

Google Cloud

Next '24

Accelerate your
AI with
Serverless



Sara Ford

Senior Developer
Relations Engineer,
Google Cloud



Agenda



01 Benefits of Serverless

02 Demo 1

Deploy a Gemini-powered chat app on Cloud Run

03 Demo 2

Use Cloud Run for Gemini function calling

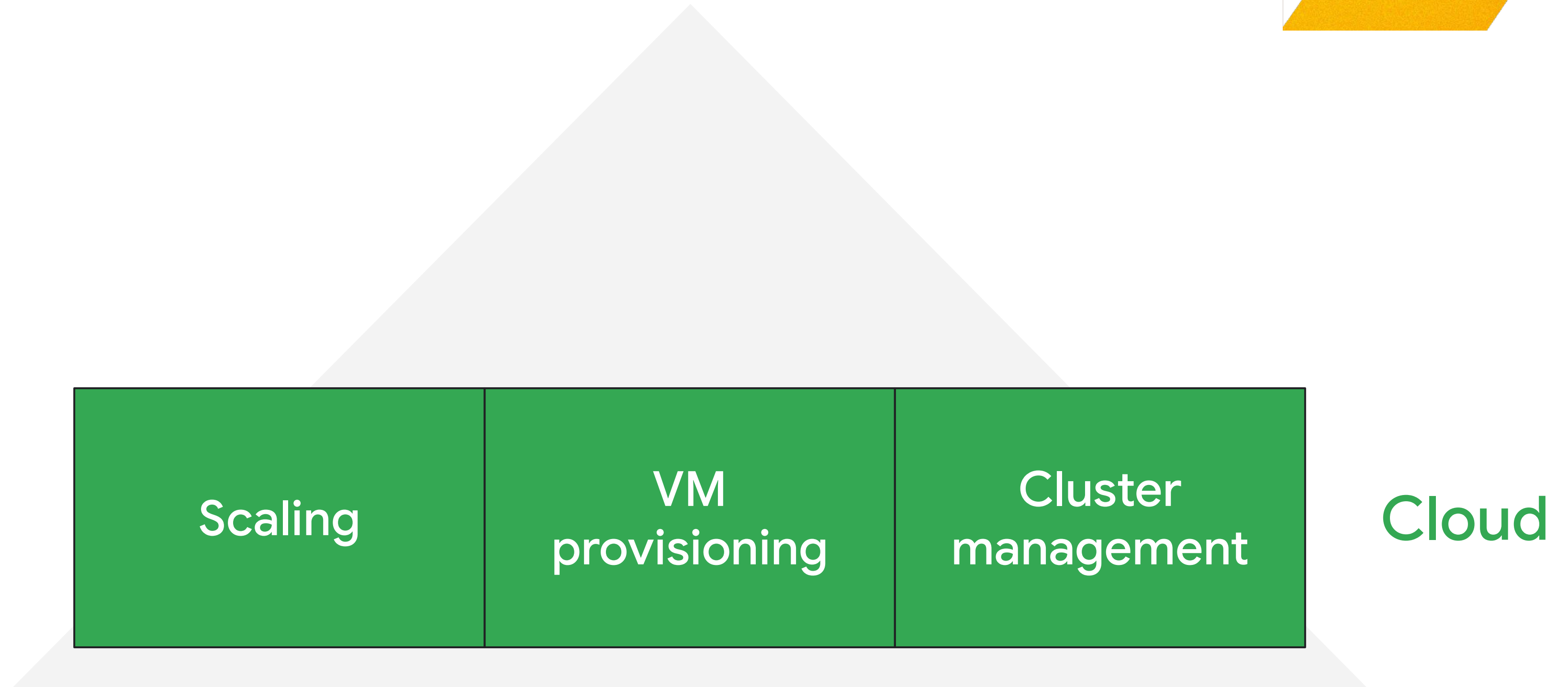
04 Demo 3

Use Cloud Run Jobs & Video Intelligence APIs to process videos

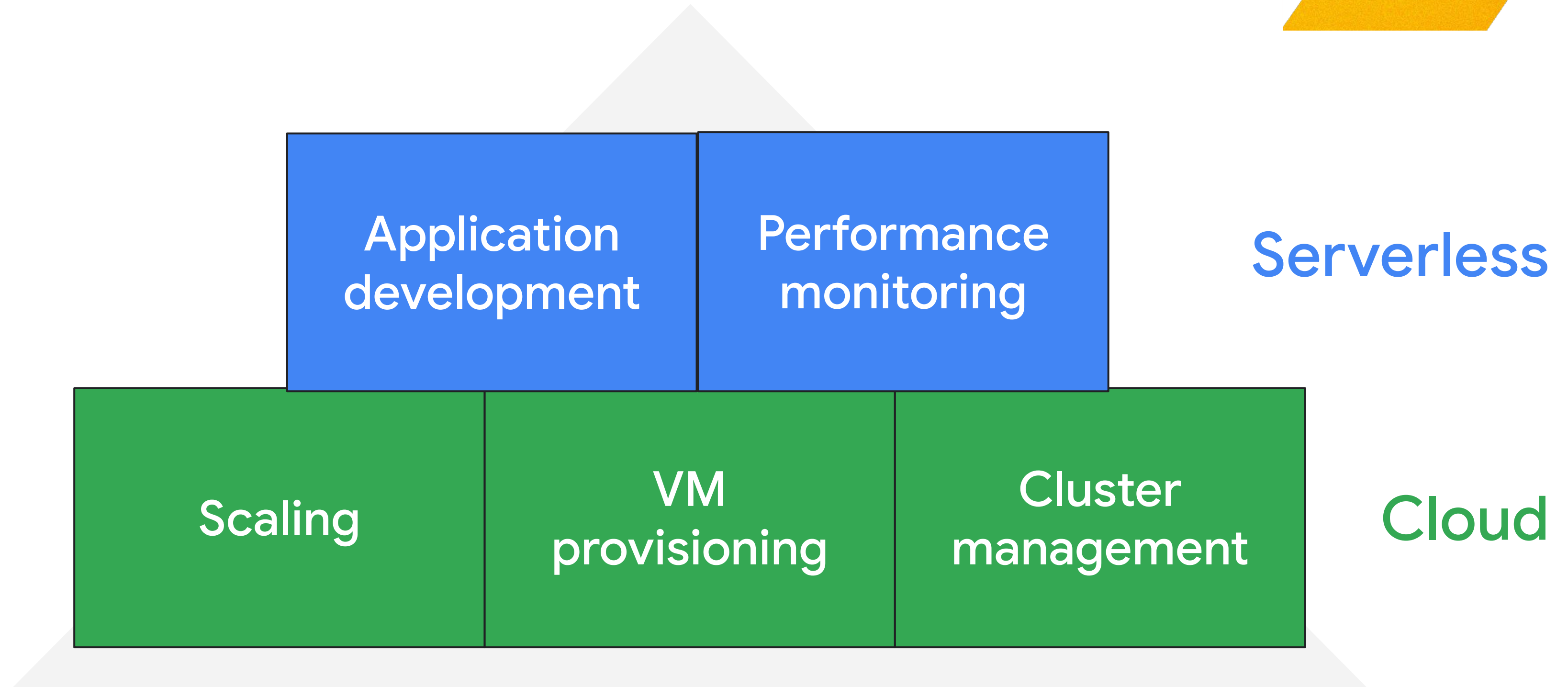
Benefits of Serverless

What serverless offers to developers

Benefits of serverless



