CANONICAL

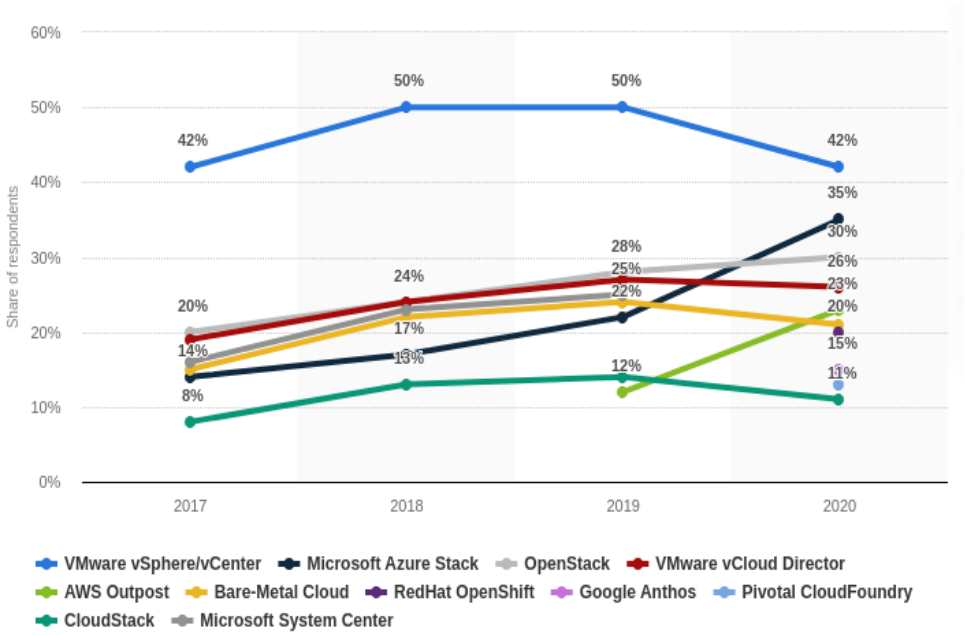
A business guide to hybrid/multi-cloud

April 2021

Introduction

The cloud computing market is growing fast. According to Forrester's "Predictions 2020: Cloud Computing" report from November 2019, the combined global market size of public cloud services was expected to reach $299.4 billion in 2020 and continue to grow at a compound annual growth rate (CAGR) of 29.2% in 2019-2025 [1]. Cloud transformation is driven by the numerous advantages of cloud infrastructure compared to the traditional data centres, such as better economics, scalability and improved DevOps agility.

While public cloud services still account for a significant portion of the overall cloud market, the share of private cloud infrastructure is growing at a similar pace. According to the independent report by Statista, enterprises spent $72.9 billion on private cloud solutions in 2020 and those spendings are going to grow at a CAGR of ~28% in 2021-2027 [2]. The most popular private cloud platforms used by enterprises in 2020 were VMware vSphere, Microsoft Azure Stack, OpenStack, VMware vCloud Director and AWS Outposts [3].



Among various cloud trends, hybrid/multi-cloud architectures are gaining momentum these days. In their "Cloud Trends in 2020: The Year of Complexity, and its Management" report, 451 Research said that hybrid/multi-cloud is emerging as the predominant strategic posture to manage digital-era information technology (IT) and business transformation, with 62% of enterprises pursuing a hybrid IT strategy [4].

This approach is familiar to Canonical customers, such as Cisco, Daimer and SBI BITs, who have implemented their private cloud infrastructure with Canonical's Charmed OpenStack. They now effectively operate their workloads in multi-cloud environments benefiting from the frictionless experience provided by Ubuntu across all infrastructure areas.

This increasing demand for multi-cloud architecture is mostly driven by economics. Organisations are constantly trying to optimise their infrastructure costs to make sure that they only use as much resources as needed and pay less for them. But what is the right approach and solution to truly ensure the total cost of ownership (TCO) reduction in the long-term? In the following whitepaper, we thoroughly analyse common challenges and propose a methodology and solutions to ensure maximum CapEx and OpEx efficiency. We also demonstrate that using a multi-cloud architecture, consisting of common public cloud infrastructure and cost-effective private cloud, is essential to achieve infrastructure cost optimisation.

