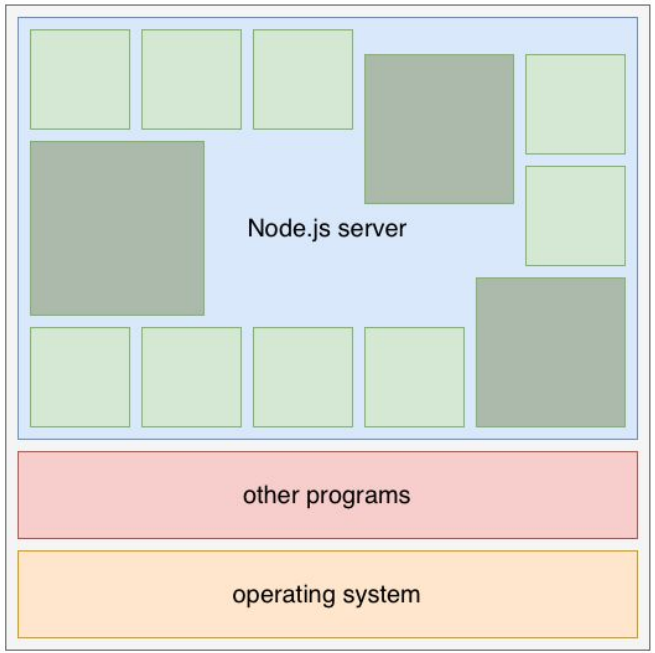
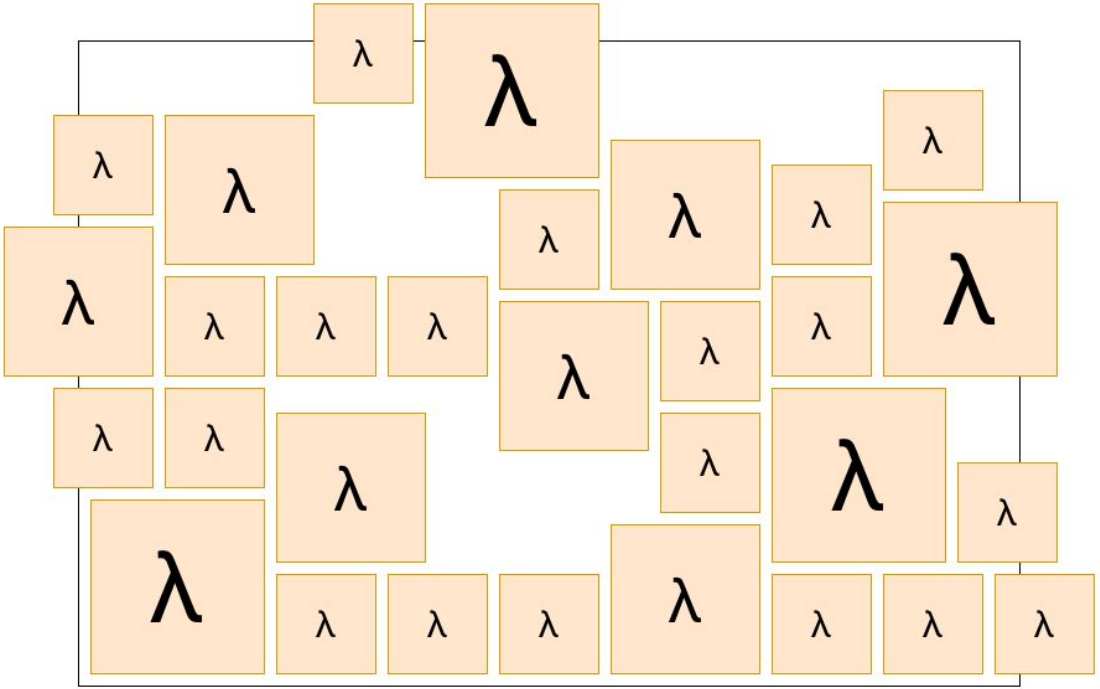
AWS Serverless:
Scaling small and large applications

