CLOUDIMPACT by Cygnet.One

Case Study

$117k Estimated Monthly AWS Cost Reduction

GPU infrastructure shifted from always-on capacity to demand-based consumption, reducing idle runtime while preserving peak performance.

Challenge

- GPU instances ran continuously despite low utilization.
- Idle compute capacity increased AWS spend.
- Overprovisioning reduced cost-to-serve efficiency.
- Infrastructure costs did not match user demand.

Solution

- Analyzed usage patterns, concurrency, and demand trends.
- Shifted to on-demand GPUs triggered by user activity.
- Right-sized GPU instances for workload needs.
- Added auto-scaling and scale-to-zero for idle periods.

Results

- **$117k monthly AWS cost savings**
- **Significant reduction in idle GPU hours**
- **Improved infrastructure utilization**
- **Peak performance was maintained under demand**

Tech Stack

AWS | CloudWatch | New Relic | Argo CD | Jenkins | Python | PostgreSQL | Redshift

hello@cloudimpact.ai