The Stack

This article was published on: 05/4/23

Featured

Amazon Prime Video team throws AWS Serverless under a bus: Saves 90% by ditching Lambda, microservices

Amazon Prime Video has dumped its AWS distributed serverless architecture and moved to what it describes as a "monolith" for its video quality analysis team in a move that it said has cut its cloud infrastructure costs 90%.

The shift saw the team swap an eclectic array of distributed microservices handling video/audio stream analysis processes for an architecture with all components running inside a single Amazon ECS task instead.

Prime Video blasts both barrels at AWS serverless.

(Whether this constitutes what some may think of as a monolith or instead is now one large microservice is an open question; ultimately it is an improved approach to the job that has saved it an awful lot of money…)

Senior software development engineer Marcin Kolny said on Prime's technology blog that toolings built to assess every video stream and check for quality issues had initially been spun up as a "distributed system using serverless components" but that this architecture "caused us to hit a hard scaling limit at around 5% of the expected load" and the "cost of all the building blocks was too high to accept the solution at a large scale."

Strikingly, in one discussion about this decision on Twitter, a senior product engineer at Amazon piped up to tell the world that "We don't use serverless in-house for production loads and no company at sufficient scale should. Pretty sure the docs even say that." (Ladies, gentlemen, non-binary readers; we can't see that in the docs…)

> *"please don't use and spend lots of money on our services for production loads at sufficient scale"*
> *—AWS docs, somewhere, maybe*

— April King 🌀 (@CubicleApril) May 4, 2023

The initial setup had seen the Prime Video team analysing frames and audio buffers using machine-learning algorithms, with AWS Step Functions used as a primary process orchestration mechanism to coordinate the execution of several serverless Lambda functions.

All audio/video data was stored in AWS S3 buckets and an AWS SNS topic was used to deliver analysis results but the cost of passing data around racked up fast.

Join peers following 👉 The Stack on LinkedIn 👈

As Kolny's blog spells out, initial microservices-based tools built for video stream defect detection had hit all kinds of issues: "The main scaling bottleneck in the architecture was the orchestration management that was implemented using AWS Step Functions. Our service performed multiple state transitions for every second of the stream, so we quickly reached account limits. Besides that, AWS Step Functions charges users per state transition. The second cost problem we discovered was about the way we were passing video frames (images) around different components. To reduce computationally expensive video conversion jobs, we built a microservice that splits videos into frames and temporarily uploads images to an… S3 bucket. Defect detectors (where each of them also runs as a separate microservice) then download[ed] images and processed it concurrently using AWS Lambda. However, the high number of Tier-1 calls to the S3 bucket was expensive."

(To some observers, "the design in the PV [Prime Video] article is problematic. Misusing services doesn't fix architecture issues, it exposes them" as Lambros Petrou, a senior software engineer at DataDog added on Twitter; a view to some degree shared by former CTO Steve Chambers, who told *The Stack*: "Basically, they [now] use the same architecture but condense components into containers so they don't have expensive calls and data transfers across a network between different cloud services… [it's] kind of an obvious optimization!"

One engineer added in a Reddit debate on the shift: "Microservices have overhead. What used to be a simple inter-process communication or even an in-memory call between two small parts of a system becomes a full HTTPS, OAuth, JSON encoding/decoding exercise every time one of those short conversations needs to happen. When your system is blown apart into 500,000 pieces and each communication requires that setup, AND you're being billed for each transaction, the cost and complexity adds up. The reaction against monoliths was the need to replace the entire application in one shot,

meaning developers would actually need to test stuff. DevOps means there's no more testing and we fail forward in production, and the only way you can do that is by having tiny functional pieces so you can find/fix stuff fast. I don't think there's anything wrong with saying these super-chatty parts of the application belong together without the need to open millions of connections all the time…"

Amazon Prime Video: Microservices to Monolith? Well, not *quite*

The new architecture.

"The relentless drumbeat of a distributed, microservices-based platform that decouples everything from data, network endpoints to segregated UX with various protocols was maddening without context" commented one global CTO on LinkedIn after reading the post, which was originally filed in March but just attracted attention across the engineering and broadly technology community this month, adding drily: "I wonder if cloud providers are now going to patternize and sell full stack monoliths on their platform."

"This isn't a dig against Lambda as that platform helped the team build the service fast and get to market" as one observer, Kelsey Hightower, noted. "But it is a testament to the overhead of microservices in the real world.

"Moving data around is typically an underestimated cost. A monolithic architecture doesn't mean a spaghetti code base. You should be writing modular code regardless of the deployment model…"

He added: “The initial design was great. It also helped them get to market quickly. Which is huge. Now they get to step back and analyze the next phase. That’s what makes this post so good” – a view echoed by former CTO Steve Chambers who told The Stack: “I think the Prime team are awesome for posting this…”

Amazon invested $7 billion in 2022 across Amazon Originals, live sports and licensed third-party video content included with Prime, its earnings show. Its recent earnings calls have also revealed significant pressure on growth as customers optimise their AWS workloads to cut cloud costs. Even it’s not immune.

See also: AWS Support will no longer fall over with US-EAST-1, after architectural rebuild

Share

Ed Targett

Ed Targett is the founder of The Stack. He has 15 years of experience in newsrooms and consultancies, spanning business technology, capital markets, energy, and sustainability. You can reach him on ed@thestack.technology

Insurance market in mammoth DC migration: 70b rows of data, 200 critical apps, 1 weekend...

Luna ML and the lingerie you didn’t know your life needed

Related Articles

Slack outage: “It’s always DNS” (and, well…)

“We didn’t want to be railroaded into decisions” – Cutting your ERP OpEx and planning the future, with Rimini Street.

Trio of new SolarWinds vulnerabilities gives RCE.

Leave a Reply

Your email address will not be published. Required fields are marked *

Name *

Email *

Website

Save my name, email, and website in this browser for the next time I comment.

Post Comment

Check Also

Luna ML and the lingerie you didn't know your life needed

Amazon Prime Video team throws AWS Serverless under a bus: Saves 90% by ditching Lambda, microservices