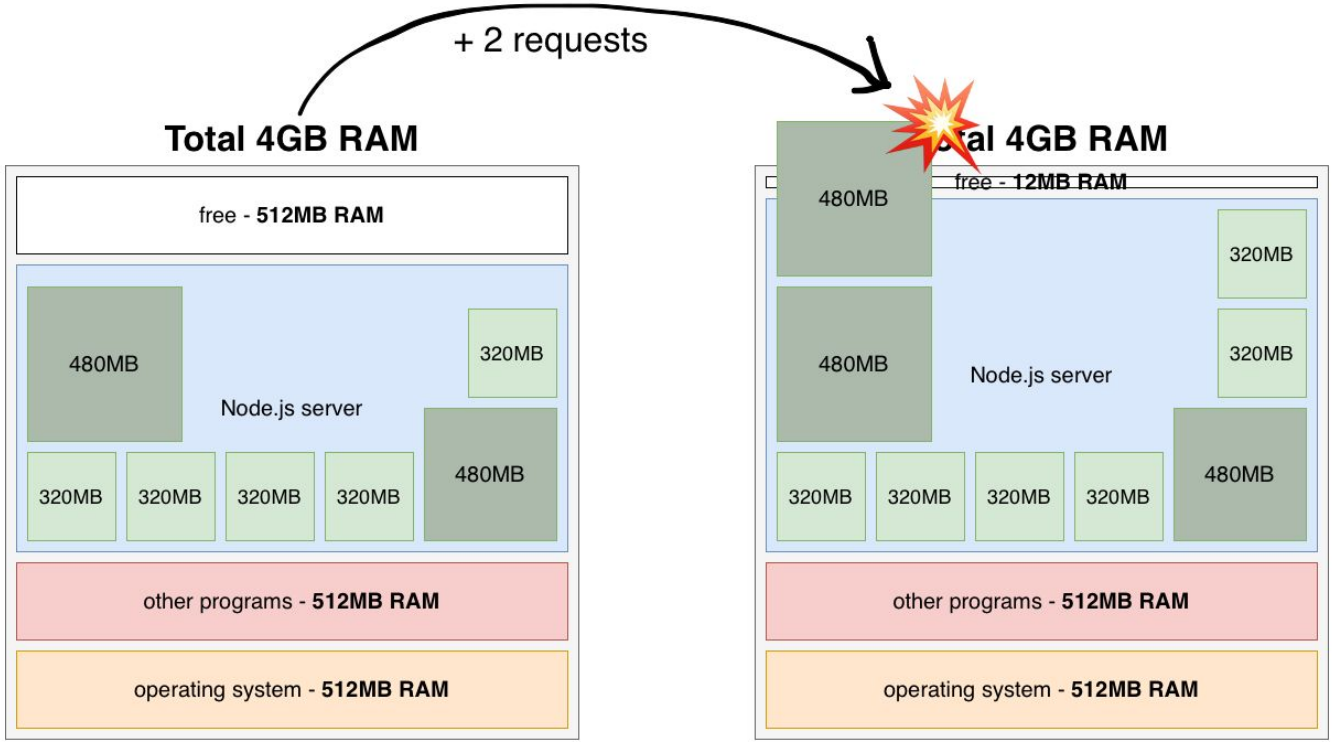
👋 Show of hands

Traditional VM vs Lambda

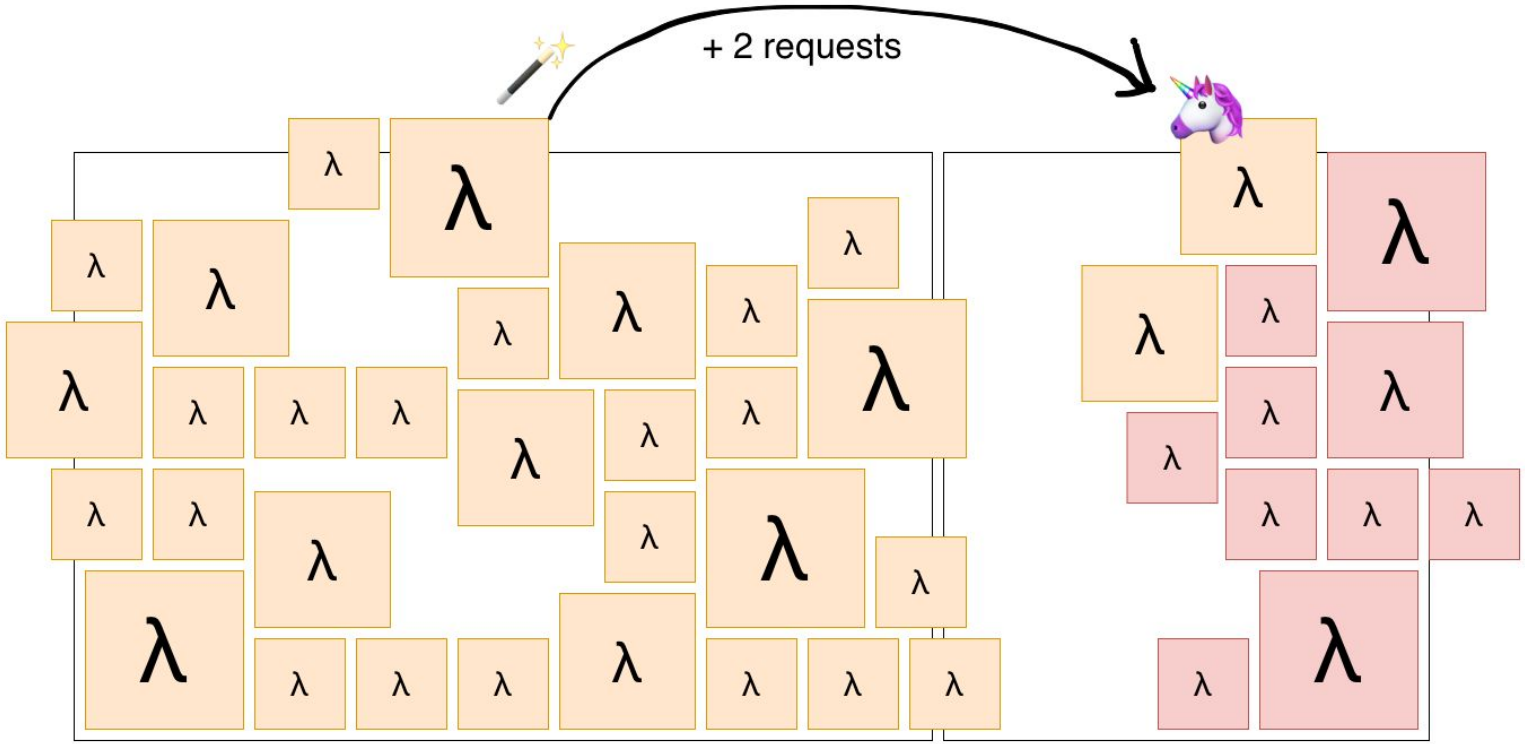




Traditional VM scaling



Lambda scaling



The biggest secret of Serverless

(and many other serverless services)

Hundreds of thousands of
EC2 instances under-the-hood

Usman Khalid

He/His/Him

Head of Serverless Compute

linkedin.com/in/ahmedusmankhalid/

hundreds of thousands of servers.

Lambda limits

15m hard timeout
128 MB - 10240 MB RAM
1 - 6 vCPU
cold starts
telemetry (ADOT has its own limitations)
limited official runtimes
isolation - public service (not great for highly regulated subjects)

Breaking the limits, slowly and steady

15m hard timeout	Durable functions
128 MB - 10240 MB RAM	Lambda Managed Instances (LMI)
1 - 6 vCPU	up to 32 GB RAM + 16 vCPU
cold starts	SnapStart (Java, Python, .Net)
telemetry (ADOT has its own limitations)	OpenTelemetry (Otel) Support #20
limited official runtimes	Container Images, Lambda Web Adapter
isolation - public service (not great for highly regulated subjects)	LMI - runs inside your VPC (not same as VPC support - accessing services in VPC)

AWS Lambda — A Decade+ of Serverless Innovation

43 milestones · Nov 2014 → Apr 2026

2014

NOV 13, 2014
AWS Lambda Announced
Unveiled at re:Invent — Node.js runtime, S3 / DynamoDB / Kinesis event triggers.

2015

APR 9, 2015
General Availability
Production-ready event-driven compute service.

APR 9, 2015
SNS Event Source
SNS notifications can trigger Lambda functions.

JUN 15, 2015
Java Runtime
Java 8 added as a managed runtime.

JUL 9, 2015
API Gateway
HTTP invocation via Amazon API Gateway.

OCT 8, 2015
Python + Versioning
Python 2.7 runtime, versioning, 5-minute execution timeout.

2016

FEB 11, 2016
VPC Access
Functions can reach private VPC resources.

NOV 18, 2016
SAM + Env Variables
Serverless Application Model; environment variables.

DEC 1, 2016
C# + Lambda@Edge
.NET Core runtime; Lambda@Edge runs at CloudFront edge.

2017

APR 19, 2017
AWS X-Ray Tracing
Distributed tracing for Lambda functions.

