CASE STUDY

# How Aneta handles bursty GPU workloads without overcommitting

A pre-seed AI startup used Runpod to support volatile ingest and inference workloads.

| 90% | 200ms | 1 hour |
| --- | --- | --- |
| cost reduction | cold start times | migration time |

## The challenge

Aneta's workloads could require very large GPU capacity for a few weeks, followed by periods of zero usage. Traditional providers pushed the team toward long-term commitments or expensive pay-as-you-go economics.

## The Runpod approach

Aneta started with on-demand GPU pods and began moving toward serverless infrastructure, using existing containerized workloads to deploy quickly.

## The result

Runpod gave the team burst-friendly compute, lower infrastructure cost, fast startup times, and less DevOps overhead so engineering could stay focused on product.