Challenges with infrastructure cost optimisation

The main promise of cloud transformation was to provide lower TCO compared to the legacy IT infrastructure. In the past, organisations used to maintain dedicated servers for individual services and operate the infrastructure in conjunction with the applications. This was leading to inefficient resources utilisation and additional time spent on maintaining dependencies between the underlying operating system (OS) and running applications. By leveraging virtualisation and containerisation technologies, placing applications inside of isolated silos, distributing them equally across the underlying servers for more efficient resource consumption, and finally, provisioning both the infrastructure and applications in a fully automated way so that they would be used only when needed, they were expecting to achieve TCO reduction.

However, the reality is that not many organisations were able to achieve this goal. There are multiple reasons for that, but the most evident one is that the number of cloud workloads is constantly growing. The same is true for the amount of data organisations are collecting, storing and analysing. This leads to increasing infrastructure costs as more and more resources are needed to host applications. Moreover, the latest trends in the industry, such as cloud native computing, put an additional pressure on infrastructure teams. This is because cloud native computing leverages the microservices architecture which requires application decomposition into atomic units. As a result, each of them is running in a separate virtual machine (VM) or a container, leading to thousands of cloud workloads and an increased resource consumption if not designed properly.

This challenge is well recognised by leading public cloud providers who are constantly lowering their service fees. According to their official blog, Amazon Web Services (AWS) has reduced prices 67 times since AWS launched in 2006 [5]. The presence of multiple public cloud providers on the market helps maintain healthy competition, drive innovation, improve the quality of services and lower prices. However, it does not help to avoid the growth of demand for resources.

As a result, many organisations have recently started exploring private cloud solutions. Since owning is usually more profitable than renting, especially in the long-term and at scale, they were expecting to achieve TCO reduction by running the cloud infrastructure themselves. Private cloud is not an option for everyone, however, due to relatively high initial CapEx costs. Moreover, leading private cloud providers, such as VMware, require expensive licenses to be purchased upfront, before the cloud can even be deployed. Finally, the private cloud always comes with a limited capacity and opinionated technology choices which limit an organisations' ability to run all of their workloads in the private cloud.

For all of these reasons, infrastructure cost optimisation in the cloud environment is not as easy as it seems. Canonical's mission is to guide organisations through this process to make sure that they are maximising their return on investment (ROI) from the cloud transformation process. This usually translates into running workloads where it makes the most sense from an economical point of view. In order to achieve this goal, Canonical has pioneered a number of solutions over the last few years, focused on ensuring cost-effectiveness of cloud infrastructure. These include model-driven operations and price-performance optimisation techniques for cloud infrastructure implementation.

Cloud pricing comparison: a methodology

Although the vast majority of organisations are usually able to easily determine the overall TCO associated with their cloud infrastructure maintenance, some of them struggle to estimate the TCO per application or per service. This is because they do not really know which portion of the infrastructure is utilised by a particular service. As a result, they are not able to compare the prices between various cloud providers to make sure that their workloads always run on the cloud infrastructure that provides the best value for money. This leads to cost-ineffective workload placement and increased infrastructure costs.

In order to make sure that workloads always run where it makes the most sense from an economical standpoint, organisations should adapt certain cloud pricing comparison methodology and be able to move workloads between various clouds once the cost conditions change. In the cloud environment, a proven methodology for pricing comparison are cost-per-resource metrics. Those include:

- cost per VM,
- cost per TB of persistent storage,
- cost per input/output operations per second (IOPS),
- cost per 1Gbps of network bandwidth,
- … and more.

As leading public cloud providers implement so-called pay-as-you-go (PAYG) billing, those metrics are usually counted per hour or per month. Therefore, as the first step, organisations should estimate how much resources (instances, instance types, storage, network, etc.) are needed to run their applications and for how long.