NOV 28, 2017
Traffic Shifting
Aliases enable canary and linear deployments.

2018

JAN 15, 2018
Go + .NET 2.0
Go and .NET Core 2.0 added as managed runtimes.

JUN 28, 2018
SQS Event Source
Amazon SQS as a Lambda event source.

OCT 10, 2018
15-minute Timeout
Maximum execution duration raised from 5 to 15 minutes.

NOV 29, 2018
Layers + Custom Runtimes
Reusable layers plus custom runtimes — bring any language.

NOV 29, 2018
Ruby + ALB Integration
Ruby runtime; Application Load Balancer target.

2019

SEP 2019
Hyperplane VPC Networking
Hyperplane ENIs deliver fast cold starts inside VPCs.

NOV 25, 2019
Async Destinations
Destinations + retry and event-age controls for async.

DEC 3, 2019
Provisioned Concurrency
Pre-warmed execution environments — no cold starts.

2020

JUN 16, 2020
EFS for Lambda
Mount Amazon EFS file systems inside functions.

DEC 1, 2020
Container Images
Package functions as OCI container images, up to 10 GB.

DEC 1, 2020
1ms Billing Granularity
Billing rounded down to the nearest millisecond.

DEC 1, 2020
10 GB Memory / 6 vCPUs
Memory raised to 10 GB, up to 6 vCPUs per function.

2021

MAY 24, 2021
Extensions GA
Lambda Extensions API GA; observability/security partners.

SEP 29, 2021
Graviton2 (arm64)
arm64 (Graviton2) support — better price / performance.

2022

MAR 23, 2022
10 GB Ephemeral /tmp
/tmp storage configurable up to 10 GB per invocation.

APR 6, 2022
Function URLs
Dedicated HTTPS endpoints for Lambda functions.

NOV 28, 2022
SnapStart for Java
Sub-second cold starts for Java functions.

2023

APR 7, 2023
Response Streaming
Stream response payloads incrementally — over 6 MB.

AUG 2023
12× Faster Scaling
1,000 concurrent executions every 10 sec per function.

NOV 9, 2023
Amazon Linux 2023
AL2023 base for managed runtimes and containers.

2024

2024
Recursive Loop Detection
Auto-detect and stop runaway recursive invocations.

2024
VS Code Console Editor
VS Code-based editor inside the AWS Lambda console.

NOV 2024
SnapStart for Python / .NET
SnapStart extended to Python 3.12+ and .NET 8+ runtimes.

2025

2025
SnapStart on arm64
SnapStart on Graviton2 arm64; additional regions added.

AUG 2025
INIT Phase Billing
Standardized billing for the function initialization phase.

DEC 2025
Durable Functions
Multi-step apps and AI flows; suspend up to 1 year, checkpoint.

DEC 2025
Managed Instances
Capacity providers for GPU and high-memory workloads.

DEC 2025
Tenant Isolation
Separate execution environment per tenant or end-user.

DEC 2025
Kafka + SQS Updates
Native Kafka schema evolution; SQS provisioned mode.

DEC 2025
1 MB Async Payload
Max payload 256 KB → 1 MB for async, SQS, EventBridge.

DEC 2025
Console-to-IDE
One-click IDE handoff; LocalStack offline testing.

2026

APR 2026
S3 Files
Mount S3 buckets as file systems inside Lambda — file API on S3 durability.

Foundational · Runtimes · Integrations / Networking · Performance / Cold-start · Architecture / Compute · Developer Experience · Limits / Pricing

Kiro + Powers

AWS Observability

Comprehensive AWS observability platform combining CloudWatch Logs, Metrics, Alarms, Ap

by AWS

Try power | Uninstall | Check for updates

Build AWS infrastructure with CDK and CloudFormation

Build well-architected AWS infrastructure with CDK using latest documentation, best practice

by AWS

Try power | Uninstall | Check for updates

Build applications with Lambda durable functions

Build resilient multi-step applications and AI workflows with automatic state persistence, retry logic, and ...

by AWS

Try power | Uninstall | Check for updates

IAM Policy Autopilot

AWS IAM Policy Autopilot analyzes your application code locally to generate and deploy identity-based p...

by AWS

Try power | Uninstall | Check for updates

Build applications with Aurora DSQL

Build applications using a serverless, PostgreSQL-compatible database with scale-to-zero and pay-per-...

