DevFest Mediterranean 2025

Accelerate AI with Cloud Run

Running your model or agent serverlessly

Google Developer Groups

https://linq.es/giulianobr

+ 20 years in the market

+ 10 years in the Cloud

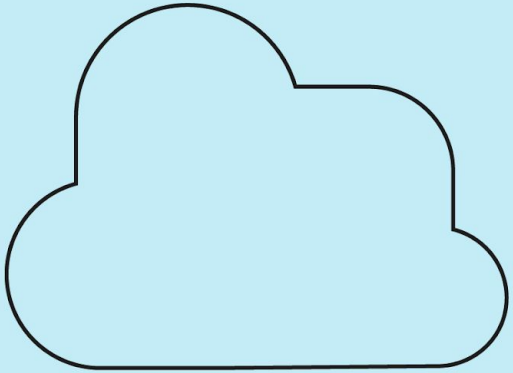
Software Development, Software Architecture, Automation, DevOps, Cloud, FinOps

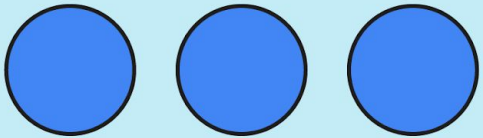
Giuliano Ribeiro

Cloud Architect
Google Cloud Architect/DevOps

DevFest Mediterranean 2025

The Problem: AI Deployment Gap





Google Developer Groups

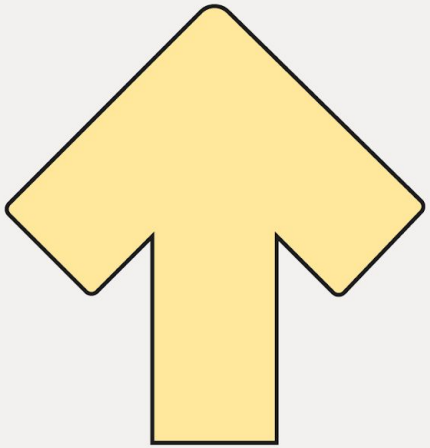
Google Developer Groups

Cost

Complexity

Scalability

Google Developer Groups

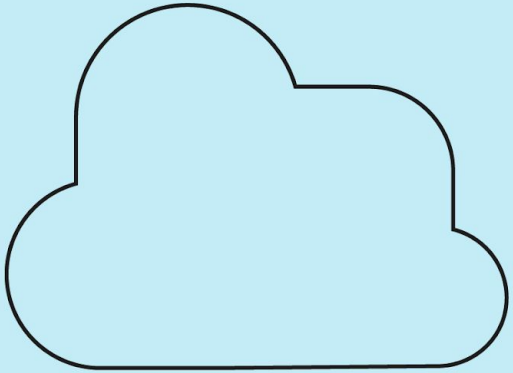


- The Traditional Way (and its pains)
- A Better Way
- Live Demo: Deploying the Gemma LLM with Ollama in minutes
- The Benefits
- When not to use it
- Q&A

