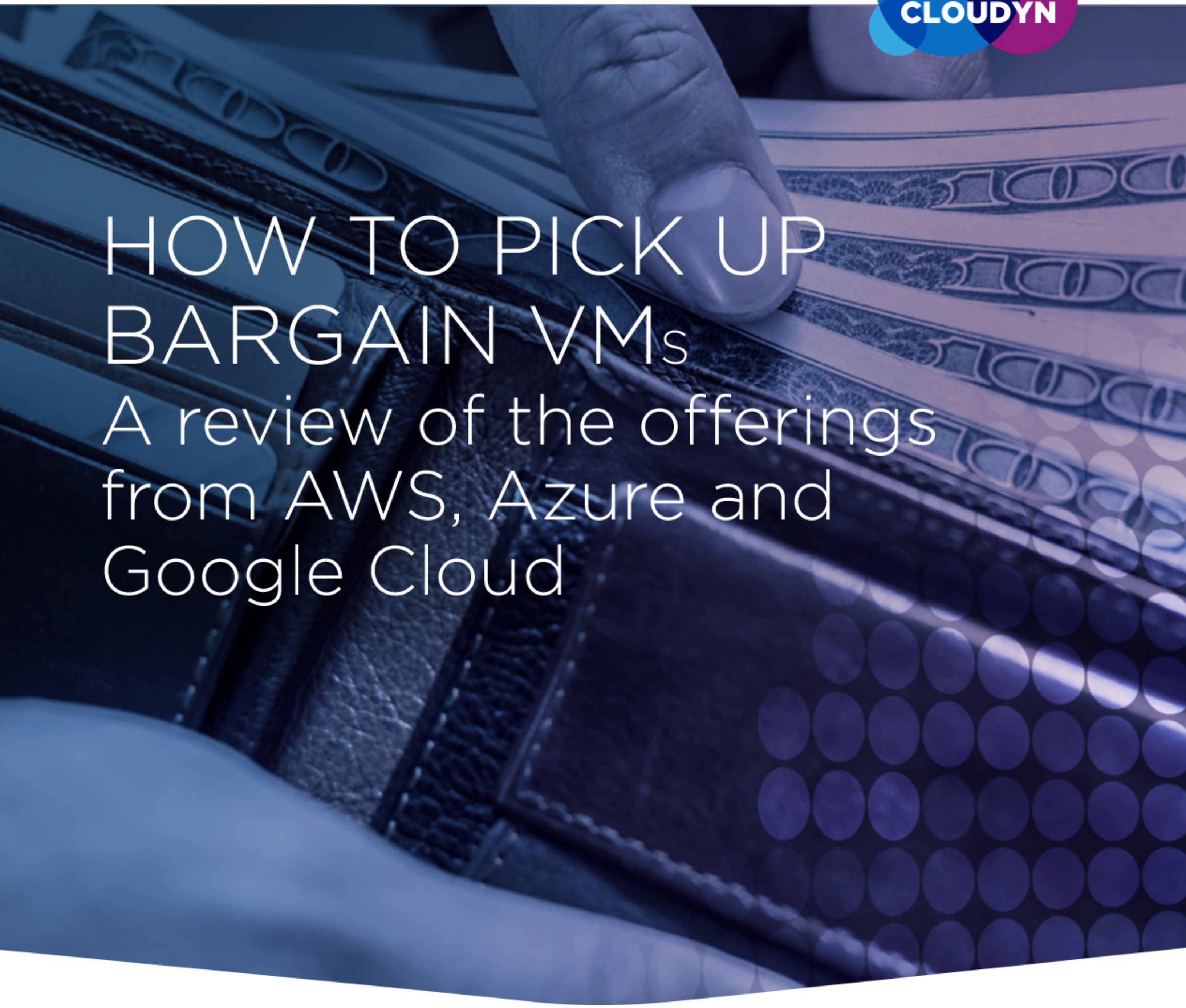
Cloudyn eBook



HOW TO PICK UP BARGAIN VMs A review of the offerings from AWS, Azure and Google Cloud

How to Pick up Bargain VMs

A REVIEW OF THE OFFERINGS FROM AWS, AZURE AND GOOGLE CLOUD

A recent announcement at Microsoft Build 2017 revealed that Microsoft is offering a new way to purchase Azure compute capacity at a much lower price; the option of purchasing **low-priority VMs**. This new offering from Microsoft is reminiscent of Google's Preemptible Instances and the Spot Instances available from AWS.