by AWS

Try power | Uninstall | Check for updates

AWS SAM

An opinionated Kiro Power to aid development with AWS Serverless Application Model (SAM). Includes ...

by AWS

Try power | Uninstall | Check for updates

https://kiro.dev

Lambda is not the only serverless compute option

ECS Fargate

- select # of vCPU and memory
- upload your Docker image
- quick scaling
- fully managed

ECS Managed Instances

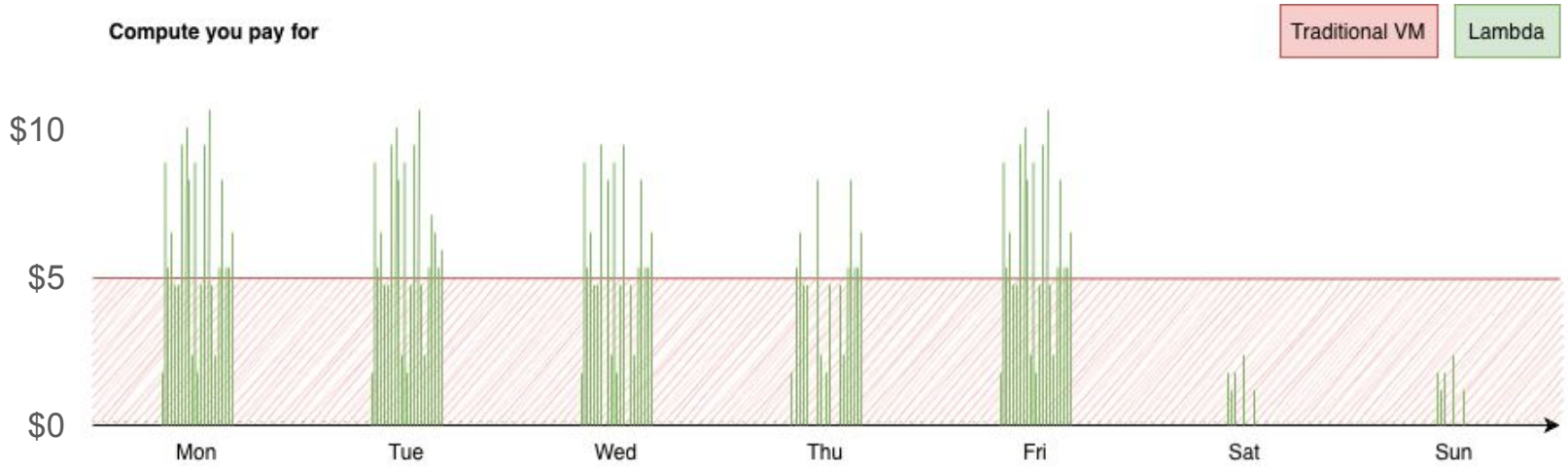
- select the instance type
- managed - no maintenance
- 15% maintenance fee

Traditional VM price comparison

Provider	vCPU / RAM / SSD	Price (original in **bold**)
Hetzner	2 / 4GB / 40GB (ARM chip)	$5.85/month **€4.99/month**
Fly.io	2 / 4GB / 40GB (SSD?)	**$32.68/month** €27.88/month
DigitalOcean	2 / 4GB / 80GB	**$24.00/month** €20.47/month
Scaleway	2 / 4GB / N/A	$19.68/month **€16.79/month**
AWS t4g.medium (70% utilization)	2 / 4GB / 40GB (ARM chip)	**$22.29/month** €19.01/month

Traditional VM limits

- scalability - painful to set up
- durability - single box
- maintenance cost
- overprovisioning - peak load provisioning



Premature optimization trap

1 GB of RAM per request?

You can handle X requests with Rust/Golang on single Raspberry Pi!

