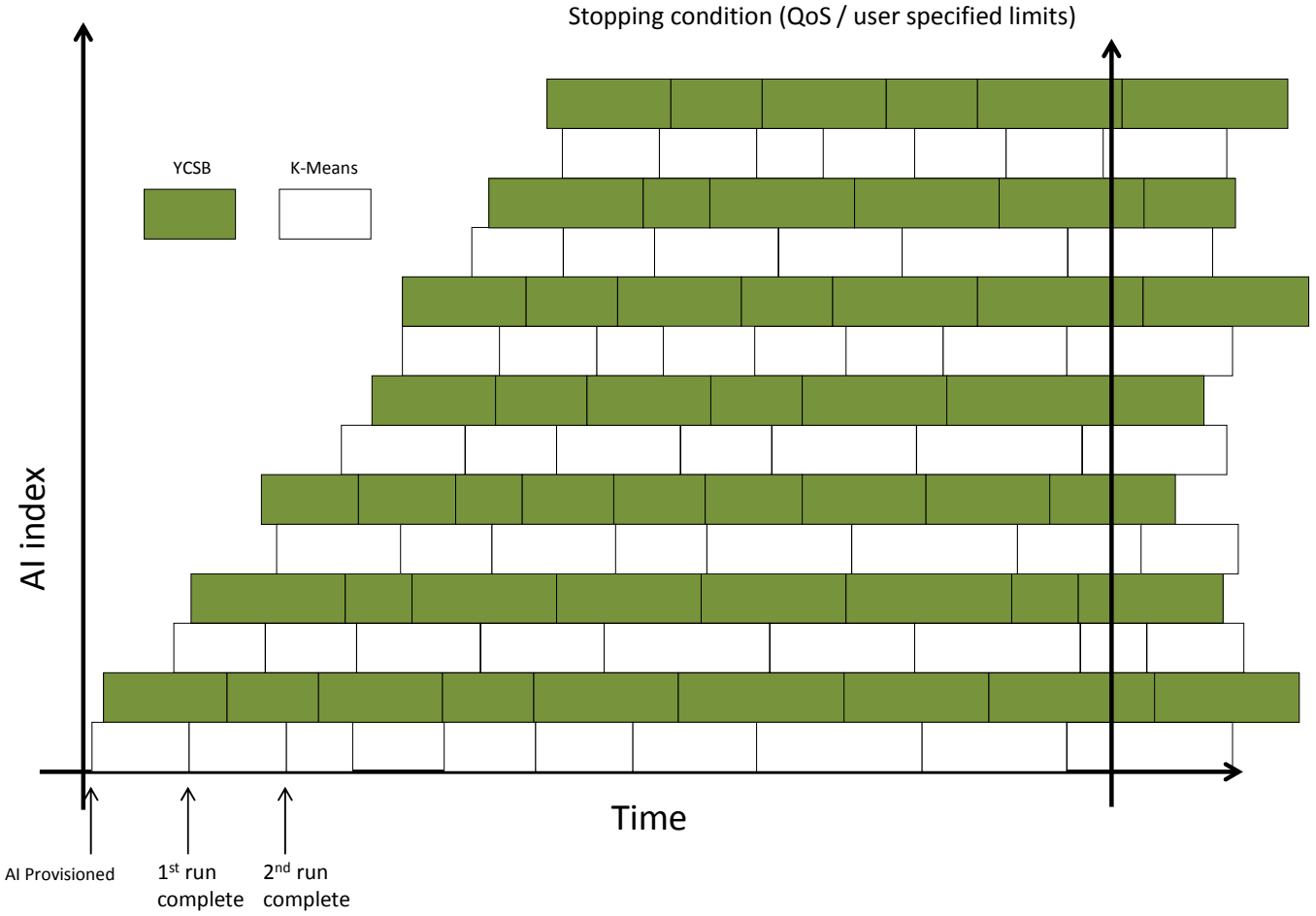
- Determine the performance of a single application instance for determining QoS criteria.
- The assumption is that only a single application instance is run in the cloud (white-box) or in the user account (black-box)

Elasticity + Scalability phase

- Determine cloud elasticity and scalability results when multiple workloads are run
- Stops when QoS criteria as determined from baseline is not met, or maximum AIs as set by a tester are reached
- Scalability results are normalized against a "reference platform"

Elasticity / Scalability phase

Each block indicates the time for a valid run within an application instance
Application instances (comprising multiple instances are provisioning on the left)