So, what are the differences between these offerings from the three main cloud providers? Let's take a look and discuss each one's advantages and disadvantages. We'll start with the recently announced low-priority VMs and work our way backwards.

What are Low-Priority VMs?

Microsoft's Low-Priority VMs are essentially their surplus compute capacity, which has now been made available to their customers for a period of up to 24 hours. Their availability varies by the time of day, day of week and general demand from Azure for their different VM sizes. Once purchased, if Azure needs the capacity for a higher priority job, they can take it back when no surplus capacity is available, and are therefore not considered to be dependable. In return for this unpredictability, Microsoft offers their low-priority VMs at a greatly discounted price - **up to 80% less than the cost of on-demand VMs**.

Clearly low-priority VMs are not suitable for all workloads, but they can provide excellent savings for the right type of job; like many batch processing jobs which can withstand interruptions and may also have flexibility in job execution time. In fact, almost all workloads that can use Azure Batch can take advantage of low-priority VMs.

Preemptible VMs

In May 2015, Google Platform announced their Preemptible Virtual Machines, offering a savings of up to 80% on the cost of its regular on-demand instances.

The move was designed to attract more enterprise customers to the Google platform and challenge the dominance of Amazon and Microsoft in the cloud infrastructure marketplace.

Like Azure's low-priority VMs, GCP's Preemptible VMs automatically shut down after 24 hours, and may be shut down earlier with a notice period of just 30 seconds, rendering them also unpredictable and only suited to specific types of workloads.

Google takes a number of factors into consideration in its preemption process. For example, it avoids stopping too many VMs from a single customer and gives preference to instances that have been running longer. In other words, instances are more at risk of preemption when they first start running. However, apart from any separate licensing costs, you're not charged if Google stops your instance within the first 10 minutes.

Charges for Preemptible VMs are typically around 20% of the full on-demand price. Based on average usage, which takes into account partial sustained-use discounts, they generally work out at about 25% of the running cost of the equivalent standard machine. The service carries no guarantee of availability and no SLA.

Preemptible Instances are really easy to purchase. Simply add **--preemptible** when creating compute instances.

AWS Spot Instances

AWS was the first cloud provider to offer a way to quickly and simply purchase extra AWS EC2 computing capacity at a discount, with their AWS Spot Instances. Spot instances do not have a fixed rate, but rather allocation is based on a bidding process where your machine will run as long as your bid price is above the current Spot price on the Spot Market. Your Spot Instances will continue to run until you choose to terminate them, or the Spot price exceeds the your maximum specified amount.

If the Spot price increases, and exceeds your specified price, you will receive a two minute warning before your Spot instance is terminated, and you will not be charged for the partial hour that your instance has run.

Amazon offers Spot Instances at more than 75% lower than their on-demand prices, and they are commonly used for high-performance computing (HPC) whereby firms that perform financial or scientific analysis spin up hundreds or thousands of machines for a short time, and other applications that can handle disruptions.

What Applications Are a Good Fit for Preemption?

Preemptible VMs are a great way to reduce the cost of processing large-scale workloads that are not time sensitive, but do require massive amounts of compute resources. They're also well suited to jobs such as **security testing, short-term scaling, media encoding, financial** or scientific analysis and **insurance risk management**, where you can spin up hundreds or thousands of machines and quickly complete a job in a short space of time. However, they're not suitable for mission-critical services such as operational databases and internet applications.

How Should Your Workloads Handle Preemption?

As already mentioned, preemptible/ low priority and Spot VMs are somewhat unstable and therefore suited to specific types of workloads. There are also a number of measures that can be taken in order to protect and recover your running application.

