Canonical

Cloud Optimisation Playbook

March 2023

Table of contents

Executive summary

One of the biggest paradoxes of cloud computing is around cost perception. While leading public cloud providers have constantly been reducing their prices every year, most organisations have seen their cloud bills rise continuously. For example, Amazon Web Services (AWS) has reduced prices a total of 107 times since it was launched in 2006 [1]. Meanwhile, more than 80% of cloud users claim that their organisation has seen an increase in cloud infrastructure spending over the last two years [2].

There are many reasons for this phenomenon. First of all, the spectacular success of cloud computing encouraged organisations to invest more in it. As a result, initial proof of concept (PoC) environments quickly became surrounded by developer virtual machines (VMs) and production workloads. Furthermore, as organisations keep growing, their demand for cloud resources also keeps growing. And finally, over time, many cloud resources usually become underutilised or even completely wasted. Precisely in the same way as any other resources.

This is where cloud optimisation comes into play. It starts with very basic steps, such as understanding current spending and using the right tools to estimate costs. When augmented with recurring cost monitoring practices, cloud optimisation enables organisations to take full control back over their budget. Since average spendings on cloud infrastructure account for 5.7% or an organisations' capital budget and 41% of their IT budget these days, cost savings resulting from a well-oiled cloud optimisation process are meaningful [3-4].

In this whitepaper we demonstrate how to establish this process leaving no single cloud resource underutilised. By following Canonical's best practices we present how to optimise cloud workloads by assigning the right resources, taking the right conditions for performance, compliance and cost into account. This whitepaper provides an overview of native cost optimisation features available across all leading public cloud platforms, cost considerations in hybrid multi-cloud environments and tips for building cost-effective private clouds.

Part I - Cloud optimisation process

Cloud optimisation can bring tangible economic benefits only when established as a recurring process. In the first part of this whitepaper we present the key ingredients of this process and the complete path to its implementation.

Understand your current cloud spending

Understanding your current cloud spending is the first step you should take before you even start optimising. Even though it sounds trivial, according to Canonical's Cloud Pricing Report, more than 20% of cloud users have no idea how much their organisation spends on the cloud every year. What's more, knowing the amount you spend is just one piece of the puzzle. Knowing the structure of the spend is way more important. It is essential to understand which resources (VMs, storage, bandwidth, etc.) your organisation uses, how much and why. If you can't answer those questions, you have homework to do.

Use pricing calculators to estimate costs

Pricing calculators are one of the most effective tools to estimate cloud costs. Since multiple factors have an impact on the final bill, it is always challenging to predict an approximate total cost of ownership (TCO) yourself. Extra charges or discounts might apply depending on your use case. Pricing calculators take all of those factors into account, providing accurate estimates based on your detailed resource requirements. You can then compare estimates from different cloud platforms as well as your current expenses to check potential savings from adopting cloud optimisation techniques. Pricing calculators are available from leading public and private cloud providers [6-9].

Monitor your cloud spending on a regular basis

Once you get an understanding of your current cloud spending and learn how to use pricing calculators to estimate costs, it's essential to repeat those activities on a regular basis. Cloud optimisation is an ongoing process. It's essential to monitor how much you spend regularly, analyse trends, understand them and take reactive actions if needed. It's not uncommon in big enterprises to hire a dedicated person or a team to take care of those responsibilities as their full-time job. This also correlates closely with FinOps practices, making everyone in the organisation accountable for their cloud resource usage.

Take reactive actions if needed

Monitoring and analysing your cloud spend on a regular basis is just the beginning of the cloud optimisation process. Taking reactive actions based on your analysis is what takes you to a fully optimised cloud environment. Resulting actions might be as simple as implementing an automation to shut down your developers' VMs at the end of the day. Others might turn into large, multi-year projects (e.g. re-designing your business applications to be cloud-native). It is essential to estimate potential savings resulting from taking such actions and evaluate whether there is business justification for implementing them. You should keep in mind, however, that there might be other reasons for undertaking such projects, aside from cost reduction that is.

Take proactive actions

Finally, even before things get worse and you lose control over your cloud bills, you should take proactive actions. Organisations should constantly review the various optimisation techniques available out there and plan whether and when to implement them. It is not uncommon to build a well-structured roadmap or set some high-level TCO reduction goals and keep working until they're achieved. All of that takes us to the second part of this whitepaper which provides an overview of the most common cloud optimisation techniques.

Part II - Cloud optimisation techniques

There are many cloud optimisation techniques that businesses can adopt to drive their costs down. Not all of them might be applicable to your organisation. However, all of them should be taken into consideration as a part of the process roll-out.

Generic rules

We are going to start with an overview of some generic rules. These should be followed regardless of the cloud environment being used.

Terminate unused cloud resources

According to Flexera's 2022 "State of the Cloud Report" , one-third of cloud resources are wasted these days [10]. This includes unused VMs, data which is no longer needed, servers with no VMs running on them, yet consuming power, etc. It is important to review all your cloud resources in use and terminate non-essential ones on a regular basis. This can either be done manually or thanks to automation tools; either those available natively in your cloud platform or implemented by your cloud operations team.