Benefits of serverless

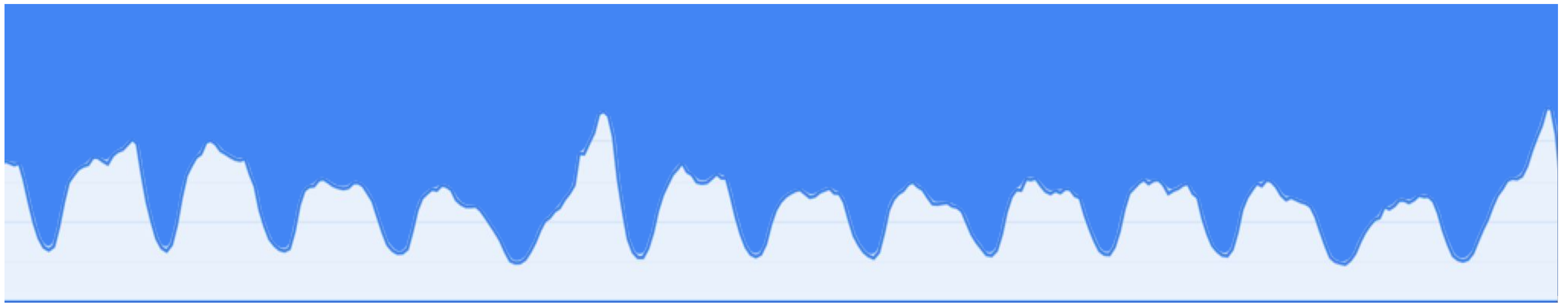


In traditional systems, you need to over-provision resources



50k

Users



Time

But you also don't want to under-provision



150k

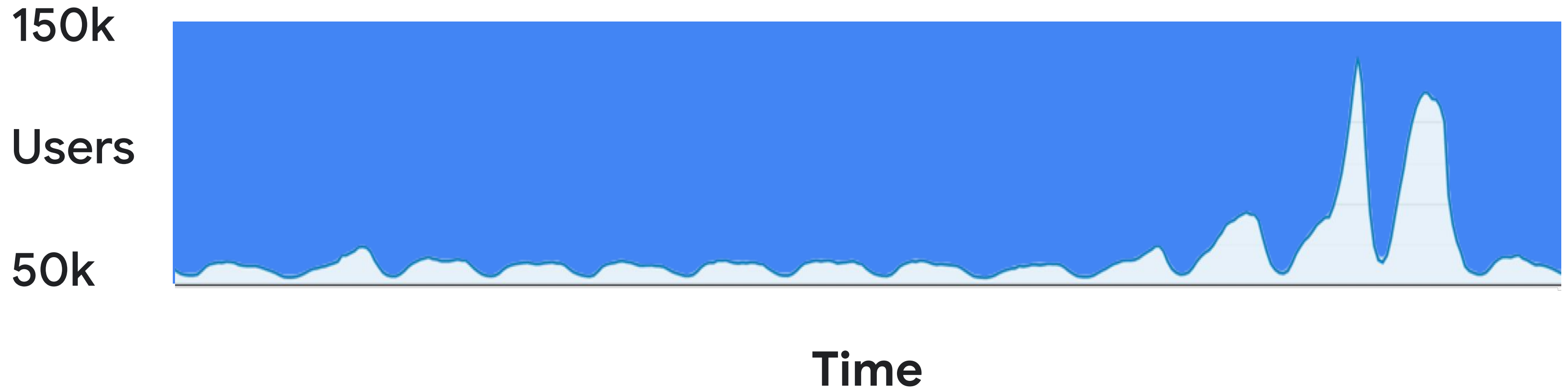
Users

50k