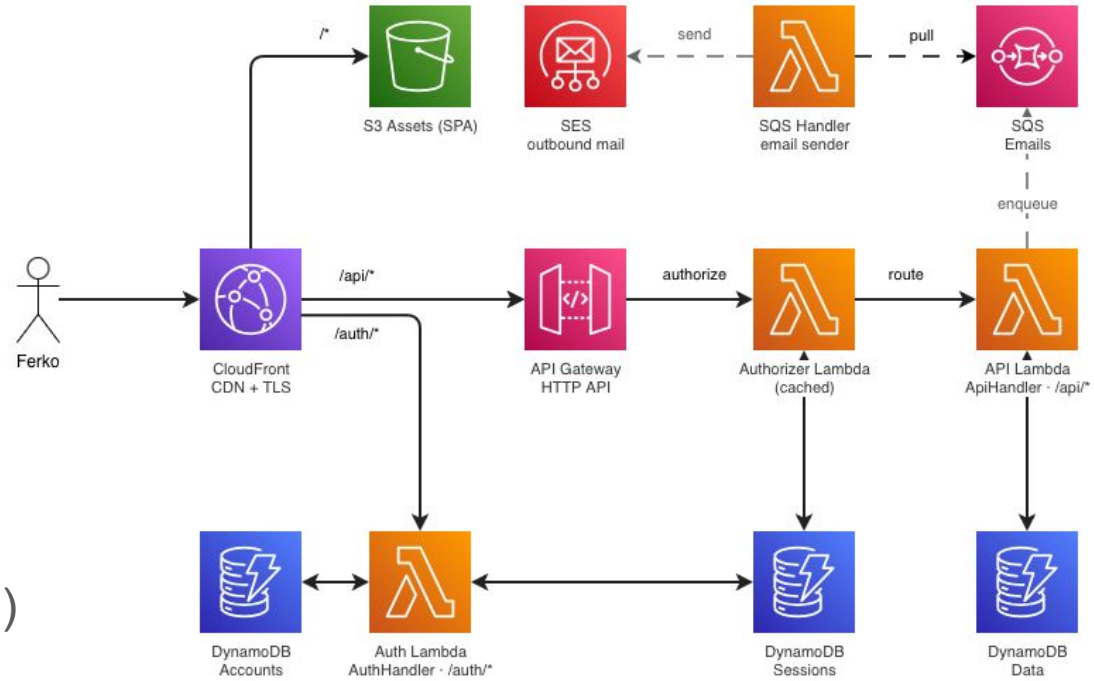
Instead:

- set up billing alerts
- set up budgets for Lambda/DynamoDB
- consider cost of infrastructure per paying customer

Reporting system - simplified architecture

- quarterly reports
- 600 active users
- fully serverless

- TanStack router + MUI
- tRPC
- valibot + formisch
- AWS SDK + CDK
- react-email (pre-rendered)
- Node.js 24

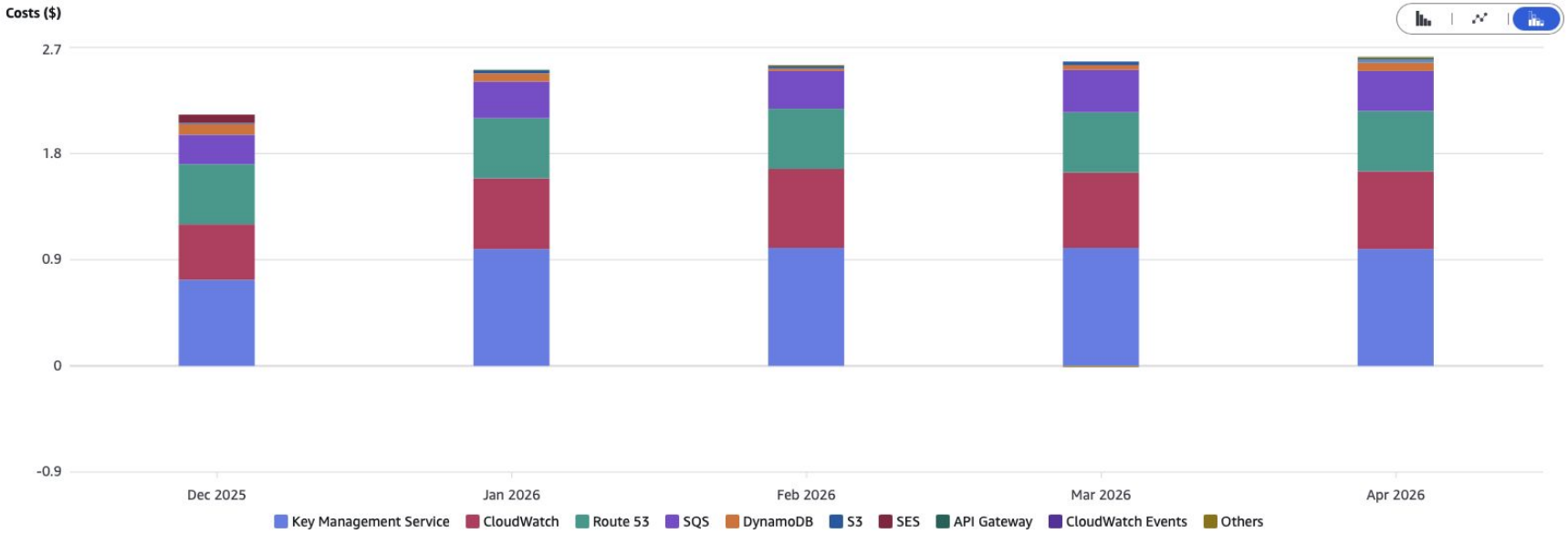


Cost over time (last 5 months)

