Performance & Price/Performance Benchmark of IaaS Providers 2020

This report is based on the benchmark testing and report of independent analyst company 'Cloud Mercato'.

Introduction

The IaaS segment of the cloud market is subject to a permanent race for efficiency where providers redouble their efforts to push the level of performance or the best match with their users' workloads. Virtual machines (VMs) being one of the main components of infrastructures, the constant pursuit of the best products is fast and it can be observed that several offerings are released by providers each quarter.

During 2020 cloud service providers created new lines of VMs based on the latest generation of available CPUs improving performance and as such offering a better price/performance ratio. This means that customers will receive additional compute power for their money.

This benchmark focuses on a comparison of adequate 16 vCPU flavors and block storage of the leading hyperscalers Amazon Web Services (AWS), Microsoft Azure and Google Cloud Platform (GCP). Additionally a European Provider, Deutsche Telekom's Open Telekom Cloud is included. It starts with the performance results and evaluates the price/performance of the evaluated VMs.

Product specifications

For the comparison the following virtual machines of the various providers were used:

Provider	Name	Series	CPU	RAM	Hourly
AWS	m5.4xlarge	General purpose 5	16	65536	$ 0.92
AWS	r5.4xlarge	Memory optimized 5	16	131072	$ 1.22
Google	n1-highmem-16	N1 High Memory	16	106496	$ 1.04
Google	e2-standard-16	E2 Standard	16	65536	$ 0.69
Azure	Standard_D16s_v3	Standard Ds v3	16	65536	$ 0.69
Azure	Standard_E16s_v3	Standard Es v3	16	131072	$ 1.28
OTC	c3.4xlarge.4	Elastic compute 3	16	65536	$ 1.19
OTC	c4.4xlarge.4	Elastic compute 4	16	65536	$ 1.18
OTC	s2.4xlarge.4	General purpose v2	16	65536	$ 0.79
OTC	s3.4xlarge.4	General purpose v3	16	65536	$ 0.94

To run a VM an allocation of block storage volumes to the VM is necessary. To apply a fair benchmark methodology, the necessary block storage was sized with a rule of thumb: 50 GB per CPU. This rule mimics a traditional scaling up with growing computing power. It allows to go through the different performance tiers applied to some storage classes.

The following block storage was used for the volumes attached to VMs, but as well for the specific block storage benchmark.

Provider	Storage	Hourly
Amazon Web Services	General purpose SSD	$ 0.11
Google Cloud	Persistent Disk SSD	$ 0.20
Microsoft Azure	Premium LRS	$ 0.20
Open Telekom Cloud	Ultra-High I/O	$ 0.10

To minimize errors, duplicate VMs were deployed during testing.

Source: IaaS Benchmark study by Cloud Mercato 2020

Performance testing and data

CPU Performance

The VM performance is mainly qualified by computing power delivered by CPU model. This value derives from a lot of other characteristics bound to virtualization and CPU specifications such as frequency, built-in instructions and more. Facing a real-life workload in the cloud, the simple MIPS (million instructions per second) or FLOPS (floating-point operations per second) declared by CPU vendors aren't enough to declare if one machine performs better than another.

To collect a synthetic value benchmark suite Geekbench 5 was used. This software runs workloads with integer, floating point and cryptographic domains. The set of tests includes various kinds such as compression, machine learning or compute vision and each test is performed with single and multi-thread modes.