DevFest Mediterranean 2025

Google Developer Groups

DevFest Mediterranean 2025



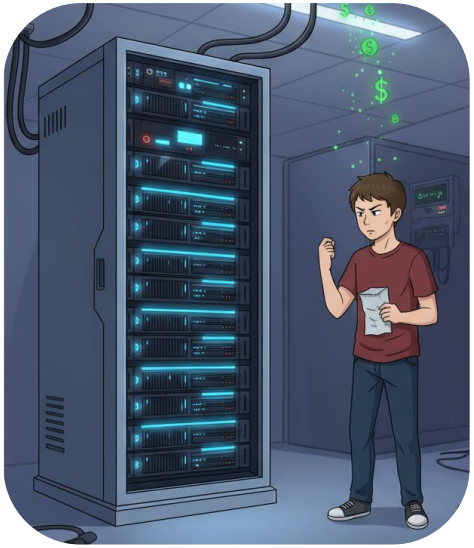
The "Traditional" Approach

Google Developer Groups

The "Traditional" Approach

Google Developer Groups

Why this is painful for AI



Wasted GPU Costs

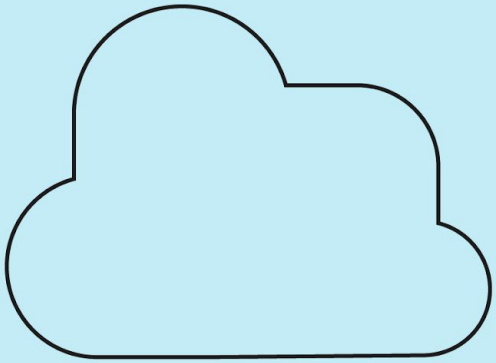
No flexibility

DevOps Overhead

Google Developer Groups

What if you could just focus on your code and let Google handle the rest?

DevFest Mediterranean 2025



Cloud Run

Benefits of Cloud Run

Higher Velocity & Productivity.
Cloud Run allows developers to spend more time **writing code** and **less** time managing **infrastructure**.

Higher Reliability.
Cloud Run is **redundant** by default. Google is your SRE.

Lower Cost.
Cloud Run **autoscales** to meet your needs and scales to **zero**. Pay only for what you use.

95% faster deployment than legacy platforms

98% fewer interruptions to service

15% - 50% cheaper than provisioned platforms

75% cheaper than on-prem

“

Our initial concern about choosing serverless was cost.

It turns out that using **Cloud Run is significantly more cost-effective than running the number of VMs** we would need for a system that could survive reasonable traffic spikes with a similar level of confidence.

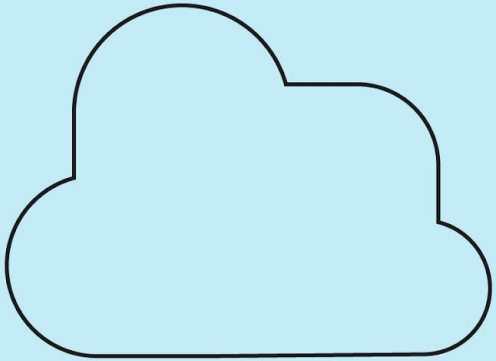
BBC

Cloud Run with GPUs

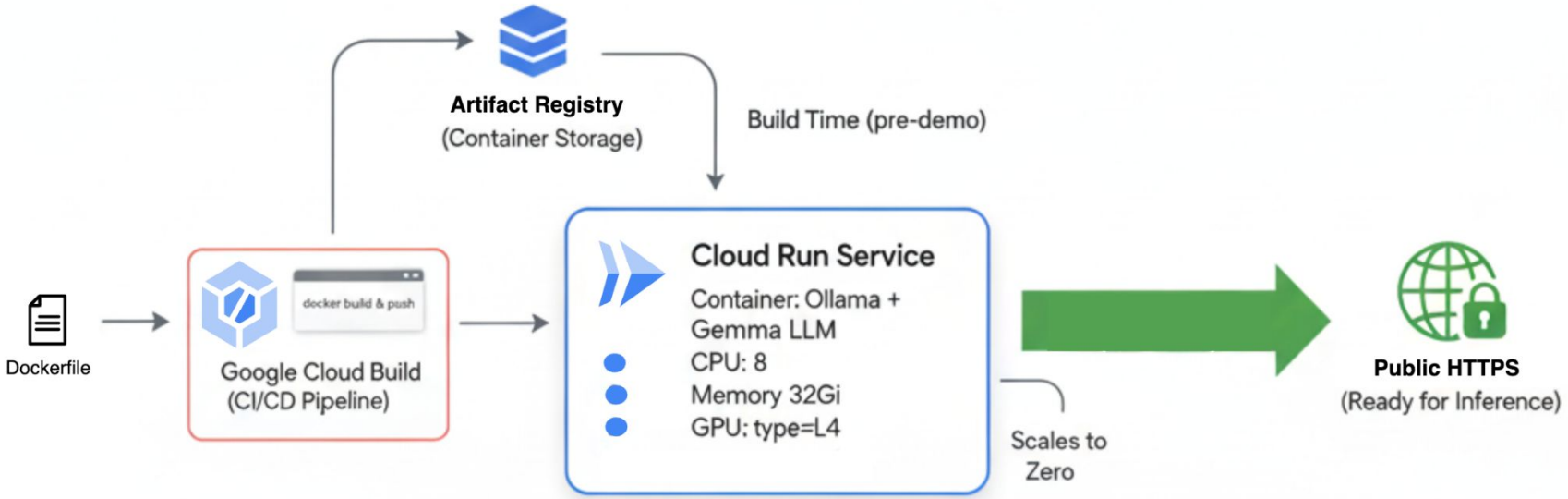
- On-demand
- Hyper-elastic
- Fast starting
- Pay by the second

