Groopview Accelerates Live Engagement with a Dual-Nova AI Avatar on AWS

Project Overview

Groopview, a real-time social-streaming start-up, wanted an AI avatar that could understand text and image in the middle of fast-moving group sessions. Avahi built an API for "Dualy," for an event-driven AWS solution that orchestrates two Amazon Nova models—Nova Pro for precise reasoning and Nova Lite for rapid responses—to match each user query to the best service and reply in near-real time. The result is a scalable, low-latency experience that boosts audience interaction and unlocks new revenue streams.

About the Customer

Philadelphia-based Groopview, Inc. operates a mobile platform where creators stream simultaneous front- and rear-camera views, enabling authentic, "in-the-moment" conversations in sports, entertainment, and live events.

The Problem

As Groopview's user base grew, hosts struggled to answer every question, surface contextual insights, and moderate chats quickly enough to keep viewers engaged. They needed an AI co-host that could:

- process mixed-media inputs (text, images) in real time,
- fetch answers from external APIs, and
- deliver sub-second interruption control for hosts.

Failing to solve these challenges risked higher churn and limited monetization.

Why AWS

Groopview selected AWS for its fully managed generative-AI stack and global, low-latency infrastructure. Amazon Bedrock provided turnkey access to the Nova family of LLMs—giving Groopview both speed (Nova Lite) and reasoning depth (Nova Pro) under a single API.

Why Groopview Chose Avahi

As a Premier-tier AWS partner with early-access experience on Nova models, Avahi combined reference architectures, accelerator libraries, and event-driven best practices to deliver a

production-grade solution in just six weeks. Its ability to orchestrate multiple LLMs for cost-and-latency control set Avahi apart.

Solution

Avahi implemented a serverless, multi-layer architecture:

- **Ingress & Control** – Amazon API Gateway routes text and image streams to AWS Lambda for preprocessing.
- **AI Orchestration** – Within Amazon Bedrock, Avahi's agent framework intelligently orchestrates calls to a collection of external APIs, ensuring low latency while preserving contextual continuity:
  - *Nova Lite* + *Nova Pro* for simple queries (<3 s average latency).
  - *Nova Lite* + *Nova Pro* for complex queries (~7 s).

    Lambda merges model outputs, adds citations, and streams responses back to the app.
- **Data Layer** –Session metadata are stored in an Amazon RDS; secrets live in AWS Secrets Manager.
- **Scalability & Observability** – GPU-backed Amazon EC2 g6e instances auto-scale for bursty inference; Amazon CloudWatch provides unified metrics and alerts.

Key Deliverables

- Dual-model Nova orchestration (Nova Pro + Nova Lite)
- Real-time AI avatar with text and vision channels
- Context-management logic
- Event-driven ETL pipeline (Lambda → S3 → RDS)
- IaC, CI/CD runbooks, CloudWatch dashboards & alerts

Project Impact

Groopview launched Dualy on schedule and saw a surge in session stickiness and creator satisfaction. The dual-Nova design cut complex-query latency by more than 80 percent versus Claude Sonnet, while preserving high accuracy and reducing compute cost per interaction.

Metrics

- **~2.5 s** average response for simple queries (Nova Lite and Nova Pro)
- **~7 s** average response for complex queries (Nova Pro and Nova Pro) vs **~45 s** on Claude Sonnet — **80% faster**
- **2** fully supported interaction modes: text, and vision
- **6-week** end-to-end delivery from kickoff to production

Client Name: Groopview, Inc.
Client Business City Location: Philadelphia, PA
Client Business Industry: Social Media / Live Streaming
Services & Tech: Amazon Bedrock (Nova Pro, Nova Lite), Amazon API Gateway, AWS Lambda, Amazon S3, Amazon RDS, Amazon EC2 g6e, Amazon CloudWatch, AWS IAM, AWS Secrets Manager, AWS Systems Manager Parameter Store