Duolingo Reduces Compute Costs by over 60 Percent in One Quarter on AWS

"At its heart, Duolingo is a technology and engineering company. Our main focus is to invest as much of our time and resources as possible in developing the front and backend of our user application while having our lean team manage infrastructure and services as little as possible."

- *Severin Hacker, Co-Founder and CTO at Duolingo*

Embracing Microservices to Drive Rapid Development and Experimentation

Learning a new language is traditionally a time-consuming and challenging experience. Duolingo's mission is to put technology to use and fundamentally change the way individuals learn new languages. Hundreds of millions of people around the world have embraced Duolingo's unique approach to language education through Duolingo's web and mobile apps.

"At its heart, Duolingo is a technology and engineering company," says Severin Hacker, chief technology officer and co-founder at Duolingo. "Our main focus is to invest as much of our time and resources as possible in developing the front and backend of our user application while having our lean team manage infrastructure and services as little as possible."

Duolingo chose to build on Amazon Web Services (AWS) from inception. The company first built a monolithic application architecture on AWS taking advantage of Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Relational Database Service (Amazon RDS). Duolingo has continuously evolved its use of AWS to embrace new managed services, like AWS Elastic Beanstalk, and third-party technologies such as Terraform for infrastructure-as-code management.

As the company continued to scale at a rapid pace while also releasing new features, its teams ran into issues with its architecture. "Our monolith was becoming a bottleneck in terms of scalability and velocity and started to cause deployment headaches," says Hacker. "We also found groups were siloed, with some working on the monolithic application and some on other services."

As a company focused on experimentation, agility, and scalability, a monolithic approach no longer worked, so Duolingo decided to transform its application from a monolithic to a Docker-based microservices architecture. The company began a large-scale migration to Amazon Elastic Container Service (Amazon ECS) managed by Terraform.

About Duolingo

Duolingo is a popular language-learning platform and the most downloaded education app on both the App Store and Google Play with more than 300 million users worldwide. The company's mission is to make education free, fun, and accessible to all.

Challenge

While scaling rapidly, Duolingo decided to transform its application from a monolithic to a Docker-based microservices architecture on AWS. As the Duolingo infrastructure team transitioned to Amazon ECS, they noticed their costs rising and found that some workloads and applications within the same ECS cluster required different types of compute resources.

Solution

Duolingo chose to use Spotinst on AWS to optimize and streamline its infrastructure on AWS. Duolingo uses Spotinst's Elastigroup service, a cluster orchestration and scaling service, ensuring maximum availability for minimum cost.

Benefit

Today, Duolingo manages over 100 microservices on AWS, giving its teams the ability to deploy their own services with speed and ease. Using Spotinst, Duolingo reduced its overall compute costs by over 60 percent in one quarter and its total AWS costs by 25 percent. The money that Duolingo saves using Spotinst is put toward new product development.

Optimizing Compute Usage and Costs using Spotinst on AWS

While Duolingo knew it could solve many pain points by moving from a monolithic to a microservices approach, the transformation created new challenges in terms of cost and instance optimization.

As the Duolingo infrastructure team transitioned to Amazon ECS and became familiar with the service, they noticed their costs rising and found that some workloads and applications within the same ECS cluster required different types of compute resources. The company sought to find a solution that could seamlessly work with Terraform to simplify workload deployment on AWS, reduce costs, optimally mix Amazon EC2 Instance types and sizes, and more efficiently utilize Reserved Instances (RIs) and Amazon EC2 Spot Instances.

After being introduced to Spotinst by AWS and testing the service, Duolingo knew it had found the right solution for the company. "What stood out to me about Spotinst was its pricing model," says Max Blaze, staff operations engineer at Duolingo. "We only pay for Spotinst when the product saves us money."
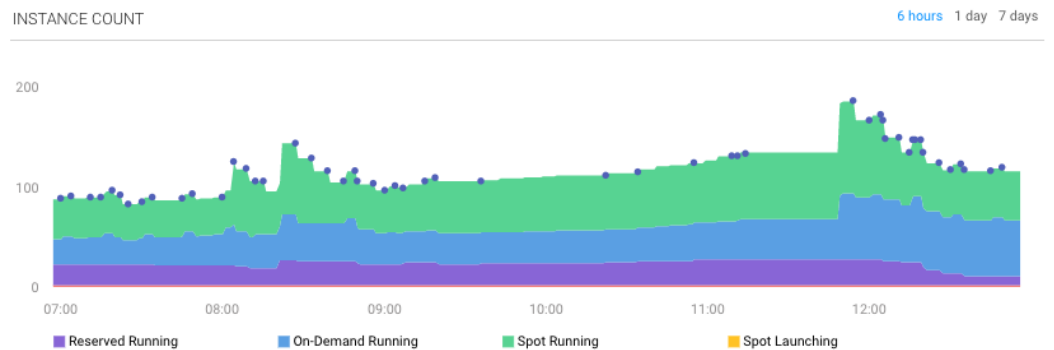

Image 1: Spotinst's Elastigroup automatically finds all unutilized Reservations and prioritizes Reservations usage before launching a Spot Instance.

Duolingo uses Spotinst's Elastigroup service, a cluster orchestration and scaling service, ensuring maximum availability for minimum cost. The service integrates with Amazon ECS and provides support for Terraform. Through its technology's design, Spotinst prides itself on making its customers' lives easier. "What makes us unique is our ability to automate and innovate in our customers' environments without asking them to change anything," says Tomer Hadassi, solutions architecture team lead at Spotinst.

Spotinst's Elastigroup ECS Autoscaler makes real-time decisions for Duolingo about available Spot instance types and automatically detects Availability Zones (AZs) to find the optimal Spot capacity pools based on ECS task requirements. "We've found Spotinst is very flexible and simplifies quite a bit of our setup," says Blaze. "It's helped us to streamline our infrastructure."

Saving Money and Exploring AI Technology to Embolden Language Learners

Today, Duolingo manages over 100 microservices on AWS, giving its teams the ability to deploy their own services with speed and ease. Using Spotinst, Duolingo reduced its overall compute costs by over 60 percent in one quarter and its total AWS costs by 25 percent. The money that Duolingo saves using Spotinst is put toward new product development.

"Spotinst is a rare and fantastic service," says Hacker. "Our team started using the product, and—within a week—we started realizing significant cost savings. We've also found Spotinst's customer service to be great." Duolingo continues to expand its use of Spotinst. The company is now using Spotinst Eco, a new flexible reserved capacity management service, to hand off the existing management of its RIs and not only free up team time to focus on higher-value projects, but to also tap into real-time, deep analytics-based RI management to deliver savings beyond Duolingo's initial savings capabilities.

Looking forward, Duolingo sees massive opportunity in using its data for new development. "We're focused on developing our product to help advanced users and continuously provide all users with better learning tools," says Hacker. "A big area we're investing in is machine learning. We collect so many data points and are asking ourselves, how can we make use of the data we have to improve the products we provide?"

Learn how Duolingo is using deep learning and artificial intelligence on AWS to improve its language learning offerings.