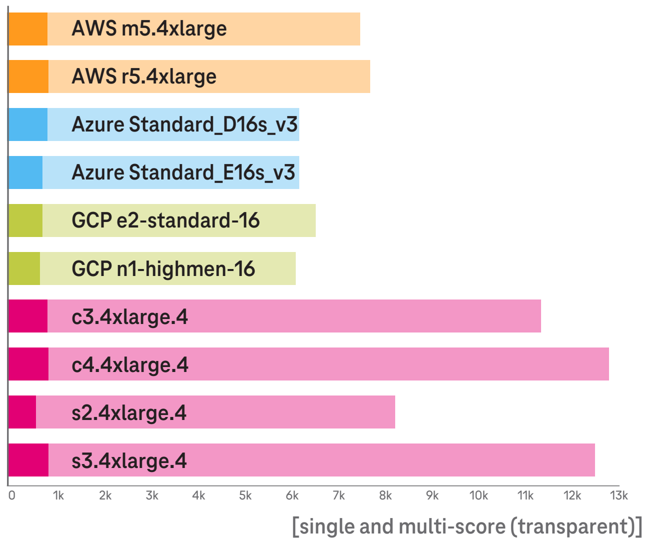
AWS shows the best performance results of the established hyperscalers. Google and Azure are nearly on par. Azure flavors D16 and E16 don't show much of a difference in performance while Google's Standard 16 shows an increased performance of roughly 7 percent compared to the highmem VM. Open Telekom Cloud flavors offer about 60 percent additional performance in the test environment.

CPU Steal

CPU steal is a real phenomenon that occurs from time to time in cloud environments, especially if VMs are not accurately separated. By nature, VMs mandatorily share resources with other tenants or at least with the hypervisor. That could allow CPU steal, but overcommitting could be an additional reason for CPU steal. As a rule of thumb: The older the hypervisor and the more occupied the data center, the higher the risk for CPU steal. CPU sharing can be measured by collecting a Linux kernel counter called "CPU steal".

Expressed here in percentage, this number represents the amount of time that a task was not able to be done by CPU because of someone else's usage. When measuring CPU steal CPUs are loaded with Prime95. This software stresses all CPUs up to 100 percent and allows capturing the amount of tick stolen during the load.

Results show that CPU steal is a rather rate phenomenon when using the watched cloud providers.



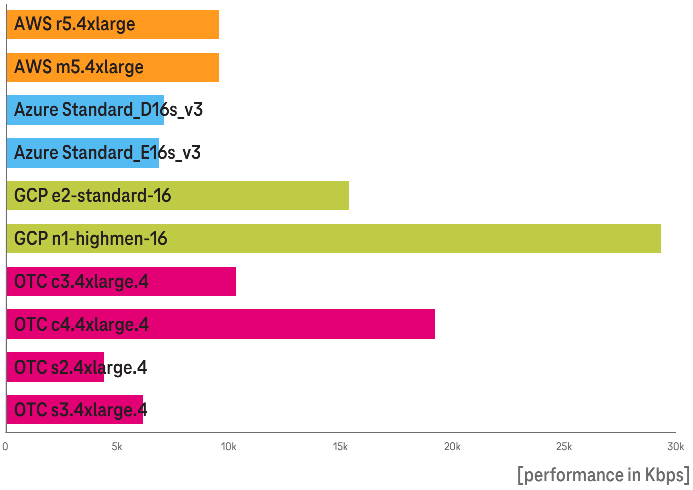
		max	mean	std
Provider	Flavor			
AWS	m5.4xlarge	0.1	0.1	0
AWS	r5.4xlarge	0.1	0	0
Google	e2-standard-16	0.1	0	0
Azure	Standard_D16s_v3	0	0	0
Azure	Standard_E16s_v3	0	0	0
OTC	c3.4xlarge.4	0	0	0
OTC	c4.4xlarge.4	0	0	0
OTC	s2.4xlarge.4	0	0	0
OTC	s3.4xlarge.4	0	0	0

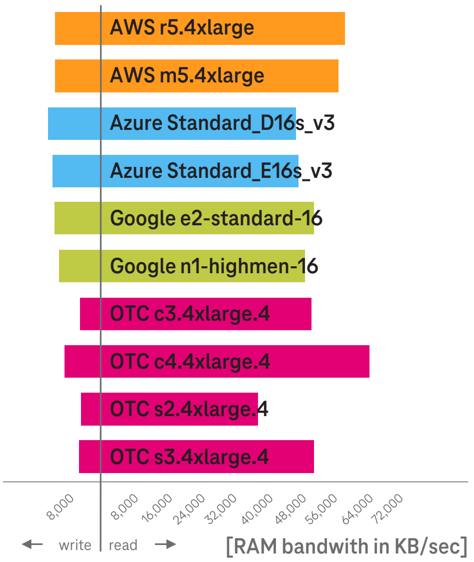
Source: IaaS Benchmark study by Cloud Mercato 2020

RAM Performance

For some use cases the availability of efficient RAM is even more important than CPU power, e.g. for memory-intensive workloads caused by in-memory databases like SAP's HANA®. The volatile memory is the fastest one present on a system. To evaluate RAM performance the bandwidth between CPU and memory is tested using Sysbench. Its test scenarios are based on random access to memory with 1 k blocks.

