Enterprises are missing out by not optimizing cloud spending, not going multicloud

Analysts - Owen Rogers, Jean Atelsek

Publication date: Tuesday, September 7 2021

Introduction

Across the global IaaS market, businesses are missing out on significant savings by sticking to on-demand pricing. Cloud providers have created an array of cost-saving mechanisms that many, but certainly not all, businesses are taking advantage of. The benefits of optimization often outweigh the benefit of migrating. Although pursuing a multicloud model for cost arbitrage can be challenging, direct cloud costs can plummet as a result.

The 451 Take

For new applications, enterprises should choose the cloud service that best suits their needs in terms of both cost and business requirements. Optimizing – by purchasing capacity in advance or setting up workloads to scale resources dynamically – can squeeze costs even further without impacting performance. However, our research suggests businesses are leaving $6bn of cost savings on the table by using only on-demand pricing. A bit of time and effort can deliver huge savings, and cloud providers already make tools available to do this. For the providers, the benefits are better cash flow, greater predictability and lower costs. Many third-party tools can also optimize cloud use, even across multiple clouds, and building an application that spans venues can yield vast savings on direct cloud costs. However, this isn't easy, and companies face a raft of technical, process and people challenges in doing so. The first step for all cloud users should be to look at what they do today and see if optimization can work for them.

A lost opportunity

451 Research's Market Monitor service values the global infrastructure-as-a-service market at $48bn for 2021, and 36% of respondents to our Voice of the Enterprise: Cloud, Hosting & Managed Services, Organizational Dynamics survey are consuming cloud only at on-demand rates – the most

This export was generated by user caterina.iacoviello@spglobal.com at account S&P Global on 9/9/2021 from IP address 147.161.128.254.

© 2021 S&P Global Market Intelligence. All rights reserved.
Using Interactive Document Server technology from Publish Interactive

expensive option. That's a shame, because our Cloud Price Index finds that, across the five hyperscalers, average savings of 36% on the cost of a simple application (consisting of virtual machines, storage and networking) can be made just by using commitment discounts. The Cloud Price Index interactive tool allows enterprises to specify a 'basket' of cloud services that make up an application, and to see how their costs compare with the benchmark for that region and with a 'best case' average that takes discounts into account. This means that, across the globe, cloud consumers are missing out on $6bn of savings, equivalent to 13% of all IaaS revenue.

Why aren't enterprises optimizing? The hyperscalers provide lots of tools to recommend optimization opportunities, and there are excellent third-party software platforms that provide advice (more on this shortly), as well as a slew of new managed service-provider advisory services, such as from 2nd Watch, Effectual and Cloudreach. We think that many enterprises that default to on-demand consumption just haven't considered how significant the savings can be or invested in the time and services to explore them. Others may feel their requirements are too 'bursty' to optimize. On-demand provides huge flexibility, but the reality is that most enterprises don't necessarily need to scale up and down on a second-by-second basis. A balanced approach is to use commitment discounts for long-term baseline capacity, then supplement with on-demand as needed.

Another question: Why do hyperscalers provide enterprises with the ability to make savings? The primary purpose is to obtain commitments so that investments can be made in infrastructure with the knowledge that they will be paid off. For example, reserved instances, originally offered by AWS and now for sale on Microsoft Azure and Google Cloud, allow buyers to achieve 70% or higher savings on virtual machines by committing for up three years. With the capital from these aggregated purchases and the confirmation of usage for three years, the hyperscaler can make a more accurate forecast of usage. This allows it to buy bulk hardware with the knowledge that it will be used and not wasted, and allows it to negotiate power and other expenses using the capital and guaranteed cash flow as leverage.

Is optimization a threat to cloud providers? We think not. The fact that such significant discounts are available shows the value of this commitment. Although revenue could go down, commitment essentially reduces the hyperscalers' cost base, so margin is improved. Note that reserved instances are only available for compute, which is why the Cloud Price Index average discount is far lower when we factor in bandwidth and storage, which aren't subject to such big savings (although we do factor in volume-based discounts for storage and other services). Even if the market optimized overnight, it would still represent just a drop of 13%. Furthermore, cloud services are only likely to grow. The Jevons Paradox is in play here: Cost savings from cloud won't necessarily drive overall cost savings for the enterprises; the savings will likely be spent on new services to improve productivity and derive new revenue for enterprises.

In this analysis, we're not factoring in rightsizing, the use of spot instances, orphaned resources or other opportunities to reduce spending. Our estimate is conservative as a result, and the opportunity is likely to be much larger.