Live Demo

DevFest Mediterranean 2025



Demo Architecture



Google
Developer
Groups

Dockerfile

```
FROM ollama/ollama:latest
# Listen on all interfaces, port 8080
ENV OLLAMA_HOST 0.0.0.0:8080
# Store model weight files in /models
ENV OLLAMA_MODELS /models
# Reduce logging verbosity
ENV OLLAMA_DEBUG false
# Never unload model weights from the GPU
ENV OLLAMA_KEEP_ALIVE -1
# Store the model weights in the container image
ENV MODEL gemma3:4b
RUN ollama serve & sleep 5 && ollama pull $MODEL
# Start Ollama
ENTRYPOINT ["ollama", "serve"]
```

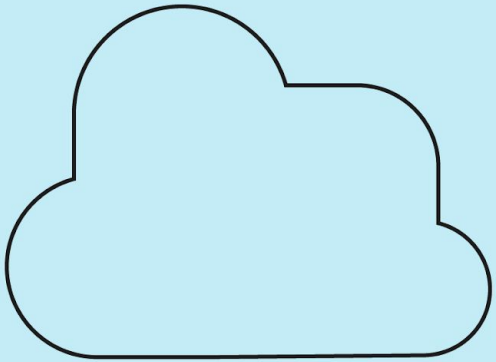
Google Developer Groups

Gcloud run deploy

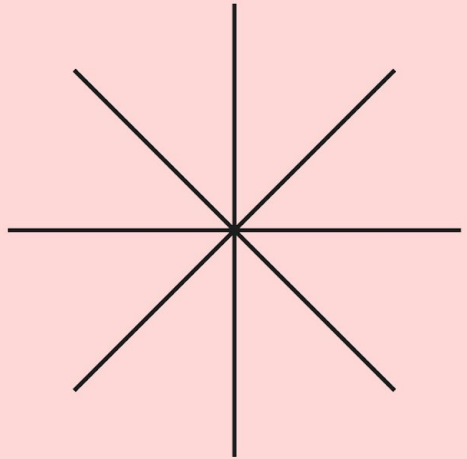
```
gcloud run deploy gemma3 \
   --image us-docker.pkg.dev/cloudrun/container/gemma/gemma3-4b \
   --concurrency 4 \
   --cpu 8 \
   --set-env-vars OLLAMA_NUM_PARALLEL=4 \
   --gpu 1 \
   --gpu-type nvidia-l4 \
   --max-instances 1 \
   --memory 32Gi \
   --no-allow-unauthenticated \
   --no-cpu-throttling \
   --no-gpu-zonal-redundancy \
   --timeout 600 \
   --region europe-west4 \
   --project your-project
```