Since cost-per-resource metrics are publicly available for leading public cloud providers, organisations can later use those metrics to estimate the TCO per application. Moreover, a number of cloud pricing comparison tools, such as TCO calculators, are available on the Internet to help obtain even more detailed estimates [6]. Calculating those estimates based on cost-per-resource metrics from various cloud providers allows organisations to choose one that provides the best value for money. Such a process should be executed on a regular basis to ensure maximum TCO reduction.

Things get complicated when it comes to public cloud vs private cloud pricing comparison. This is because private clouds do not have cost-per-resource metrics associated with them by default. This comes from the fact that almost every private cloud is different. There are various private cloud platforms available, each of them having different requirements in terms of the minimum cloud size and licensing. They run on top of different hardware, use different software-defined networking (SDN) platforms and different network topologies. Finally, different vendors provide commercial services for those clouds, each of them charging their customers different service fees. Therefore, cost-per-resource metrics have to be calculated individually for each private cloud being built.

Canonical's Charmed OpenStack is an enterprise private cloud, engineered for the best price-performance that uses the concept of model-driven operations to significantly reduce the private cloud maintenance and operations cost. Charmed OpenStack provides a cost-efficient extension to the public cloud infrastructure, empowering businesses to optimise their CapEx and OpEx costs in multi-cloud environments and to lower the TCO of maintaining their cloud infrastructure.

This is challenging, but not impossible. Leading private cloud providers, including Canonical, provide cost-per-resource estimates for their products. This enables organisations to use the same methodology across public and private clouds in their efforts towards infrastructure cost optimisation. In the following section we present cost-per-resource estimates from a baseline private cloud platform running Canonical's Charmed OpenStack.

Charmed OpenStack: cost-per-resource metrics

Calculating cost-per-resource metrics for a private cloud starts by estimating its TCO and overall capacity. The TCO is then divided by the individual capacity ingredients to extract particular cost-per-resource metrics.

In order to estimate the TCO of a private cloud, a lot of factors have to be taken into account. The cloud is a complex environment, consisting of hardware and software components running in a data centre that has to be managed on a daily basis. Private cloud TCO ingredients are shown in Fig. 1.

Since private cloud hardware has to be renewed on a regular basis, the TCO calculations should include the initial CapEx costs and the recurring OpEx costs for the duration of the renewal period.

Initial CapEx costs	Recurring OpEx costs
• Hardware costs (racks, servers, switches, cabling, etc.) • Licenses and delivery costs	• Hosting services (rent, electricity, network and hardware maintenance) • Operations and maintenance costs • Support subscriptions

Tab. 1. Private cloud TCO ingredients

The following estimates assume a baseline 12-node general-purpose Charmed OpenStack cloud, deployed on a reference hardware as a part of Canonical's Private Cloud Build (PCB) service. Since Charmed OpenStack is fully open source, no license costs apply Typical hosting charges have been included based on Canonical's hosting partners' official pricing. Moreover, to provide the most accurate estimates for the operations, maintenance and subscription costs, it has been assumed that the cloud is managed by Canonical under the Managed OpenStack service which includes the Ubuntu Advantage for Infrastructure (UA-I) support subscription. Finally, the TCO has been estimated for a 3-year hardware renewal period.

Canonical provides commercial services for Charmed OpenStack in the form of the following packages:

- **Private Cloud Build (PCB) and PCB Plus** - fixed-price consultancy packages for Charmed OpenStack implementation on reference architecture and certified hardware, including cloud design and delivery, on-prem workshops, workload analysis and migration plans.
- **Ubuntu Advantage for Infrastructure (UA-I)** - enterprise subscription for Ubuntu, covering all layers of the infrastructure which includes 10 years of security updates, phone and ticket support, production-grade service level agreements (SLAs) and regulatory compliance programmes.
- **Managed OpenStack** - fully-managed private cloud service, including cloud monitoring, maintenance, daily operations, incident and problem resolution, software updates and OpenStack upgrades that enables organisations to fully outsource the management of their private cloud.