Cost and usage overview Info

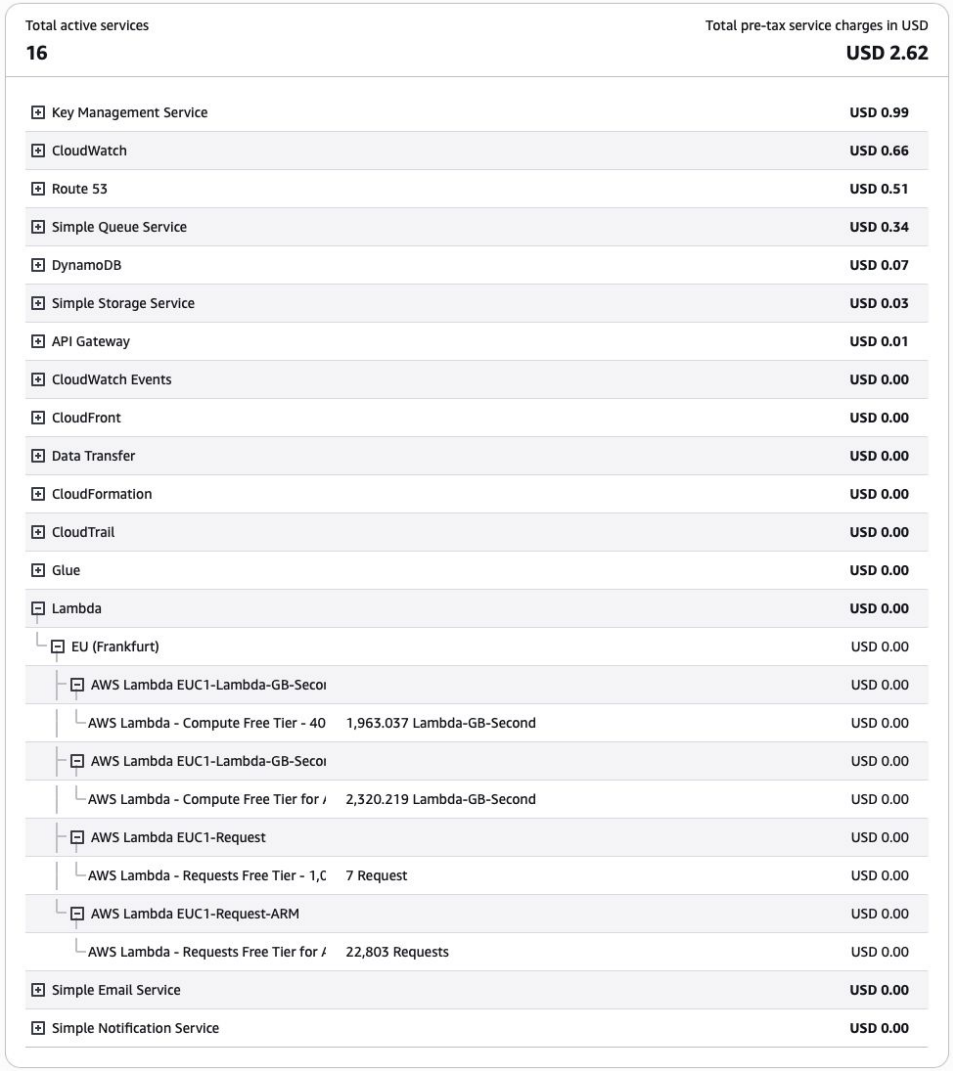
Total cost	Average monthly cost	Service count
$12.39	$2.48	19

Cost and usage graph



April 2026 bill breakdown

- **KMS** - used for DNSSec
- **Route 53** - $0.50 per hosted zone
- **SQS** - even empty queues are being polled constantly
- **S3** - ~500 write/5.5k read reqs
- **API Gateway** - 12k requests
- **CloudFront** - 22k reqs
- **SES** - 300 emails sent



