A KUBERNETES FIELD GUIDE

Tokens per Dollar

Running open-weights models like Llama 3.1 in production without lighting your cloud budget on fire.

micro-edition · 9 chapters

vLLM · KEDA · Karpenter

CONTENTS

FRONT MATTER

Who This Book Is For

You run, or are about to run, an open-weights large language model in production on Kubernetes. Maybe it's Llama 3.1, maybe Qwen, maybe Mistral. You've noticed one of three things: the GPU bill is far larger than anyone forecast, the GPUs sit idle most of the day while you pay for them anyway, or latency falls apart the moment real traffic shows up.

This book is for platform engineers, SREs, and cloud architects who are comfortable with Kubernetes — you know what a Deployment is, you've written an HPA, you can read a Helm chart. It does not assume you know anything about how LLM inference works under the hood. That part matters, because most of the money you're wasting is wasted precisely at the boundary between "people who know Kubernetes" and "people who know inference." This book lives at that boundary.

A note on scope: this is a micro-book. It's deliberately short. Every chapter exists because skipping it costs real money. There are no chapters on training, fine-tuning, or prompt engineering. There is no survey of seventeen serving frameworks. We use vLLM because it's the open-source engine most teams standardize on, it exposes the right metrics, and its economics are well understood. Almost everything here transfers to SGLang or TensorRT-LLM with minor translation.

Assumptions and caveats. Examples target vLLM 0.10.x-era releases and Kubernetes 1.30+. YAML is written for AWS in most examples (instance names like `g6.xlarge`), but nothing is AWS-specific in principle — GCP and Azure equivalents are noted where

the mapping isn't obvious. Cloud prices in this book are rounded, on-demand, US-region figures from the time of writing. They will have changed by the time you read this. The *ratios* between them change much more slowly, and the ratios are what the reasoning depends on.

CHAPTER 01

The Meter Is Running

Let's start with the uncomfortable arithmetic, because every decision in this book flows from it.

A single NVIDIA H100 rents for roughly $3–7 per hour on-demand from the major clouds (with some, notably Azure, higher still). An A100 80GB runs $3–5. Even the modest L4 — a 24GB card you'd use for an 8B model — is around $0.70–1.00 per hour once it's wrapped in an instance. Multiply by 730 hours in a month:

GPU	APPROX. $/HR (ON-DEMAND)	$/MONTH, RUNNING 24/7
L4 (24GB)	$0.80	~$584
A10G (24GB)	$1.00	~$730
L40S (48GB)	$1.90	~$1,390
A100 80GB	$5.00	~$3,650
H100 80GB	$6.00	~$4,380

A team serving Llama 3.1 70B on four replicas of 2× A100 is spending about $29,000 a month before they've served a single token to a paying customer. That's the sticker price. The real problem is what you get for it.

Utilization is the whole game

Here's the pattern I've seen at every company I've helped with this, without exception. Traffic to an LLM-backed product is diurnal and spiky. Your users are awake for ten hours a day. Within those ten hours, load swings 5–10x between quiet and busy. But GPUs were provisioned for the worst fifteen minutes of the worst day —