In order to estimate the overall capacity of the cloud, one should start by identifying the capacity of a single node and multiplying it by the number of nodes in the cloud. However, there are other factors that have an impact on the ultimate cloud capacity and need to be taken into account. Those include overcommitment ratios for both central processing unit (CPU) and random access memory (RAM), persistent storage replication factor and the desired percentage of cloud resources that should never be exceeded.

For the purpose of preparing the following estimates the default deployment options have been assumed. Those include:

- 2:1 CPU overcommitment ratio,
- no RAM overcommitment,
- persistent storage replication factor of 3,
- the desired cloud utlisation at 75%.

It has also been assumed that the cloud is implemented based on the Hyper-Converged architecture and so that its node specifications match the official recommendations.

Cost-per-VM metrics from a baseline Charmed OpenStack cloud are shown in Tab. 2. As various instance types consume different amounts of cloud resources, their cost-per-VM metrics differ depending on their size. Moreover, since the capacity of a private cloud is always limited, the number of VMs that can run in the private cloud are limited too. In order to allow for more VMs users can scale the cloud out by adding extra nodes. This affects cost-per-VM metrics, however, so they should always be calculated on the fly based on the individual requirements. To help organisations with this process Canonical maintains a TCO calculator for Charmed OpenStack [7]. It is also important to mention that exact prices may vary depending on the used hardware, cloud configuration, local hosting charges and salary costs.

Number of vCPUs	Amount of RAM [GB]	Amount of ephemeral storage [GB]	Amount of persistent storage [GB]	Number of VMs	Hourly cost per VM [USD]
1	4	8	40	3078	0.0132
2	8	8	80	1539	0.0264
4	16	8	160	769	0.0529
8	32	8	320	384	0.1059

Tab. 2. Cost-per-VM metrics from a baseline Charmed OpenStack cloud.

By comparing cost-per-VM metrics from Tab. 2 with the official pricing of AWS Elastic Compute Cloud (EC2) t3a instances, which use the same CPU family as a baseline Charmed OpenStack cloud, it is evident that Charmed OpenStack allows for ~60% cheaper VMs compared to baseline AWS EC2 pricing [8]. And those calculations do not cover any other cost-per-resource metrics that are added to the monthly invoice by public cloud billing systems. As a result, Charmed OpenStack provides a more cost-effective infrastructure for running cloud workloads. However, this is not just a private cloud, but a hybrid/multi-cloud architecture that allows organisations to fully optimise their infrastructure costs.

Cost optimisation in hybrid/multi-cloud environments

Although private clouds usually provide better cost-per-resource metrics than public clouds, it does not immediately mean that they are more cost-effective in all cases. For example, using a private cloud platform for the purpose of running just a few VMs does not really make sense. Moreover, there are additional challenges associated with private cloud use, such as limited capacity and technology options that typically force organisations to use hybrid/multi-cloud architecture. In order to better understand the overall economics of private and public clouds we have to take a closer look at their TCO ingredients (i.e. CapEx and OpEx costs).

TCO analysis

Rough estimates of the private and public cloud TCO for a cloud with a fixed capacity are shown in Fig. 1. The biggest challenge with private cloud implementation is the relatively high initial CapEx costs. Organisations have to invest in hardware, software licenses and pay for the cloud delivery. This is not something that every organisation can afford. On the other hand, public clouds come with close-to-zero CapEx costs, offering immediate access to the platform. The only thing organisations have to do is to create an account and attach their credit card to the billing system. This is why public clouds are more attractive to the majority of the businesses, at least at the beginning of their cloud transformation. In both cases CapEx costs are flat as they are not dependent on the number of the workloads.