Optimizing derives greater cost benefits than moving

Optimization is far more beneficial to enterprises than moving providers. For our small basket benchmark application, the Cloud Price Index tool reveals that enterprises can make – on average – direct cloud cost savings of 18% by changing providers. But this doesn't account for the cost and the hassle of switching providers, nor does it consider differences in product sets and capabilities. Enterprises should look to optimize their costs before even thinking about switching providers.

How about the cost benefits of multicloud? Most enterprises are pursuing a hybrid or multicloud model to allow workload migration between venues as required, although in practice this occurs

© 2021 S&P Global Market Intelligence. All rights reserved.

rarely: 49% of enterprises say they move workloads between on- and off-premises locations just once or twice a year. Even fewer build applications that span multiple environments to take advantage of lower costs and differentiated capability – just 17% of enterprises see allowing a single application to seamlessly take advantage of multiple infrastructure environments as a driver of hybrid adoption. However, that minority could make massive savings of 62% by mixing and matching cloud services across environments. This translates to untapped savings on direct cloud costs of $24bn annually. Although direct costs can be reduced substantially, the complexity of managing these multiple environments is vast. For each environment, you need support teams that can understand each platform's nuances (and the nuances of them working together), and you need developers that can build an application that is performant and resilient across competing providers. For most cloud buyers, this hassle is not likely to be worthwhile, but as noted in the cost arbitrage section below, vendors and startups are working to make this easier for commodified compute resources.

Simple ways of optimizing

Basic techniques for optimizing cloud spending can be divided into three categories: commitment discounts, rightsizing and cost arbitrage.

- **Commitment discounts** offer savings of 70% or more in exchange for making an up-front purchase or committing to a set level of monthly spending. These plans are most appropriate for predictable workloads with steady usage. AWS offers Reserved Instances for specific VM types, including database, machine learning and container instances. These not only guarantee capacity, but also lock in the lowest rate. AWS Savings Plans are more flexible, committing buyers to an hourly level of spending over a one- or three-year time horizon, but enabling the switching of instance types as applications and usage evolve. Azure also has Reserved VM Instances with similar savings, plus it offers an additional Hybrid Benefit discount for customers with Microsoft software licenses. Google Cloud Platform's committed-use discount lowers the price for compute resources in exchange for a one- or three-year contract. GCP also offers a moderate (20-30%) sustained-use discount on some VMs that automatically kicks in after using an on-demand instance for more than 25% of a month (182 hours) and increases with greater usage. Advantages of making these commitments are predictable budgeting and low cost, but they are available only on a per-cloud basis and require careful planning – there are no 'rollover' provisions, so these are use-it-or-lose-it resources.
- **Rightsizing** exploits cloud's inherent flexibility to better match resources consumed with workload demand. This approach is best suited to workloads with unpredictable or variable demand. Whereas on-premises deployments typically require provisioning to accommodate usage peaks (i.e., underutilized infrastructure and waste), cloud applications can be architected or managed to flex available resources up and down as usage ebbs and flows. This can be done during initial deployment – choosing a sensible size VM for a given workload based on business needs for performance and availability – and dynamically by setting up autoscaling to grow and shrink (or add and subtract) compute capacity on a schedule or in response to spikes and troughs in demand. All cloud providers offer tools to accomplish this with their own compute services, but a growing number of third-party vendors enable visibility into and control of spending across clouds. Some of these products look at past usage and make suggestions for rightsizing on a forensic basis, but a growing number use machine learning to predictively scale. Vendors include Turbonomic (acquired by IBM), Cloudability (acquired by Apptio), CloudHealth (acquired by VMware), CloudCheckr and Flexera Rightscale.
- **Cost arbitrage** takes advantage of differences in compute pricing – either due to idle capacity at a given hyperscaler datacenter or regional variations – to dynamically tune an application. This method lends itself to long-running workloads. The big three hyperscalers offer low-cost ephemeral instances (called spot instances on AWS, preemptible VMs on Google Cloud Platform and low-priority VMs on Azure) that are priced at steep discounts (80% or more) versus on-demand rates, but they come with a catch – when the provider needs to reclaim the capacity, it can terminate the instance, in some cases without warning. Meanwhile, cloud-native software vendors, service providers and open source projects have made progress in creating abstractions that enable successful use of multicloud resources in a holistic fashion. Examples include CAST AI, which uses a single stretched Kubernetes

© 2021 S&P Global Market Intelligence. All rights reserved.

cluster to move workloads between and across different clouds based on minute-by-minute or penny-by-penny changes in price and availability, and Volterra.

© 2021 S&P Global Market Intelligence. All rights reserved.