

Zero-Idle GPU

Running inference workloads on a T4
that costs nothing while idle

MEASURED IN PRODUCTION

~\$13/month

GPU billed on 12 of 48 days

- Real cost data, not estimates
- Dependency pinning that actually works
- Cold start, 504s, and startup order
- Every trap the docs don't mention

Zero-Idle GPU

Running inference workloads on Azure Container Apps
serverless GPU, with measured cost data

ACO

This book is available at <https://leanpub.com/zero-idle-gpu>

This version was published on 2026-08-18



This is a [Leanpub](#) book. Leanpub empowers authors and publishers with the Lean Publishing process. [Lean Publishing](#) is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

© 2026 ACO

Contents

- Introduction 1**
 - What you get 1
 - What this is not 1
 - Why zero-idle matters more than GPU price 1
 - The tradeoff you're accepting 1
 - Architecture 1
 - How to read this 1
- What It Actually Costs 3**
 - Monthly totals 3
 - The daily distribution 3
 - The outlier day 3
 - Verifying whether you're being billed 3
 - The line item nobody warns you about 3
 - Summary 3
- Prerequisites and Layout 5**
 - What you need 5
 - Region and profile availability 5
 - Variables 5
 - A note on model licensing 5
 - Deployment layout 5
 - Cost implications of these choices 5
- Pinning Dependencies 7**
 - The core conflict 7
 - The full pin set 7
 - Bundle base models at build time 7
 - The trap that only fires at inference time 7
 - Debugging strategy when things break 7
 - Summary 8

Cold Start: Startup Order and the 504	9
What happens on a cold start	9
The startup ordering that matters	9
Handling requests during warm-up	9
Startup probe configuration	9
Reducing cold start	9
Setting expectations	9
Summary	10
Building in Azure (and Two Windows Traps)	11
Trap 1: the CLI crashes on character encoding	11
Trap 2: the build survives your client	11
A workflow that survives all of this	11
Tagging	11
Summary	11
Creating the GPU App	12
Environment and workload profile	12
Creating the app	12
Confirming it exists	12
First real request	12
If it doesn't come up	12
Summary	12
Locking It Down	14
Layer 1: source IP restriction	14
Layer 2: OAuth via built-in authentication	14
Layer 3: email domain check in the app	14
Why three layers	14
Verifying	14
Summary	14
Update Deploys and Day-2 Operations	16
The update procedure	16
Rollback	16
Registry cleanup	16
Monitoring cost	17
Routine health checks	17
Summary	17

Every Trap in This Book	18
Build and dependencies	18
The silent one	18
Cold start	18
Cost	18
Access control	18
Platform	18
Closing	19
 Sample: What This Book Covers	 20
The problem	20
What it actually costs	20
What the full book covers	21

Introduction

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

What you get

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

What this is not

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Why zero-idle matters more than GPU price

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

The tradeoff you're accepting

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Architecture

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

How to read this

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

What It Actually Costs

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Monthly totals

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

The daily distribution

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

The outlier day

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Verifying whether you're being billed

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

The line item nobody warns you about

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Summary

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Prerequisites and Layout

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

What you need

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Region and profile availability

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Variables

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

A note on model licensing

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Deployment layout

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Cost implications of these choices

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Pinning Dependencies

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

The core conflict

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Why cu118 and not something newer

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

The full pin set

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Bundle base models at build time

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

The trap that only fires at inference time

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Debugging strategy when things break

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Summary

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Cold Start: Startup Order and the 504

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

What happens on a cold start

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

The startup ordering that matters

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Handling requests during warm-up

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Startup probe configuration

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Reducing cold start

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Setting expectations

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Summary

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Building in Azure (and Two Windows Traps)

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Trap 1: the CLI crashes on character encoding

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Trap 2: the build survives your client

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

A workflow that survives all of this

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Tagging

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Summary

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Creating the GPU App

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Environment and workload profile

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Creating the app

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Confirming it exists

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

First real request

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

If it doesn't come up

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Summary

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Locking It Down

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Layer 1: source IP restriction

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Layer 2: OAuth via built-in authentication

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Layer 3: email domain check in the app

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Why three layers

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Verifying

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Summary

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Update Deploys and Day-2 Operations

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

The update procedure

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

1. Build a new tag

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

2. Deploy and force it up

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

3. Verify

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

4. Scale back to zero

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Rollback

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Registry cleanup

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Monitoring cost

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Routine health checks

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Summary

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Every Trap in This Book

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Build and dependencies

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

The silent one

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Cold start

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Cost

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Access control

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Platform

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Closing

This content is not available in the sample book. The book can be purchased on Leanpub at <https://leanpub.com/zero-idle-gpu>.

Sample: What This Book Covers

This free sample contains the opening of the book and its cost data. The remaining chapters cover the full deployment walkthrough.

The problem

If you run a GPU workload that people use occasionally, you are almost certainly paying for a GPU that spends most of its life idle. An always-on T4 VM runs a few hundred dollars a month. If your actual usage is a few hours a week, most of that is waste.

Azure Container Apps supports GPU workload profiles with `min-replicas=0`. When no requests arrive, the replica count drops to zero and GPU billing stops. The next request brings it back.

This book documents a production deployment built that way, with measured cost data and every trap I hit getting there.

What it actually costs

Measured over 48 days, scoped to one resource group. Billed in JPY; USD approximate.

Service	Per month	Notes
Container Apps (GPU)	\$2-7	Charged only on days of actual use
Container Registry	~\$11	Fixed cost, and the largest line item
Log Analytics	\$0	

Total: **around \$13/month**, including the month spent actively building and breaking things.

GPU charges appeared on 12 of 48 days

Charge that day	Days
\$0 (unused)	36
\$0.01-0.20	4
\$0.20-0.80	5
~\$1.05	2
\$2.30	1

A day of real use costs between four cents and about a dollar. Against an always-on T4 VM, that is close to two orders of magnitude.

The general rule: **this architecture gets better the less you use it**. That inverts the usual advice about reserved capacity, which rewards predictable high utilization.

The outlier day was a mistake, not a spike

The \$2.30 day happened because I raised `--min-replicas 1` to read startup logs during a deploy, then forgot to set it back. In a `min=0` setup, the expensive mistake is never the GPU hourly rate. It is leaving the scale floor raised after debugging.

The surprise: the registry cost more than the GPU

Images are tens of gigabytes (CUDA base plus bundled model weights), every deploy adds a tag, and nothing removes the old ones. Storage billing applies to the total.

If you are evaluating serverless GPU on cost, model registry storage alongside GPU seconds. At low traffic it will likely dominate.

What the full book covers

- **Pinning dependencies** - the torch/transformers conflict, and a trap that passes every startup check and fails only at inference time

- **Cold start** - startup ordering that avoids a 504, probe configuration, and handling requests during warm-up
- **Building in Azure** - no local Docker, plus two Windows-specific traps including a build that outlives the CLI that launched it
- **Creating the GPU app** - the flags that matter and why `--max-replicas 1` is a spending cap
- **Locking it down** - source IP restriction, OAuth, and an app-level domain check
- **Day-2 operations** - the update procedure, automated scale-down, registry cleanup, and cost monitoring
- **Every trap in the book** - an index by symptom, for debugging

The tradeoff you are accepting is a five-minute cold start. The book is honest about when that disqualifies this architecture.