Google
Developer
Groups

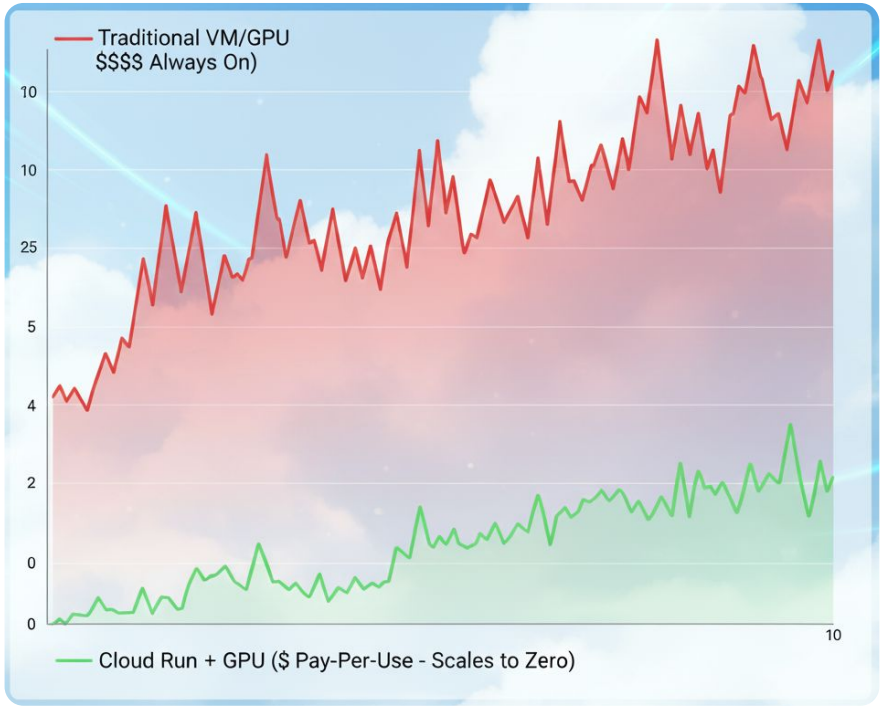
Live Demo



The Impact & Nuances



The Impact: What this means for you

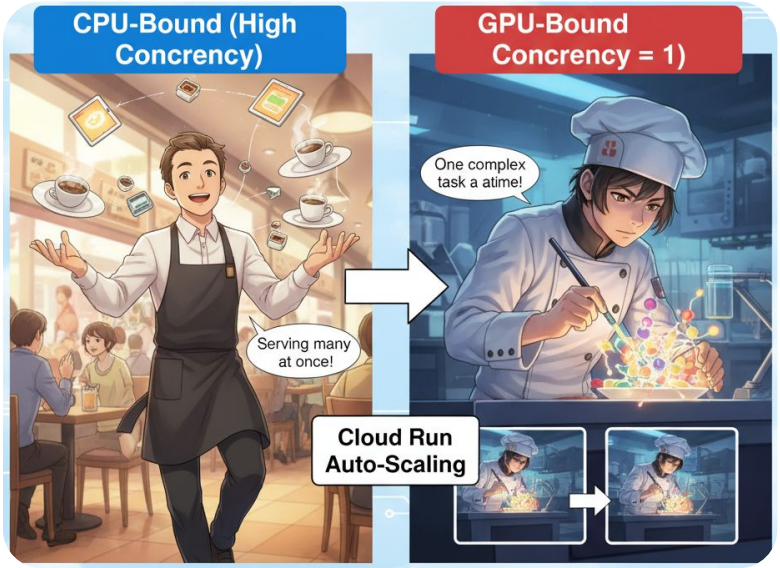


Google Developer Groups

It's Not Magic: How Concurrency Works

80

Concurrency on webapp



1

Concurrency on GPU-bound service

Google Developer Groups

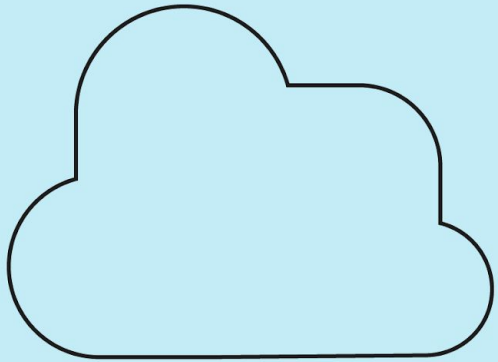
When is Cloud Run NOT the right choice?

- Not for long-running
- Constant High Traffic
- Stateful Workloads

Google
Developer
Groups

Conclusions
&
Q&A

DevFest Mediterranean 2025



Key Takeaways

- Stop Paying for Idle GPUs
- Simplify Your Ops
- Build for Scale from Day One

Google
Developer
Groups

Thank you!

https://linq.es/giulianobr