For this test read and write access have been used. By nature, write mode is slower as it suffers from latency occurred by data storing whereas read only retrieves cached data. RAM performance is time related and may decrease while the hypervisor is filling. AWS is leading the field for read bandwidth while Azure has a slight advantage in write bandwidth with its D16 VM. Open Telekom's c4-VM yields the highest read bandwidth in the test as it is not affected by neighbors (being a dedicated CPU).





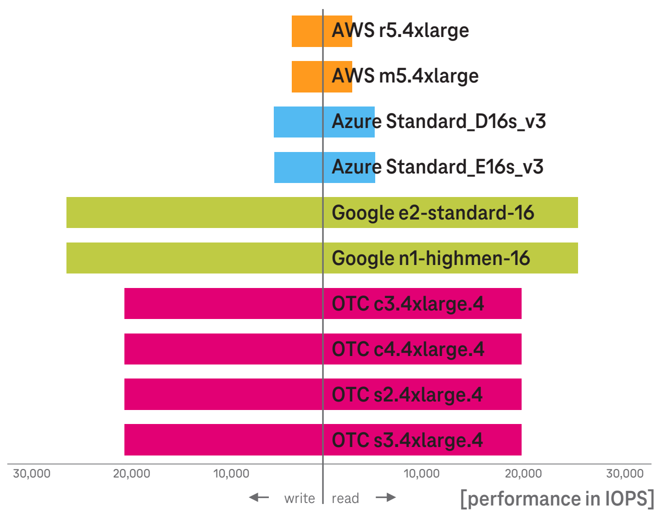
Block storage

The persistent block storage is mainly qualified by its maximum IOPS (input/output per seconds). This value helps to determine the maximum number of transactions that the virtual device is able to provide and at this service, no storage system is equal to another. Like for network bandwidth, storage is a shared resource, that is throttled by vendors. Each provider has defined its own rules about IOPS and the performance factors are: definition of base rules via storage class, maximum performance via volume size and – in the case of Azure – total volume size (performance depends on size tiers calculated on total ownership).
For the test Flexible I/O Tester (FIO) was used. AWS and Azure cap IOPS at about 3,000 while GCP and Open Telekom Cloud throttle bandwidth above 20,000 IOPS.

Network performance

Cloud providers generally have a high-performance internal network throttled in consumer usage to guarantee a certain level of services for all tenants. The maximum performance is completely virtual and is defined by vendor in the VM's network specification. For the evaluation of the maximum bandwidth iPerf3 was used. In the testing setup 2 identical VMs from the same region and availability zone were used. These were loaded with a number of threads equal to CPU to generate the maximum throughput.

Google shows the best results followed by c4 VMs of Open Telekom Cloud.



Source: IaaS Benchmark study by Cloud Mercato 2020

Price/performance value

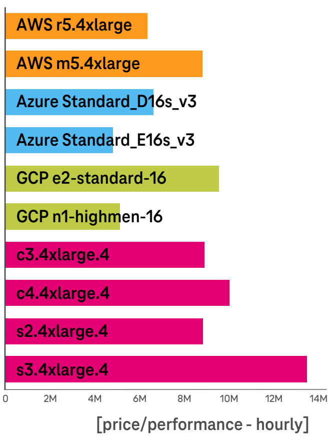
The IaaS segment of the cloud market is subject to a permanent race for efficiency where providers redouble theier efforts to push the level of performance or the best match with their users' workloads. Virtual machines being one of the main components of infrastructures, the constant pursuit of the best products is fast and it can be observed that several offerings are released by providers each quarter.