Group resources that have the same lifecycle

Following the previous point, always group resources that have the same lifecycle. For example, if your daily data analytics job requires a bunch of services to be provisioned for 1 hour, those should always be grouped together. This reduces the chance that unused resources will be left once the job is finished incurring unnecessary cost. Again, you can either use native tools available in your cloud platform, such as Azure Resource Groups, or implement something yourself.

Consider workload containerisation

Containers offer better utilisation of underlying compute resources thanks to the lack of virtualisation overhead. As a result, re-designing your business applications to be cloud-native not only improves your developers' and operators' experience; it leads to significant cost savings over time. This is especially relevant for public clouds that provide native container-as-a-service (CaaS) solutions, such as Amazon Elastic Container Service (ECS). In this case, you get competitive pricing thanks to various optimisation techniques used underneath.

Public clouds

Leading public cloud platforms provide several native features that can be used in the cloud optimisation process. The following section provides an overview of those features and considerations regarding the types of workloads they are suitable for.

Choose the right region

Cloud optimisation in public clouds starts with choosing the right region for hosting your workloads. Prices vary significantly between regions due to differences in the cost of data centre facilities and maintenance, allowing for up to 30% cost savings. On the other hand, not every region might be appropriate for your applications because of bandwidth limitations, high latency or data sovereignty concerns. As a result, choosing the right region is always a trade-off between cost, performance and compliance.

Consider using ARM instances

Ubuntu is the world's leading Linux distribution. It delivers a consistent, frictionless experience across developers' desktops, public clouds, data centres, and edge infrastructure.

ARM instances tend to be way cheaper than similar Intel or AMD instances. At the same time, they provide similar or even better performance, leading to up to 50% better price-performance and significant cost savings. ARM architecture is the first-class citizen for Ubuntu and the vast majority of open-source applications fully support ARM. You should check, however, whether this will work in your particular case.

Use reserved instances for long-running workloads

Reserved instances (aka saving plans, reservations or committed use discount) are a compelling option when running workloads in the long term. By committing to specific usage, measured in $/hour, for at least one year, you can benefit from up to 70% cost savings compared to on-demand pricing. Thus, reserved instances are best suitable for production environments with pre-defined resource requirements. Leading pubic cloud providers also offer dedicated hosts (aka sole-tenant nodes) with all their resources available exclusively.

Use spot instances for interruptible workloads

Canonical provides all necessary building blocks and commercial services for private cloud implementation, including design and delivery, enterprise support and fully-managed private cloud: from micro clouds to large-scale environments.

Spot instances benefit from unused public cloud resources which would otherwise be wasted. Spot Instances are launched whenever any extra resources are available. They are hibernated, stopped or terminated when the cloud needs those resources back. Spot Instances allow for up to 90% cost savings compared to on-demand pricing but are not suitable for workloads that need to meet service-level agreements (SLAs).

Hybrid multi-cloud

While public clouds are the most economical option to start with, standardising on a hybrid multi-cloud architecture makes much more sense once the number of workloads grows [5]. In other cases it may also be required due to compliance regulations. In the following section we highlight the tips for running applications in such environments, ensuring cloud optimisation.

Design your applications to be cloud-agnostic

The primary idea behind hybrid multi-cloud is to use more than one cloud platform at the same time and be able to move workloads between them on demand. As a result, you should design your applications to be cloud-agnostic to avoid surprises when moving them around. Standardising on common application programming interfaces (APIs) and avoiding platform-specific tools helps to achieve this goal. One example of such a solution is Juju.

Juju is an open source application management software that drives the deployment, integration and management of Kubernetes, container and VM-native applications across a wide range of cloud environments, providing multi-cloud integration capabilities.

Always run workloads where it costs less

In a hybrid multi-cloud environment workloads should always run where it costs less, unless performance and compliance requirements dictate otherwise. While public clouds are more economical on a small scale and in the short term, migrating the majority of your workloads to a cost-effective private cloud is more economical once the number of workloads grows [11]. A thorough cost analysis is key to make an informed decision regarding the placement of your workloads. You can use pricing calculators for this purpose, as mentioned earlier.

Use comprehensive and highly scalable public cloud resources whenever needed

While private clouds provide tangible economic benefits when running workloads in the long term and on a large scale, not all types of workloads might be suitable for private clouds. For example, while 90% of your workloads are x86_64 based, there might not be a business case for building another cloud for the remaining 10% that are ARM based. Ad-hoc resource-intensive jobs, such as data analytics, might also work better from an economic standpoint, if run in the public cloud. It all depends on your specific case and cost conditions you are in.

Metal-As-A-Service (MAAS) automatically discovers all physical machines available on the network, configures them and enables on-demand provisioning with an operating system (OS) of your choice, effectively turning your data centre into a bare metal cloud.

Private clouds

Standardising on hybrid multi-cloud architecture involves implementing a private cloud infrastructure. Therefore, optimising your private cloud is important too. The following section provides an overview of Canonical's best practices for building price-performance-optimised private clouds.