First of all, if you haven't done so already, you should build fault tolerance into your application. You can also mitigate against the impact of preemption by combining regular instances with Preemptible/Low-Priority/ Spot VMs in your clusters, thereby ensuring you maintain a baseline level of compute availability.

Alternatively, you can create a shutdown script that responds to preemption alerts by automatically launching regular instances to cover your shortfall in compute capacity. But, if costs are paramount and you're prepared to wait, your script should simply clean up and save your job so it can pick up where it left off.

In either case, you should test your application's response to a preemption event. You can do this by stopping the VM and checking that your script correctly completes your shutdown procedure.

However, if you want to avoid preemption in the first place, you should try running your instances at off-peak times, such as nights and weekends, when the risk of disruption is at a minimum.

AWS vs. Azure vs. Google

So, which preemptible instance offers the best value? It's hard to say. Azure's low-priority VMs are hot off the press and therefore precise details about actual cost and behavior are not yet known. Spot instances have several advantages over Low Priority VMs and Preemptible Instances, in that **they don't automatically terminate after 24 hours** and you can bid at a higher price to **reduce the risk of interruption**, thereby providing what could be seen as a more stable option.

Some enterprise customers will prefer the more predictable costs of Preemptible VMs, while some will like the flexibility and greater stability of Spot Instances.

Regarding costs, Preemptible VM costs are rounded up to the nearest minute. However, if you terminate your instance within the first 10 minutes, usage is rounded up to a full 10 minutes. By contrast, the cost for Spot Instances are rounded up to the nearest hour. But, if your Spot Instance is interrupted, you don't pay for your last partial hour of usage.

Finally, Google only gives you 30 seconds' notice of preemption, while Amazon gives you 2 minutes' notice of interruption. In addition, preemptible VMs are available to any Compute Engine instance type, whereas Spot Instances aren't available for the burstable T2 family of instances.

A comparison of the preemptible instances offered by AWS, Azure and GCP

Cloud Provider	amazon web services™	Microsoft Azure	Google Cloud Platform
Discounted Instance Name	Spot Instance	Low-priority VMs	Preemptible Instances
Purchasing Type	Bidding	Fixed price	Fixed price
Potential Savings compared to On-Demand Pricing	More than 75%	Up to 80%	Up to 80%
Cost	Rounded up to the nearest hour	As yet unknown	Rounded up to the nearest minute
Advanced warning of preemption	Two minutes	As yet unknown	30 seconds
Life of Instance	Unlimited but dependent on bidding price	Up to 24 hours or high demand at a higher price	Up to 24 hours or high demand at a higher price
Availability	All server types except for T2	As yet unknown	All

Each discounted instance has its own advantages and disadvantages just as each provider’s offerings have their pluses and minuses. What is important is that you find the best fit for your enterprise and its needs.

Identifying when you can choose preemptible instances

A cloud business management platform will analyze your pattern of cloud behavior, identify your specific use case, and recommend the most efficient ways to optimize consumption.

As an example, such a solution can detect if you are running many applications for short amounts of time while paying for on-demand instances, when you could save money and use preemptible instances instead.

ABOUT CLOUDYN

Cloudyn is an enterprise-grade, SaaS solution that pioneered the single-pane-of-glass approach to managing and optimizing multi-platform, hybrid cloud environments. Supporting Microsoft Azure, Amazon Web Services, Google Cloud, OpenStack and Docker, Cloudyn delivers measurable cloud success by enabling full visibility and accountability packaged with continuous optimization across all clouds. The solution provides insights into usage, performance and cost, coupled with actionable recommendations for smart cloud optimization. Cloudyn enables accountability through comprehensive cost allocation and management helping enterprises get to cloud ROI more rapidly. Thousands of global customers rely on Cloudyn, including Fortune 500 leaders across all major market verticals.

Visit us at **www.cloudyn.com** or contact-us@cloudyn.com and follow us on @Cloudyn_Buzz.