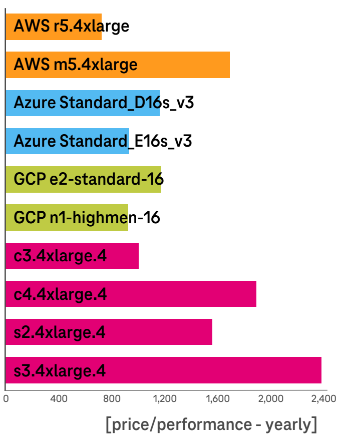
Performance is one side of the coin, the other side is price. To evaluate the real value that a user gains from a cloud provider, a consolidated view containing performance and price is the best approach. Price/performance offers a universal metric for comparing service value. This section shows price/performance calculation for compute and storage based on the prices listed and the results from performance tests. As Open Telekom Cloud is billed in €, the following exchange rate has been used: 1 US$ = 0.858 €.

Compute

For the price/performance value in the CPU area Geekbench multi-score data from the respective section above was used. As it represents a VM performance across different topics, it is well suited to evaluate the whole product capabilities. The multi-score was divided by the price. As different billing options exist across providers, the hourly (on-demand usage) and yearly consumption (long-term usage) are compared.

In the pay-as-you-go consumption model with hourly pricing Google's e2 Standard 16 leads the field followed by AWS' m5.4xlarge VM. This changes when customers opt for yearly pricing: AWS m5 is then by far the best option of the hyperscalers. Independent of the pricing/subscirption model the latest flavors of the Open Telekom Cloud give the best price/performance.

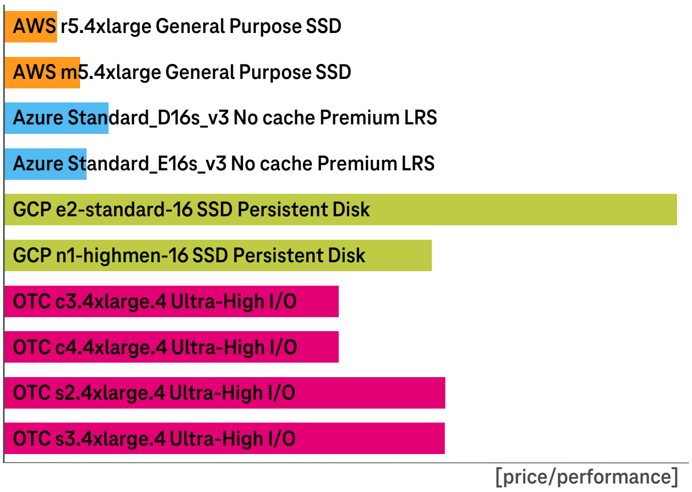




Storage

From the IOPS testing, an apple-to-apple-comparison was created gathering the VM and volume hourly prices. Read and write access performance were also taken into account to calculate the price/performance value using the following formula:

$$\frac{(\text{rIOPS} * 2) + \text{wIOPS}}{\text{Price}}$$



This results reflect the huge gap in performance resp. throttling the bandwidth. Google's persistent disk in connection with the e2 Standard 16 flavor offers by far the best result for block storage in terms of price/performance. Open Telekom Cloud is second.

Source: IaaS Benchmark study by Cloud Mercato 2020

Study notes

Testing was conducted on specific VM types for each provider. Provider VM configurations may yield different results based on underlying infrastructure, virtualization technology, setting (e.g. shared resources), and other technology factors. Furthermore, issues such as user contention or physical hardware malfunctions can also cause suboptimal performance. Cloud Mercato therefore provisioned multiple VMs with the same configuration to better sample the underlying hardware and enabling technology, as well as to improve testing accuracy and limit the effects of underlying environmental variables.

The VMs selected for this engagement were generally-available specific offerings from the various providers. While better performance can often be attained from providers when additional features or support services are purchased, the seleceted VMs used in Cloud Mercato's testing do not leverage such value-added services. This helps provide data and test results that are indicative of real-world customer choices and ensures the most direct comparison possible.

Note: This report is based on the benchmark testing and report of the independent analyst company Cloud Mercato. T-Systems commissioned Cloud Mercato to do this report.