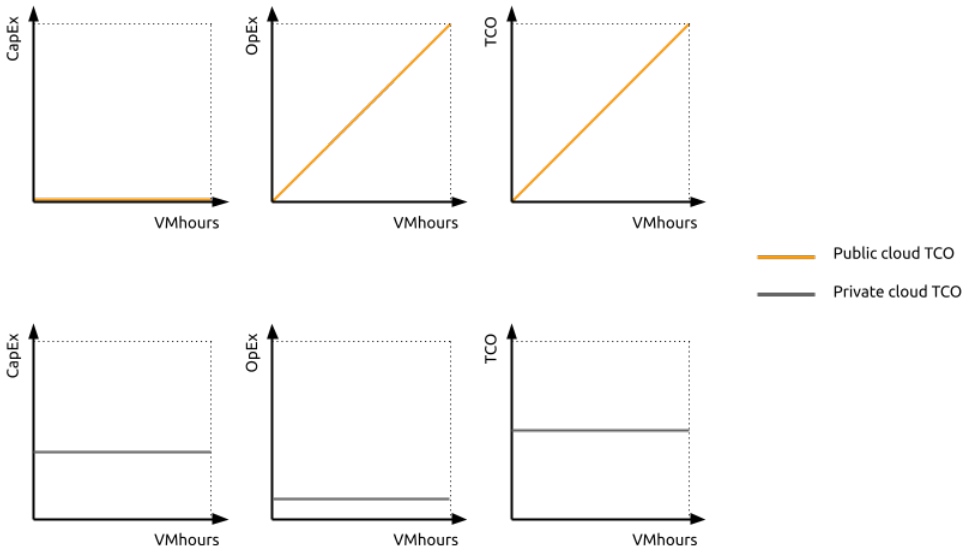


Fig. 1. TCO comparison for private and public clouds.

The situation is totally different when it comes to the OpEx costs analysis. Private cloud OpEx costs remain flat and relatively low compared to the CapEx. This is because organisations always have to pay the same amount of money for hosting services, operations and maintenance, and support subscriptions, regardless of the number of VMs running in the private cloud. In turn, public clouds implement PAYG billing. This means that public cloud OpEx costs are growing as the number of workloads increases. Since cost-per-resource metrics from public clouds are higher than cost-per-resource metrics from private clouds, public cloud OpEx costs are growing very fast.

When we put TCO graphs on the same figure, as shown in Fig. 2, it is evident that the decision on the cloud architecture should always be driven by the overall economics. While public clouds provide the best value for money for small-scale application deployments, at some point it makes more sense to invest in a private cloud. That being said, it is important to remember that private clouds have a limited capacity and come with opinionated technology choices. Those issues can be addressed by adding more nodes and node types. All of that comes with an additional cost, however. Thus, a hybrid/multi-cloud architecture is the most optimal in the majority of the cases.

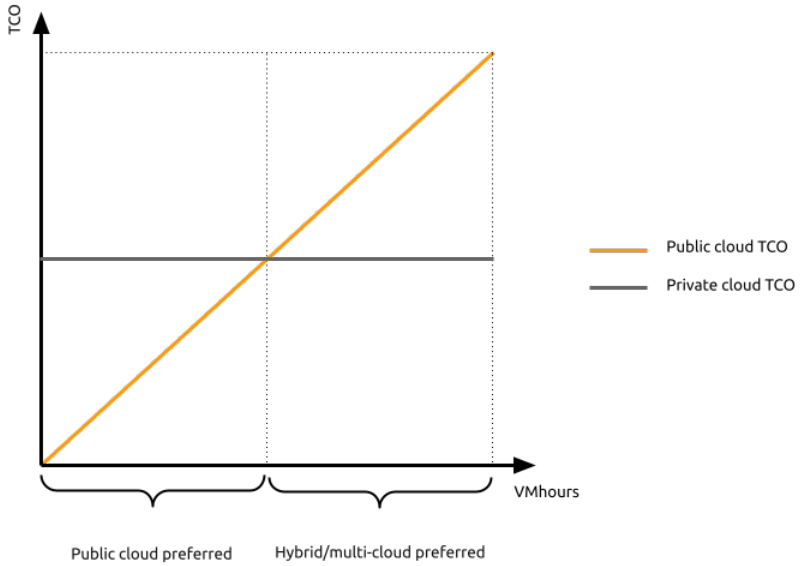


Fig. 2. Cloud architecture driving criteria.