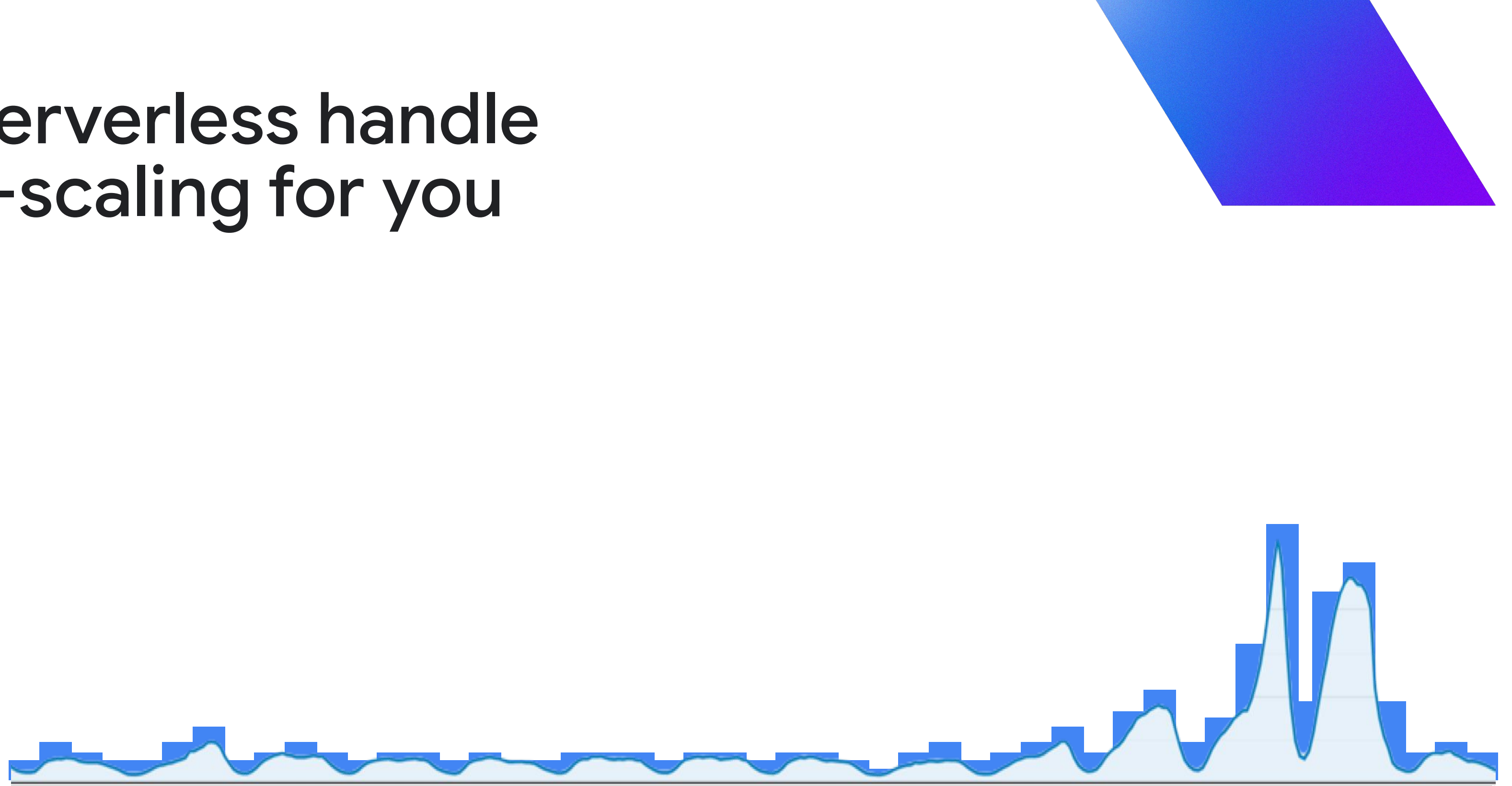
Time

And you may even have to over-provision by a lot!



Let serverless handle auto-scaling for you

150k
Users
50k



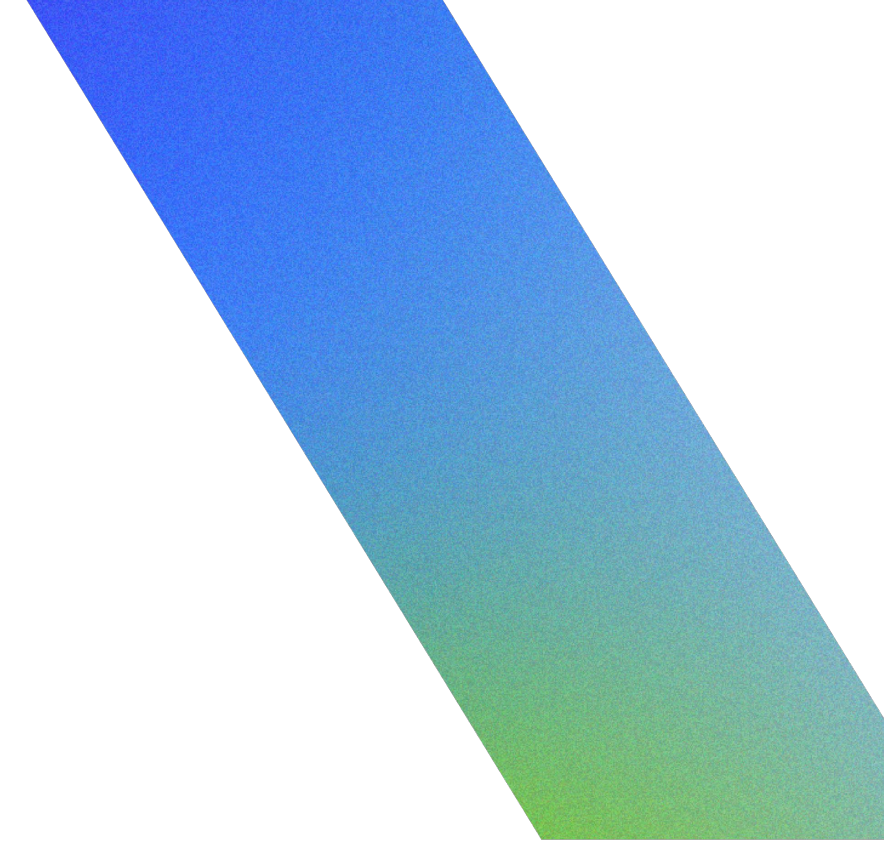
With serverless you pay for the area **under** the curve.

