The IT leader's guide to optimizing your cloud budget:
Understand how CPU technology factors into cost savings and business performance
Today, with even greater global uncertainty added to the usual business uncertainty, it’s even more important that you use your existing tech resources wisely, and do as much as you can with what you have. That is no less true now that companies are migrating to the cloud to gain the agility and resilience they need to compete.

But there’s a lack of understanding of the role CPU technology plays in optimizing cloud spend. Often, companies don’t realize they have a hardware choice as they are migrating (or once they’ve migrated) their infrastructure. The good news is that best of breed cloud platforms and CPU vendors offer a number of tools you can use to optimize the cost of cloud computing. Getting more out of your cloud resources can translate into more customers served, more issues resolved, and more adaptability for the overall business.

Cost optimization best practices: The basics

Start by auditing your environment to make sure you are following best practices. The goal is to maximize business value while optimizing costs, keeping in mind the most effective and efficient use of cloud resources. When it comes to optimizing costs, there are lots of tools and techniques that organizations can use. But tools can only take you so far. In our experience, there are several high-level principles that organizations, no matter the size, can follow to make sure they’re getting the most out of the cloud.

“Best of breed cloud platforms and CPU vendors offer a number of tools you can use to optimize the cost of cloud computing.”
Cost visibility. This includes knowing what you spend in detail and how specific services are billed, and having the ability to display how (or why) you spent a specific amount to achieve a business outcome. Key capabilities include the ability to create shared accountability, hold frequent cost reviews, analyze trends, and visualize the impact of your actions in real time. Using a standardized strategy for organizing your resources, you can accurately map your costs to your organization’s operational structure to create a showback/chargeback model. You can also use cost controls like budget alerts and quotas to keep your costs in check over time.

Resource usage optimization. Reduce waste in your environment by optimizing usage. The goal is to implement a specific set of standards that draws an appropriate intersection between cost and performance within an environment. This is the lens to look through when reviewing whether there are idle resources, better services on which to deploy an app, or even whether launching a custom VM shape might be more appropriate. Most companies that are successful at avoiding waste are optimizing resource usage in a decentralized fashion, as individual application owners are usually the best equipped to shut down or resize resources due to their intimate familiarity with the workloads. In addition, you can use Google Cloud’s Recommender tool to help detect issues like under- or over-provisioned VM instances or idle resources.

Pricing efficiency. This includes capabilities such as sustained use discounts, committed use discounts, flat-rate pricing, per-second billing, or other volume discounting features that allow you to optimize rates for a specific service. These capabilities are best leveraged by more centralized teams within your company, such as a Cloud Center of Excellence (CCoE) or a FinOps team that can lower the potential for waste while optimizing coverage across all business units.
Optimizing compute costs

When companies migrate to the cloud, the first step is typically to set up virtual machines (VMs) to run workloads on the cloud platform. And the key to optimizing compute costs is to evaluate the price-to-performance ratio.

Choose the right CPU for your VM

It may come as a surprise that you actually have a choice when selecting the type of physical CPU to run your cloud workload by selecting the optimal VM. In fact, that choice can have a measurable effect on performance and cost. You can gain immediate cost efficiencies by choosing a CPU with a lower price-performance ratio than alternatives. For example, N2D VMs powered by 3rd generation AMD EPYC™ processors offer 13% lower cost than comparable N-series VMs. In addition, running your workload on a more performant CPU can reduce run time, thus saving you ongoing compute costs. When evaluating VM types, make sure the underlying CPU complements your workload needs.

General purpose is suited for most enterprise apps, medium databases, and web or app serving. Look for a flexible, complete features set, balanced between customization, performance, and TCO. Google Cloud N2D general-purpose VMs powered by 3rd gen AMD EPYC™ processors are the largest general-purpose VMs available, with up to 224 vCPUs. In addition, N2D VMs powered by 3rd gen AMD EPYC™ processors deliver over 30% better price-performance over 2nd gen AMD EPYC™ processors.

Compute-optimized is built for high compute needs such as electronic design automation (EDA), high-performance computing (HPC), high-performance web and ad serving, media transcode, scientific modeling, AI/ML, and AAA gaming. Google Cloud C2D compute-optimized VMs are based on 3rd gen AMD EPYC™ processors. Each VM supports up to 112 vCPUs and 896 GiB memory, along with advanced networking up to 100Gbps, and up to 3TB of Local SSD.

Scale-out workloads are designed to efficiently accommodate cloud deployments with highly variable workloads that require the flexibility to scale seamlessly without concern for over- or underprovisioning. Google Cloud Tau T2D VMs are powered by 3rd gen AMD EPYC™ processors. Tau VMs deliver 42% better price performance over comparable general-purpose offerings and 56% higher absolute performance. The x86 compatibility provided by T2D VMs gives you market-leading performance improvements and cost savings, without having to port your applications to a new processor architecture.