Sample scenarios

In the following section, we demonstrate sample scenarios and provide guidance on the cloud architecture choices. Since such decisions should always be driven by the up-to-date data, we highly recommend checking the actual estimates using the official TCO calculators depending on your actual needs.

Scenario 1: Internal CRM system

In this scenario a software company is hosting a third-party customer relationship management (CRM) system for their internal purposes. The platform consists of a database and web application. The roughly estimated requirements for this system are shown in Tab. 3.

Purpose	Number of vCPUs	Amount of RAM [TB]	Amount of persistent storage [TB]	Network requirements	Additional information
Database	24	0.2	2	No	No
Web application	24	0.1	0	No	No

Tab. 3. Internal CRM platform sample requirements.

In this case, there is no economic justification for private cloud implementation. The organisation should host all workloads in the public cloud, but continue monitoring resource consumption and be able to move to a private cloud once the number of workloads grows.

Scenario 2: Online banking system

In this scenario a financial institution is hosting their own online banking system in the cloud. The system consists of a database, web application and a number of developer VMs. The roughly estimated requirements for this system are shown in Tab. 4.

Purpose	Number of vCPUs	Amount of RAM [TB]	Amount of persistent storage [TB]	Network requirements	Additional information
Database	500	4	40	Moderate inter-AZ traffic	No
Web application	1,000	4	0	Moderate inbound traffic	Daily 16-hour load spike
Developer VM	200	0.8	5	No	Used during business hours only
Developer VM	20	0.08	0.5	No	ARM CPUs required; Used during business hours only

Tab. 4. Online banking system sample requirements.

In this case, it makes more sense to deploy a private cloud infrastructure and host the majority of the workloads there. As the implementation of additional ARM-based hypervisors for the purpose of hosting a small number of developer VMs is not economically feasible, the organisation can use a hybrid/multi-cloud architecture to host those VMs in the public cloud.

Scenario 3: Video streaming system

In this scenario an entertainment company is hosting their own video streaming system. The system consists of a data warehouse, data lake, data analytics, video transcoding engine and web application. The roughly estimated requirements for this system are shown in Tab. 5.

Purpose	Number of vCPUs	Amount of RAM [TB]	Amount of persistent storage [TB]	Network requirements	Additional information
Data warehouse	32,000	512	5,120	Moderate inter-AZ traffic	No
Data lake	4,000	32	256	No	No
Data analytics	12,000	24	0	No	Run daily for 8 hours
Video transcoding engine	24,000	48	0	No	Daily 8-hour load spike
Web application	6,000	24	0	High outbound traffic	Daily 8-hour load spike

Tab. 5. Video streaming system sample requirements.

In this case, a hybrid/multi-cloud architecture is a must as running all of those workloads in the public cloud is extremely expensive. The organisation can leverage the public cloud infrastructure during heavy load periods. At this scale it also makes more sense to hire a dedicated cloud operations team rather than relying on fully-managed services provided by the private cloud vendor.

Multi-cloud best practices

The following section briefly summarizes the best practices for cost optimisation in multi-cloud environments. Following those practices ensures maximum CapEx and OpEx efficiency and helps to augment ROI.

Workloads design

It is not unusual that organisations simply move their legacy monolithic workloads to the cloud without redesigning underlying applications. However, such an approach leads to suboptimal resource consumption and challenges with application maintenance and operations. Cloud workloads should always be designed to consume as many resources as needed. They should be able to scale out once the demand for the resources increases. A proven approach to achieving this is the cloud-native computing. Cloud-native leverages the microservices architecture to decompose applications into atomic units and run them inside of containers. As a result, resources can be carefully optimised according to the actual needs, leading to controlled consumption and lower costs.

Workloads placement

Ignoring the economics aspects of the infrastructure always has a negative impact on the business. Wrong decisions regarding workloads placement can quickly lead to a steady increase in TCO. Workloads should always run on the infrastructure that provides the best value for money. In order to estimate infrastructure costs and make data-driven decisions regarding workloads placement, organisations can use cost-per-resource metrics and TCO calculators. If making an investment in a private cloud does not make sense, they should continue using public cloud infrastructure. As public cloud costs continue to grow, they should leverage multi-cloud architecture and move the vast majority of their workloads to the private cloud.