Cloud Run is Google Cloud's serverless engine.

Run applications fast and more secure in a fully
managed environment.

Services

Jobs



Services

Serve HTTP traffic under a dedicated endpoint

- ➔ Automatic scaling with min/max
- ➔ Out-of-the-box URL with TLS
- ➔ Pay when code is processing requests
- ➔ HTTP, events, websockets, HTTP/2, gRPC
- ➔ Built-in traffic splitting for gradual rollouts
- ➔ Revision history

Jobs

Services

Serve HTTP traffic under a dedicated endpoint

- ➔ Automatic scaling with min/max
- ➔ Out-of-the-box URL with TLS
- ➔ Pay when code is processing requests
- ➔ HTTP, events, websockets, HTTP/2, gRPC
- ➔ Built-in traffic splitting for gradual rollouts
- ➔ Revision history

Jobs

Execute tasks to completion

- ➔ No HTTP endpoint
- ➔ No HTTP endpoint
- ➔ Automatic scaling with max
- ➔ Executed manually, or on a schedule
- ➔ Runs a specified number of tasks, up to 24h
- ➔ Execution history

Cloud Run offers additional fine-tuning for auto-scaling



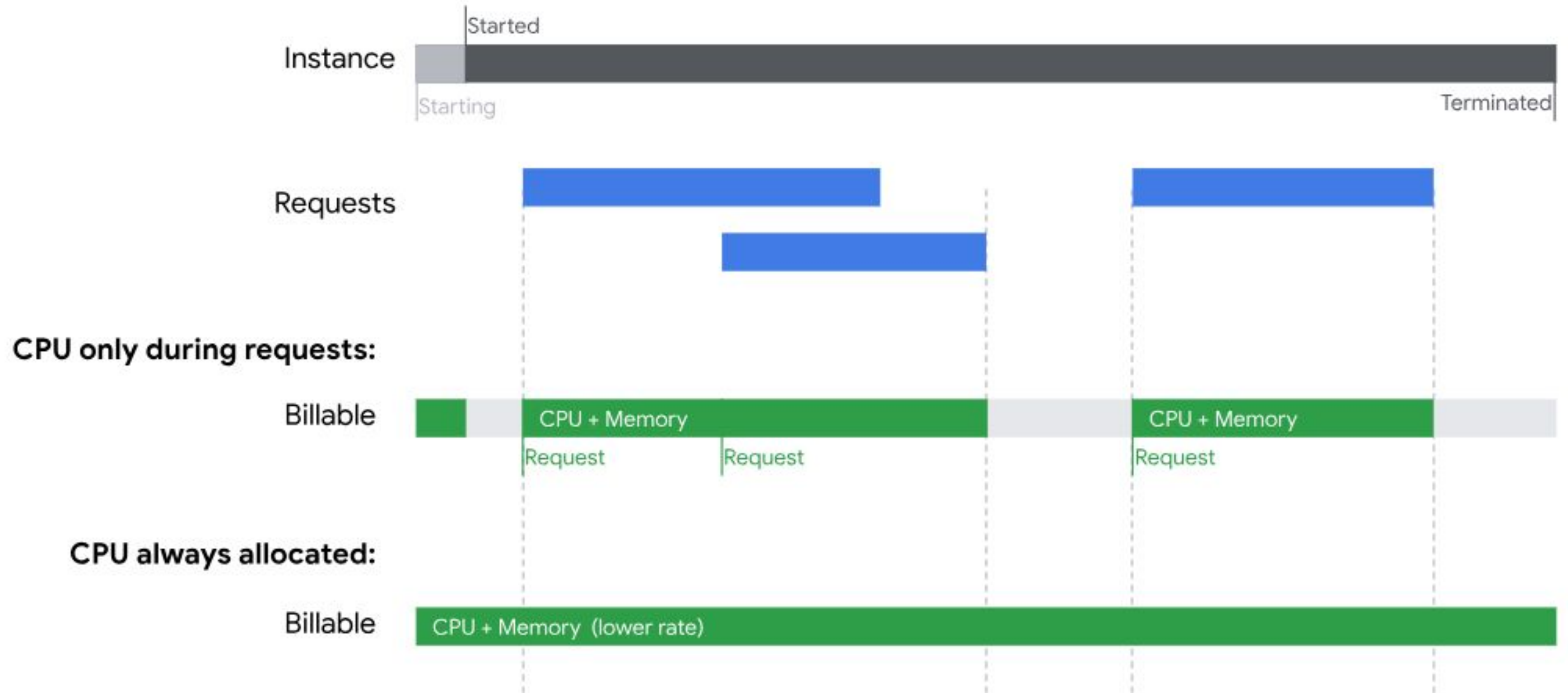
Keep instances warm

Configure minimum instances to avoid scaling to zero instances and improve latency

One instance == many requests

Concurrency refers to the number of requests that can be served **at the same time** from a Cloud Run instance. A service can have many instances

Billing



<https://cloud.google.com/run/pricing>

Demo 1 codelab



Demo 2 codelab



Demo 3 codelab



Demo 1

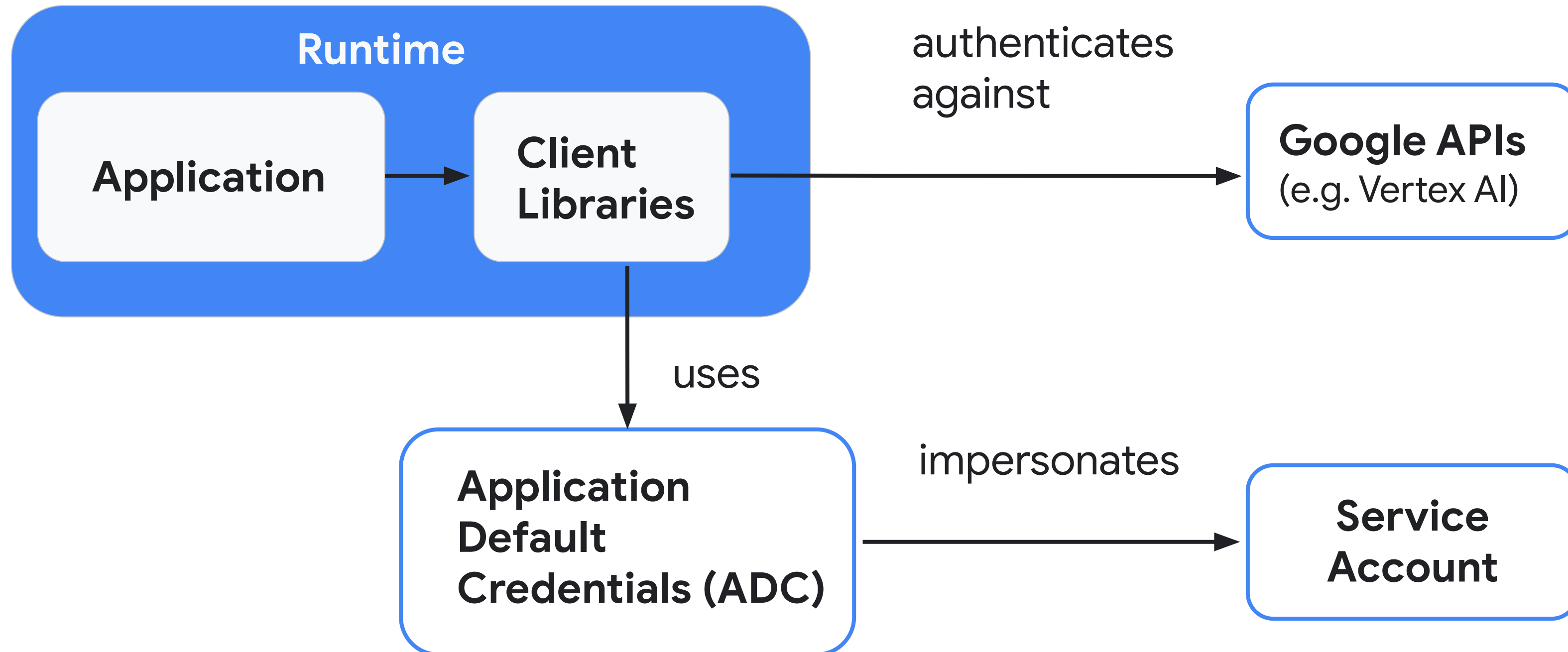
Deploy a Gemini-powered chat app on Cloud Run