Improved price performance delivered by Tau VMs over comparable general-purpose offerings
Only pay for what you need

1. **Rightsize VMs.** Workload requirements can change over time. Instances that were once optimized may now be serving fewer users and traffic — or, conversely, they may be struggling to accommodate growing usage. By monitoring usage, you should be able to distinguish long-term trends that suggest it’s time to downsize or upsize your machine type based on changes in vCPU and RAM usage.

2. **Identify idle VMs.** The easiest way to reduce your cloud vendor’s bill is to remove resources that are no longer being used, such as proof-of-concept projects that have since been deprioritized, or zombie instances that nobody bothered to delete.

3. **Schedule VMs to auto start and stop.** Production systems tend to run 24/7. However, VMs in development, test, or personal environments tend to only be used during business hours, and turning them off can save money.

Take advantage of pricing incentives

Understanding the incentives your cloud vendor offers can help save money while managing your VM fleet. For example, Google Cloud offers committed use discounts that are ideal if you have a predictable steady-state workload. You can purchase a one- or three-year commitment in exchange for substantial savings on your VM usage.

Automate cost optimizations

The best way to make sure that your team is always following cost optimization best practices is to automate those practices. Labels can help you classify instances according to usage requirements. For example, you could label instances that only developers use during business hours separately from production instances. Autoscaling gives you the ability to automatically increase capacity only when you need it. You can scale up gracefully to handle an increase in traffic, and then automatically scale down again when the need for instances decreases.
Google Cloud and AMD: Better together

VMs powered by AMD EPYC™ processors deliver reliable workload-optimized performance with outstanding flexibility, advanced security features, and impressive power and cost efficiencies. A combination of x86 architecture and leading core density creates an attractive price-performance ratio that results in a disruptive pricing model to help customers optimize their infrastructure. Improved Google Cloud VM performance powered by AMD EPYC™ processors allows you to complete tasks more quickly, which then allows your VM to move on to the next query faster, reducing the total usage of compute time.

In fact, AMD-powered VMs on Google Cloud provide up to 13% cost savings over comparable alternatives, which means you can reallocate resources to business-critical needs and innovations. Convert your existing workloads to AMD EPYC™ VMs with a few clicks and count on an industry-standard x86 ecosystem.

"VMs powered by AMD EPYC™ processors deliver reliable workload-optimized performance with outstanding flexibility."
**Tools of the cost optimization trade**

**AMD Cloud Cost Advisor:** Get real-time insights into estimated cost savings when switching to cloud instances powered by AMD within Google Cloud.

**Committed use discount analysis report** in the Cloud Console: Helps you understand and analyze the effectiveness of the commitments you’ve purchased and even estimate what your resource floor looks like based on historical data.

**Custom dashboards:** Get more granular cost views of your environment by attributing costs back to departments or teams using labels. This approach allows you to label a resource based on a predefined business metric, then track its spend over time.

**Google Cloud Deployment Manager:** Automate the creation and management of Google Cloud resources including Compute Engine to create guardrails before you deploy a cloud resource.

**Recommenders:** Google Cloud offers several recommenders that can help you optimize resources, including an **idle VM recommender** that identifies inactive VMs and persistent disks based on usage metrics. The **rightsizing recommender** can show you how to downsize your machine type based on changes in vCPU and RAM usage.

**Cloud Billing reports:** Get an at-a-glance view of your Compute Engine costs.
Lytics improves efficiency to deliver higher margins

Lytics, a customer data platform, requires powerful compute and high storage capacity for its sophisticated orchestration engines. Lytics migrated to Google Cloud specifically to reduce costs. The company saw immediate cost savings per byte and per core as well as improved performance. However, Lytics’ pursuit of efficiency is ongoing.

When Google Cloud released AMD EPYC™ CPU-powered instances, it was time to review the company’s options. Lytics did a three-way test of its existing Intel-based Google Cloud Platform N1 instances against N2 instances and the AMD-based N2D. Based on the results, Lytics began to shift its workloads to N2D instances. “The N2D instances featuring AMD EPYC™ CPUs are really helping platform margins,” says Kathryn Gruenefeldt, VP of Engineering at Lytics. “Looking at the cost to support a customer, if we can bring our costs down, we can pass this on to them to price competitively.”

“The N2D instances featuring AMD EPYC CPUs are really helping platform margins.”