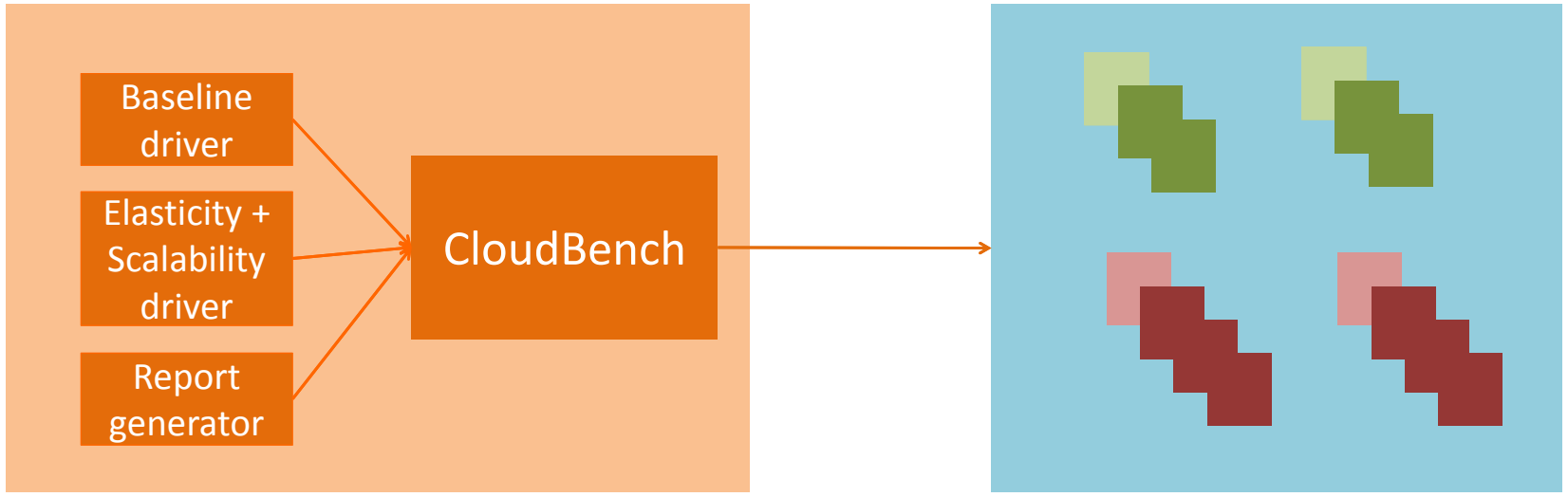
Stopping Conditions

20% of the AIs fail to provision
10% of the AIs have any errors reported in any of the AI runs
50% of the AIs have QoS condition violated across any run

- QoS conditions are defined per workload and are relative to baseline
- YCSB: 50% of AIs have one or more run where
  - Throughput is less than 33% of baseline throughput
  - 99% Insert or Read latency for any run is 3.33 times greater than Insert and Read latency in baseline
- K-Means: 50% of AIs have one or more run where
  - Complete time is 3.33 times greater than completion time in baseline phase

Maximum number of AIs as set by the cloud provider has been provisioned.

Benchmark Setup



Benchmark harness

Cloud SUT

Benchmark harness. It comprises of Cloud Bench (cbtool) and baseline/elasticity drivers, and report generators. cbtool creates application instances in a cloud and collects metrics. cbtool is driven by baseline and scalability driver to execute the baseline and elasticity + scalability phases of the benchmark, respectively.

For white-box clouds the benchmark harness is outside the SUT. For black-box clouds, it can be in the same location or campus.

Group of boxes represents an application instance.
Different color within a group indicates workload generators/ different VM sizes.

Primary Metrics

Scalability

- Measures the total amount of work performed by application instances running in a cloud. Scalability is reported as a unit-less number.
- Example: If throughput of two application instances is 5000 ops/s, then normalized to reference platform, scalability is 2.0

Elasticity

- Measures whether the work performed by application instances scales linearly in a cloud. It is expressed as a percentage (out of 100).

Secondary Metrics

Mean Instance Provisioning Time

- Measures the interval between instance creation request sent by the benchmark harness and the moment when the instance is reachable on port 22 (SSH port).

AI Run Success

- Measures the percentage of runs across all application instances that successfully completed.

AI provisioning success

- Measures the percentage of application instances that provisioning successfully.

Phase start and end date/time and region (not a metric but reported as part of fair usage)

- Date/time when the elasticity phase was started, and the region in which the test was performed.

Benchmark Web Page

Includes documentation, published results, order form, and more
https://www.spec.org/cloud_iaas2016/

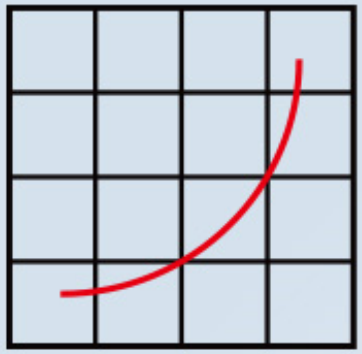
Considerations for the future

Enhancements to SPEC Cloud IaaS

- Benchmark as a service
- Cross data center performance

SPEC Cloud for:

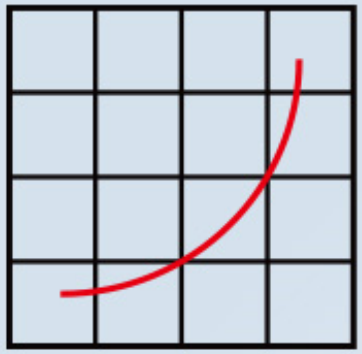
- Object storage
- Serverless systems



spec®

SPEC 2016亚洲峰会
SPEC 2016 ASIA SUMMIT

Q&A



spec®

SPEC 2016亚洲峰会
SPEC 2016 ASIA SUMMIT

Thank you!