What you'll see

- ✓ **Google Client Libraries for Vertex AI**
- ✓ **No dockerfile**

Google Client Libraries handle auth to Google APIs



What you'll see

- ✓ Google client libraries for Vertex AI
- ✓ No dockerfile

Demo



Demo 2

Use Cloud Run for Gemini function calling



What you'll see

- ✓ Gemini chatbot app
- ✓ Uses function calling to a Cloud Run service

"what's the weather like today in Seattle?"

Chat bot app

user prompt +
getWeather(string) function contract

call getWeather("Seattle") for me! 🙏

getWeather("Seattle")

**external API
or service**

{ temp: 40F,
conditions: rainy }

function response is { temp: 40F,
conditions: rainy }

Weather in Seattle is 40F and rainy!

Gemini

Use cases

- ✓ Running asynchronous operations that take more than a few seconds
- ✓ Enhance chatbots with the ability to access and process information from external sources in real-time, e.g. weather, stock prices, etc.
- ✓ Interact with SQL databases using natural language

Demo



Demo 3

Use Cloud Run Jobs & Video Intelligence
APIs to process videos



What you'll see

- ✓ Cloud Run Jobs is cron jobs for the Cloud
(up to 24 hours processing time!)
- ✓ Video Intelligence API and Vertex AI API
(not a Gemini demo)

Ex
De
fo
[1,

Final Output

```
[  
  {  
    timestamp: 1,  
    description:  
      "what is google cloud vision api? is written on a white background"  
  },  
  {  
    timestamp: 3,  
    description:  
      "a woman wearing a google cloud vision api shirt sits at a table"  
  },  
  {  
    timestamp: 8,  
    description:  
      "a woman wearing a red shirt that says "what is cloud vision api"  
  }  
];
```


Demo



Recap

- ✓ How Google Cloud does Serverless
- ✓ 3 ways you can incorporate Serverless into your AI workloads



Ready to build what's next?

Tap into **special offers** designed to help you **implement what you learned** at Google Cloud Next.

Scan the code to receive personalized guidance from one of our experts.



Or visit g.co/next/24offers

Continue your learning journey!

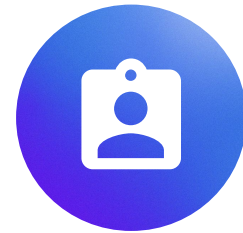


Codelabs

goo.gle/next24-serverless-demo1

goo.gle/next24-serverless-demo2

goo.gle/next24-serverless-demo3



Sessions

DEV253 - Building generative AI apps on Google Cloud with LangChain

DEV205 - Cloud Run: What's new



More Sessions!

DEV236 - Ford Motor Co.'s acceleration to Google Cloud fueled by Cloud Run

ARC104 - The ultimate hybrid example: A fireside chat about how Google Cloud powers (part of) Alphabet

Thank you