Use optimal architecture for private cloud implementation

Designing your private cloud for price-performance starts with choosing an optimal architecture for its implementation. This involves making decisions about network topology being used, server types and their components, including processors, memory and storage devices. A case study from Canonical's internal cloud demonstrated that well-architected 40-node private cloud provides up to 80% cheaper VMs compared to public cloud on-demand pricing, while ensuring high performance [12].

Use open-source software to avoid high licence costs

Although there are many proprietary virtualisation and private cloud solutions available on the market, their economic conditions don't make them attractive in the cloud optimisation process. This is because proprietary solutions require expensive licences to be purchased upfront, making CAPEX unnecessarily high. Open-source solutions do not require any licences. At the same time they ensure feature parity with the leading proprietary solutions, serving as a reasonable alternative.

Use automation to cut spendings on operations

The biggest portion of private cloud OPEX is spent on internal operations and maintenance. Therefore, reducing those expenses has a direct impact on TCO reduction. Your biggest friend here is automation. By automating typical every-day tasks, such as bare metal provisioning, backups or upgrades, you can effectively offload your cloud operations team, using their expertise somewhere else. Canonical's recommended solutions for data centre automation are MAAS and Juju.

Monitor resource consumption and shutdown unused machines

In the same way as you should monitor virtual resources being used, you should keep track of physical resources in use when operating the private cloud. It will rarely run at full throttle, leaving many servers underutilised. Therefore, it is essential to migrate your workloads to the lowest number of servers possible, maximising workload density. It's also best practice to shutdown unused servers so that they don't consume power. Leading private cloud platforms offer fully automated capabilities like that, but you can also design your own framework and practices.

Consider fully-managed vs self-managed option

Finally, you should consider whether you're going to operate your private cloud yourself or outsource all its operations to a managed service provider (MSP). While a self-managed private cloud is a more economical option when running hundreds of nodes, a fully-managed private cloud works better from the economic standpoint on a smaller scale [13]. This is because operating the private cloud usually requires hiring dedicated personnel and training them. It is important to mention, however, that a fully-managed private cloud might not always be an option because of data sovereignty concerns.

Conclusions

Adopting cloud optimisation techniques enables organisations to significantly lower their cloud spending while ensuring the desired performance and necessary compliance. Since average spending on cloud infrastructure accounts for a significant portion of organisations' capital budget, any savings resulting from cloud optimisation are meaningful from the budget point of view. This is especially important in times of unrest and economic instability.

However, cloud optimisation can only lead to long-term cost savings when established as a recurring process. It is important to understand current spendings, keep monitoring them and try to optimise them on a regular basis. Moreover, adopting cloud optimisation practices may sometimes require making long-term investments. Those include re-designing applications to be cloud-native, replatforming them to be cloud-agnostic or building cost-effective private cloud infrastructure.

As a result, while some of the presented techniques bring immediate economic benefits, others turn out to be profitable only after months or even years. However, doing nothing is still worse than doing small incremental improvements. A fully optimised cloud is not something that can be achieved in a day. It is a moving target that all organisations should start chasing as soon as possible.

Learn more

- Check Canonical's Cloud Pricing Report to get instant access to cloud list prices and service fees from leading public and private cloud providers, sample cost scenarios and exclusive commentary from industry experts.
- Read our OpenStack Deployment Guide to learn how to quickly get started with the world's leading open-source private cloud platform and how to design it for cost efficiency.
- Refer to a whitepaper about private cloud price-performance for exact tips on how to design an optimal architecture for building cost-effective private cloud infrastructure.
- Get in touch with Canonical experts.

References

[1] https://aws.amazon.com/blogs/aws-cost-management/amazon-ec2-15th-years-of-optimizing-and-saving-your-it-costs/

[2] https://ubuntu.com/engage/cloud-pricing-report

[3] https://avasant.com/report/it-budgets-surge-ahead-despite-choppy-economic-waters/

[4] https://www.gartner.com/en/newsroom/press-releases/2022-02-09-gartner-says-more-than-half-of-enterprise-it-spending

[5] https://ubuntu.com/engage/hybrid-cloud-business-guide

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

[7] https://azure.microsoft.com/en-us/pricing/calculator/

[8] https://cloud.google.com/products/calculator

[9] https://ubuntu.com/openstack/pricing-calculator

[10] https://info.flexera.com/CM-REPORT-State-of-the-Cloud

[11] https://ubuntu.com/engage/hybrid-cloud-business-guide

[12] https://ubuntu.com/engage/architecting-price-performance-private-cloud

[13] https://ubuntu.com/engage/managed-services-overcoming-cio-challenges-2021

© Canonical Limited 2023. Ubuntu, Kubuntu, Canonical and their associated logos are the registered trademarks of Canonical Ltd. All other trademarks are the properties of their respective owners. Any information referred to in this document may change without notice and Canonical will not be held responsible for any such changes.

Canonical Limited, Registered in Isle of Man, Company number 110334C, Registered Office: 2nd Floor, Clarendon House, Victoria Street, Douglas IM1 2LN, Isle of Man, VAT Registration: GB 003 2322 47

Canonical Ubuntu