

COBY BENVENISTE

Staff / Principal Software Engineer - Backend, Platform & Infrastructure

Tel Aviv-Yafo, Israel · U.S. Citizen · Open to U.S./Europe relocation · coby.benveniste@gmail.com · +972 58 681 8698
· <https://github.com/probably-not> · <https://linkedin.com/in/coby-benveniste>

PROFILE

Backend and platform engineer with 12+ years designing and operating high-scale production systems across Go, Elixir, distributed services, cloud infrastructure, and data-intensive architectures. Currently a hands-on VP of Engineering at Marketeam.ai, where I architected and built an agentic AI platform from the ground up; previously Founding Engineer at theGist, Principal Backend Engineer at Arpeely, and Staff Engineer at OpenWeb. Track record includes over 1M RPS low-latency services, event-sourced architecture, cloud infrastructure, reliability, and technical leadership from design through production.

WORK EXPERIENCE

Marketeam.ai

VP of Engineering

Tel Aviv-Yafo
July 2023 - Present

- Architected and built Marketeam.ai's agentic AI platform from the ground up in Elixir/Phoenix, owning backend architecture, cloud infrastructure, deployment, reliability, and core technical direction.
- Designed a fault-tolerant agent harness, runtime and conversation orchestration layer supporting LLM tool execution, provider-agnostic model integration, asynchronous workflows, and real-time client interactions.
- Built production workflow and data infrastructure spanning background jobs, event-driven pipelines, semantic search, AWS, Terraform, CI/CD, and end-to-end observability with OpenTelemetry.
- Designed multi-tenant security around OAuth2 and PostgreSQL Row-Level Security, spearheaded the company's SOC 2 Type II audit, and remained hands-on while leading a cross-functional engineering and product team.

The Gist

Founding Engineer

Tel Aviv-Yafo
December 2021 - July 2023

- Prototyped the initial Slack summarization bot in 35 minutes and productionized it into a product installed across thousands of workspaces with roughly 5,000 active users, earning a Slack "Brilliant Bots" Staff Pick and Product Hunt Product of the Day.
- Architected event-driven ingestion and data pipelines across Go services, SQS/Kafka, EventBridge, Step Functions, Lambda, S3, and Athena, keeping infrastructure costs extremely low while processing Slack workspace data.
- As the first hire and founding engineer, acted as chief architect through the product's evolution from Slack bot to Electron/Next.js/NestJS workspace, owned AWS/EKS/Terraform/CI/CD infrastructure, and served as the engineering interface to the Head of AI.

Arpeely

Principal Backend Engineer

Tel Aviv-Yafo, Israel
March 2021 - December 2021

- Architected and operated latency-sensitive Go services for real-time ad bidding on Google's ad exchange, sustaining over 1M requests per second while targeting sub-10ms latency.
- Replaced expensive managed cloud load balancers with an HAProxy fleet on GCP VMs, using floating-IP/DNS failover and autoscaling to handle multi-gigabyte-per-second traffic and save hundreds of thousands of dollars per month.
- Built the company's cross-cluster SRE and observability platform across GKE using Prometheus, Thanos, Grafana, and centralized alerting, including operational and business telemetry for bidding spend and win rates.

OpenWeb (Formerly Spot.IM)

Backend Engineer / Tech Lead / Staff Engineer

Tel Aviv-Yafo, Israel
July 2018 - March 2021

- Progressed from Backend Engineer to Tech Lead to Staff Engineer, leading a small cross-functional squad before moving into a Staff IC role spanning backend architecture, infrastructure, code review, and hands-on development across Ruby, Elixir, and Go.
- Co-designed and migrated major parts of OpenWeb's 2.0 architecture to event-driven Go microservices, using contract-first BFF APIs to isolate clients from underlying service changes across the conversation admin panel, moderation platform, and core conversation system.

- Led the design and delivery of Moderation 2.0 for a network that reached roughly 50M registered users, combining real-time APIs handling several thousand requests per second with SQS/Kafka processing for tens to hundreds of thousands of messages per day, depending on traffic.
- Built moderation workflows using MySQL, Redis, Google Jigsaw's API, and custom rules for abuse prevention, and led Backend Labs work on fault-tolerance improvements in core message ingestion and Kubernetes performance and cost optimization.

Signals Analytics

Software Developer, Data Scientist, Tableau Specialist

Netanya, Israel

January 2016 - April 2018

- Built and maintained an unstructured-text classification system based on analyst-defined rules that became used across company projects, and developed data models for extracting insights from large datasets.
- Became the company's point person for Tableau development, performance, and optimization.

Jamestown, L.P.

Software Developer (Freelance)

Atlanta, GA

September 2015 - December 2015

- Developed a system to automate federal tax reporting for a real-estate investment company.

Leapfrog Services, Inc.

Software Developer

Atlanta, GA

June 2014 - September 2015

- Developed proprietary internal tools used across multiple departments to automate and streamline day-to-day work.

SELECTED OPEN SOURCE & SPEAKING

LiveLoad Elixir

<https://github.com/probably-not/live-load>

Open-source load-testing framework for simulating real, distributed, live load on an application.

Declarative AI Flows with gen_statem Elixir, Erlang/OTP, gen_statem

<https://youtu.be/duEZxrsTUA4>

CodeBEAM Europe 2025 talk about building AI agents and declarative workflows with Elixir and OTP.

A Murder of LiveViews Elixir, Erlang/OTP, AMoC, FLAME

<https://bit.ly/45rs8n1>

ElixirConf EU 2026 talk about running distributed load tests in Elixir and OTP with AMoC and FLAME.

TECHNICAL SKILLS

Languages:	Go, Elixir, TypeScript, Python, SQL, Ruby
Backend & Architecture:	Distributed Services, Event-Driven Architecture, Event Sourcing, Microservices, Serverless, API Design, Contract-First APIs, Backend-for-Frontend (BFF), High-Throughput Systems, Low-Latency Systems, Multi-Tenant Applications
AI & Agentic Systems:	AI Agents, Agentic Harness Orchestration, Agent Sandboxing and Runtime, Fault-Tolerant LLM Tool Calling, Semantic Search
Elixir / BEAM:	Erlang/OTP, Phoenix, LiveView, gen_statem, PubSub, Oban, Broadway, AMoC, FLAME
Infrastructure:	AWS, GCP, EC2, Auto Scaling, GitHub Actions, Kubernetes, Docker, Terraform, CI/CD
Data & Messaging:	PostgreSQL, pgvector, ETS, MySQL, Redis, Kafka, SQS, Elasticsearch
Observability:	OpenTelemetry, Prometheus, Grafana, CloudWatch
Security:	OAuth2, PostgreSQL Row-Level Security, SOC 2 Type II