Capacity monitoring

In practice, the number of workloads running in the cloud is never static. Demand for services changes depending on the day of the week, time of the day, etc. As a result, business applications are usually implemented in a way that they can scale out and in automatically or be fully reprovisioned depending on the current demands. Therefore, it is important to constantly monitor the capacity of the private cloud as the number of workloads changes. In a hybrid/multi-cloud architecture, organisations can burst their workloads to the public cloud during heavy load periods, benefiting from additional resources being available on-demand. It is important to continue monitoring the actual resource consumption though. This is because once the demand for resources continues to grow steadily, at some point it becomes more cost-effective to scale out the private cloud rather than continue using public cloud resources.

Workloads orchestration

In multi-cloud environments business applications are distributed across various cloud providers. While database instances may run in a private cloud, front-end applications may run in a public cloud, assuming it is more cost-effective. Therefore, being able to provision applications across various cloud providers and integrate them regardless of their location becomes crucial for operations sustainability. Among various tools, model-driven operators are gaining momentum in that field. An operator encapsulates a single application as well as all the code and know-how it takes to operate it, such as how to combine and work with other related applications or how to upgrade it. Model-driven operators enable a composition of even very complex application topologies, consisting of containerised and traditional workloads that can span across multiple cloud providers at the same time.

Fully-managed services

Since many organisations do not have enough knowledge and experience to manage a private cloud themselves and they do not have a budget to hire and train new staff, they are sometimes resistant to deploy their own cloud. In such a case, fully-managed services for the private cloud are a valuable option. Leading private cloud providers, including Canonical, offer fully-managed services as an additional layer on the top of their support subscription. Fully-managed services enable organisations to completely offload cloud maintenance and operations tasks to the vendor and benefit from a cost-effective private cloud platform without additional overhead. They also tend to be more economical than hiring a dedicated staff for small-scale deployments. In order to meet the increasing demands for applications operations, Canonical provides fully-managed services for applications too.

Executive summary

As the cloud computing market continues to grow, more and more applications are running in the cloud. This unstoppable trend puts an additional pressure on organisations' IT departments as the demand for resources increases constantly. While public clouds provide an immediate access to theoretically infinite resources, their pricing structure has a direct impact on the budget, resulting in a steadily increasing TCO. As a result, infrastructure costs continue to grow, draining business outcomes and lowering the budget for innovation.

In order to optimise infrastructure costs so that they always pay less for the same amount of resources, organisations usually leverage hybrid/multi-cloud architecture. Using the existing public cloud infrastructure and a cost-effective private cloud platform at the same time enables making careful decisions regarding workloads placement to make sure that they always run where it makes the most sense from an economical standpoint. A proven methodology for cloud pricing comparison are cost-per-resource metrics. Those, in conjunction with advanced tools such as TCO calculators, provide exact TCO estimates. Canonical's Charmed OpenStack minimises those metrics to enable for much cheaper VMs. As a result, Charmed OpenStack serves as a cost-effective extension to the public cloud infrastructure.

References

[1] https://www.forrester.com/report Predictions+2020+Cloud+Computing/-/E-RES157593

[2] https://www.statista.com/statistics/507973/worldwide-true-private-cloud-spending/

[3] https://www.statista.com/statistics/511526/worldwide-survey-private-coud-services-running-application/

[4] https://go.451research.com/2020-mi-cloud-trends-year-of-complexity-and-its-management.html

[5] https://aws.amazon.com/blogs/apn/new-research-from-tso-logic-shows-aws-costs-get-lower-every-year/#:~:text=If%20you've%20been%20paying,since%20AWS%20launched%20in%202006.

[6] https://calculator.aws/#/

[7] https://ubuntu.com/openstack/tco-calculator

[8] https://aws.amazon.com/ec2/pricing/on-demand/

ubuntu
Delivered by Canonical