Total active services		Total pre-tax service charges in USD
16		USD 2.62
Key Management Service		USD 0.99
CloudWatch		USD 0.66
Route 53		USD 0.51
Simple Queue Service		USD 0.34
DynamoDB		USD 0.07
Simple Storage Service		USD 0.03
API Gateway		USD 0.01
CloudWatch Events		USD 0.00
CloudFront		USD 0.00
Data Transfer		USD 0.00
CloudFormation		USD 0.00
CloudTrail		USD 0.00
Glue		USD 0.00
Lambda		USD 0.00
EU (Frankfurt)		USD 0.00
AWS Lambda EUC1-Lambda-GB-Secor		USD 0.00
AWS Lambda - Compute Free Tier - 40	1,963.037 Lambda-GB-Second	USD 0.00
AWS Lambda EUC1-Lambda-GB-Secor		USD 0.00
AWS Lambda - Compute Free Tier for /	2,320.219 Lambda-GB-Second	USD 0.00
AWS Lambda EUC1-Request		USD 0.00
AWS Lambda - Requests Free Tier - 1,C	7 Request	USD 0.00
AWS Lambda EUC1-Request-ARM		USD 0.00
AWS Lambda - Requests Free Tier for /	22,803 Requests	USD 0.00
Simple Email Service		USD 0.00
Simple Notification Service		USD 0.00

Lambda logs breakdown (last 2 weeks)

- runtime in ms
- cold starts
- total invocations
- memory usage in MB

Showing 8 of 3,489 records matched

10,077 records (3.2 MB) scanned in 2.2s @ 4,624 records/s (1.5 MB/s)

Hide histogram

Filter table results (case insensitive)...

#	@log	avg_runtime_ms	p50_runtime_ms	p95_runtime_ms	avg_init_time_ms	p95_init_time_ms	total_invocations	cold_starts	avg_memory_mb	p50_memory_mb	p95_memory_mb
▶ 1	LambdaKeepActiveLambdaKeepActiveLogGroup	1743.24	1642.36	2278.64	383.622	415.78	5	5	93.2693	93.4601	94.4138
▶ 2	EmailsSqsHandlerLogGroup	708.8049	933.12	1060.286	339.3181	405.6	41	25	100.8104	100.1358	102.9968
▶ 3	AuthAuthHandlerLogGroup	225.0782	158.309	433.268	240.3548	294.558	47	25	95.9559	97.2748	100.1358
▶ 4	NotifyHandlerLogGroup	188.697	188.697	188.697	418.136	418.136	1	1	102.0432	102.0432	102.0432
▶ 5	AuthSdkHandlerLogGroup	184.9714	55.003	630.454	230.6532	317.694	32	13	81.5392	82.016	82.9697
▶ 6	ApiHandlerLogGroup	153.7809	86.2088	478.6239	653.2916	838.5169	1723	177	179.0727	182.1518	203.1326
▶ 7	AuthApiAuthorizerHandlerLogGroup	120.9565	51.8848	668.3073	220.28	267.7906	1490	160	85.9261	84.877	91.5527
▶ 8	EventBusHandlerLogGroup	67.8353	2.4122	508.2051	352.3026	430.425	150	30	95.3039	95.3674	99.1821

DynamoDB

- manage your schema in application layer (with `valibot`)
- add schema version to each item for easier management
- filter/sort data in memory (for small datasets)
- implement counters/aggregations with transactions/DynamoDB streams
- combine with OpenSearch/Athena when needed

Cost-Aware Modeling in DynamoDB: Designing for Performance & Spend

by Alex DeBrie

DSQL

- distributed PostgreSQL-compatible database
- same principles for distributed systems as DynamoDB applies
- highly available and scalable
- fully serverless
- compute on demand + durable storage

not a PostgreSQL drop-in replacement

- missing foreign keys
- everything must be transaction (optimistic concurrency control)
- async index creation

Serverless tips

presented at

AWS Community Day Slovakia 2026

Be honest with yourself

What is the actual scale of the project?

Are you going to deploy internationally or regional?

How many users are you going to serve?

Do you really need relational database?

Are you building for the purpose or out of habit?

Are you the only contributor?

Ivan Barlog

AWS Solutions Architect

Github ivanbarlog
beesolve

Email ivan@barlog.sk

Web barlog.sk
beesolve.com

LinkedIn ivan-barlog

Go build something!