- Kathryn Gruenefeldt, VP of Engineering, Lytics

Results

13% reduction in costs

14–20% better performance
FullStory fuels growth with greater speed and capacity

FullStory provides real-time analytics to thousands of customers across eCommerce retail, SaaS and other digital services. That means running hundreds of instances across thousands of cores. The company wanted to improve the service it provided to its customers by reducing the latencies with their existing platforms and win new business with large clients. FullStory was looking at an expensive expansion of its cloud infrastructure. That is, until AMD EPYC™ CPU-powered Google Cloud instances offered a better strategy. After evaluating the options, the company migrated its database clusters, which can run into hundreds of terabytes each, onto a single N2D instance. According to Jaime Yap, FullStory’s Director of Engineering, the migration was seamless: “There was very little software alteration needed because both are 64-bit x86 architectures, and the results were outstanding.”

The N2D cluster addressed completely the latency concerns for FullStory’s biggest customers, and the company could continue taking on larger clients — without a costly instance expansion and a whole new cluster on top of the existing one. That wasn’t the only savings, however. “When we compared the costs,” says Yap, “the N2Ds were about 13% less expensive than N1.”

“There was very little software alteration needed because both are 64-bit x86 architectures.”
– Jaime Yap, Director of Engineering, FullStory

Results

40% reduction in latency for large clients’ queries

13% reduction in costs

No need for additional database cluster
Qubit gains bullet-proof stability with reduced cost

Qubit’s AI-powered product recommendation engine delivers 9.5 million personalized experiences per hour to 364 eCommerce stores, including some of the biggest luxury brands in retail. This kind of volume requires reliable bullet-proof computing infrastructure, particularly during peaks in traffic when retailers earn most of their revenues. But Qubit had to maintain this reliability while optimizing cost. The company’s Google Cloud representative recommended trying the AMD EPYC™ processor, which offered a like-for-like alternative to the N2- highmem-4 Intel-based instances at a lower cost.

But could they really provide the stability the company’s reputation relied on? To find out, Stephen Boyle, Infrastructure Team Lead, migrated Qubit’s services live to the staging servers, then ran them for three weeks before Black Friday. “We had no issues whatsoever,” says Boyle. “It was super easy. And then the migration started saving us money!” This trial gave Qubit the confidence to move its entire recommendations service over the week before Black Friday, which included 160 N2D instances with four AMD EPYC™ processor cores and 32GB of memory each, running Google Kubernetes Engine (GKE). The migration took just five hours.

“With the savings from migrating to AMD EPYC-powered Google Cloud Platform instances, we can buy new cutting-edge technology and bring that to market quickly.”

- Sergio Iacobucci, Director of Marketing and Partnerships, Qubit

Results

- Rock-solid stability
- Rapid migration
- 13% reduction in costs
Maximize your cloud spend

Migrating your mission-critical workloads to Google Cloud VMs running on AMD EPYC™ processors is the first step toward digital transformation that will give your organization greater agility, flexibility, and efficiency.

In order to get the most bang out of your cloud computing buck, you should incorporate cost optimization strategies into every step of your cloud journey. Powerful and cost-effective, AMD EPYC™ processors are the foundation for cloud cost optimization, and combined with the powerful technology and expertise offered by DoiT, they enable significantly greater savings and performance. In addition, Google Cloud and AMD offer an entire set of cost optimization tools to maximize your cloud spend.

Get started today.
References and sources

1 Blog: New AMD EPYC-based Compute Engine family, now in beta
2 Blog: N2D VMs with latest AMD EPYC CPUs enable on average over 30% better price-performance
3 New Tau VMs deliver leading price-performance for scale-out workloads
4 AMD-powered VMs on Google Cloud provide up to 13% cost savings over comparable alternatives is based on:

- N2D Rome is 13% lower cost than comparable N-series - https://cloud.google.com/blog/products/compute/announcing-the-n2d-vm-family-based-on-amd

- N2D Milan is priced the same as N2D Rome - https://cloud.google.com/blog/products/compute/3rd-gen-amd-epyc-comes-to-compute-engine-n2d-machine-family

- T2D is priced the same as N2D - pricing available at https://cloud.google.com/compute/vm-instance-pricing - for example, T2D on-demand price in Iowa is $0.027502 / vCPU hour, and N2D on-demand price in Iowa is $0.027502 / vCPU hour.

- C2D is up to 13% lower cost than C2 - pricing available at https://cloud.google.com/compute/vm-instance-pricing - for example, C2 on-demand price in Iowa is $0.03398, and C2D on-demand price in Iowa is $0.029563. Using a percentage difference formula, you will find C2D is ~13% lower cost than C2.

5 FullStory case study
6 Lytics case study
7 Qubit case study