Google Cloud

サーバーレスを活用した、立ち上げ期から、拡大期にわたる「アーキテクチャの変遷」

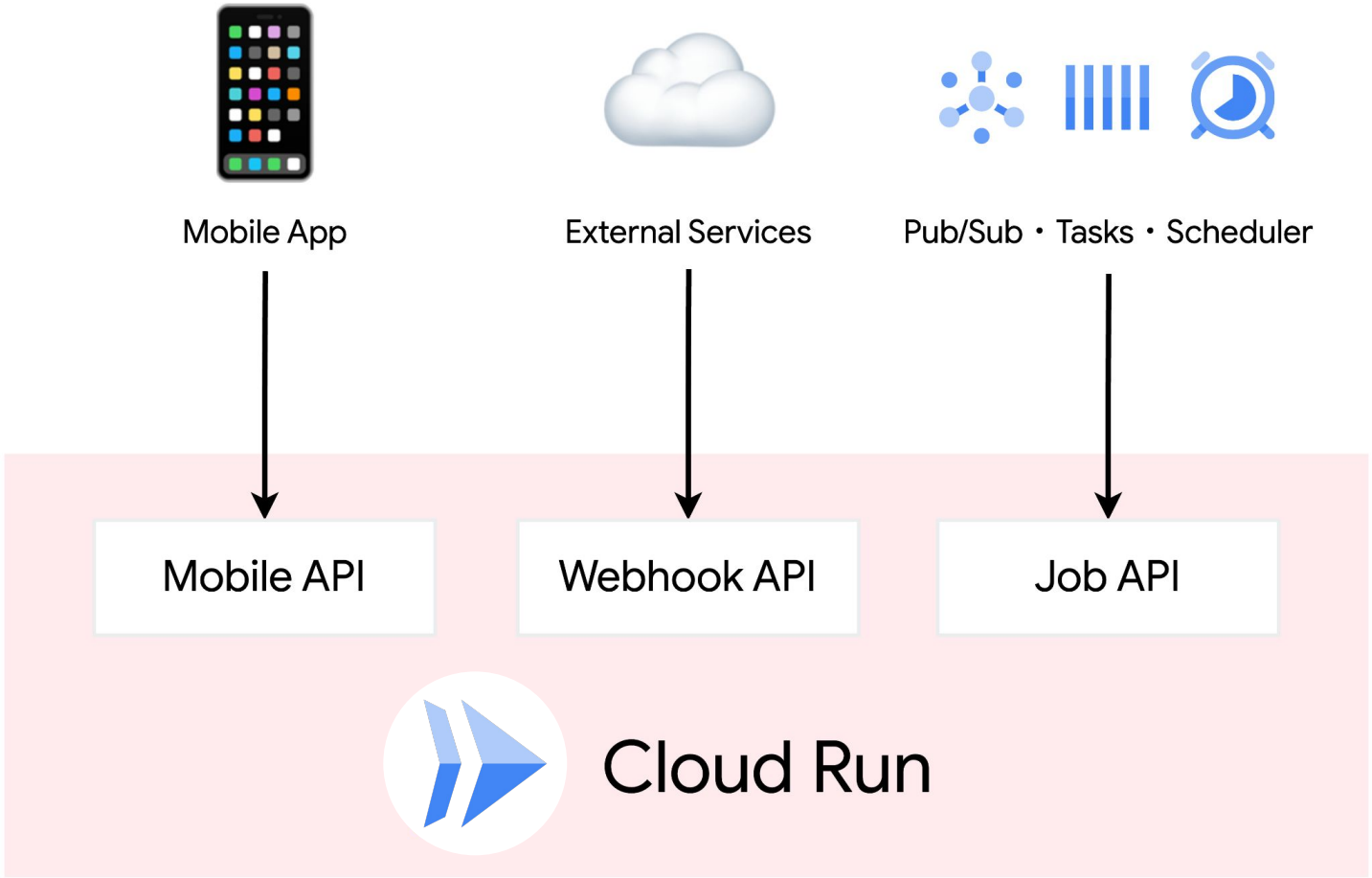
株式会社 カウシェ
Platform Team / Architect
伊藤 雄貴

#シェア買いで
みんなでお得に
ショッピング

Table of Contents

01

Application Architecture


Mobile App
External Services
Pub/Sub • Tasks • Scheduler
Mobile API
Webhook API
Job API
Cloud Run

Cloud Run

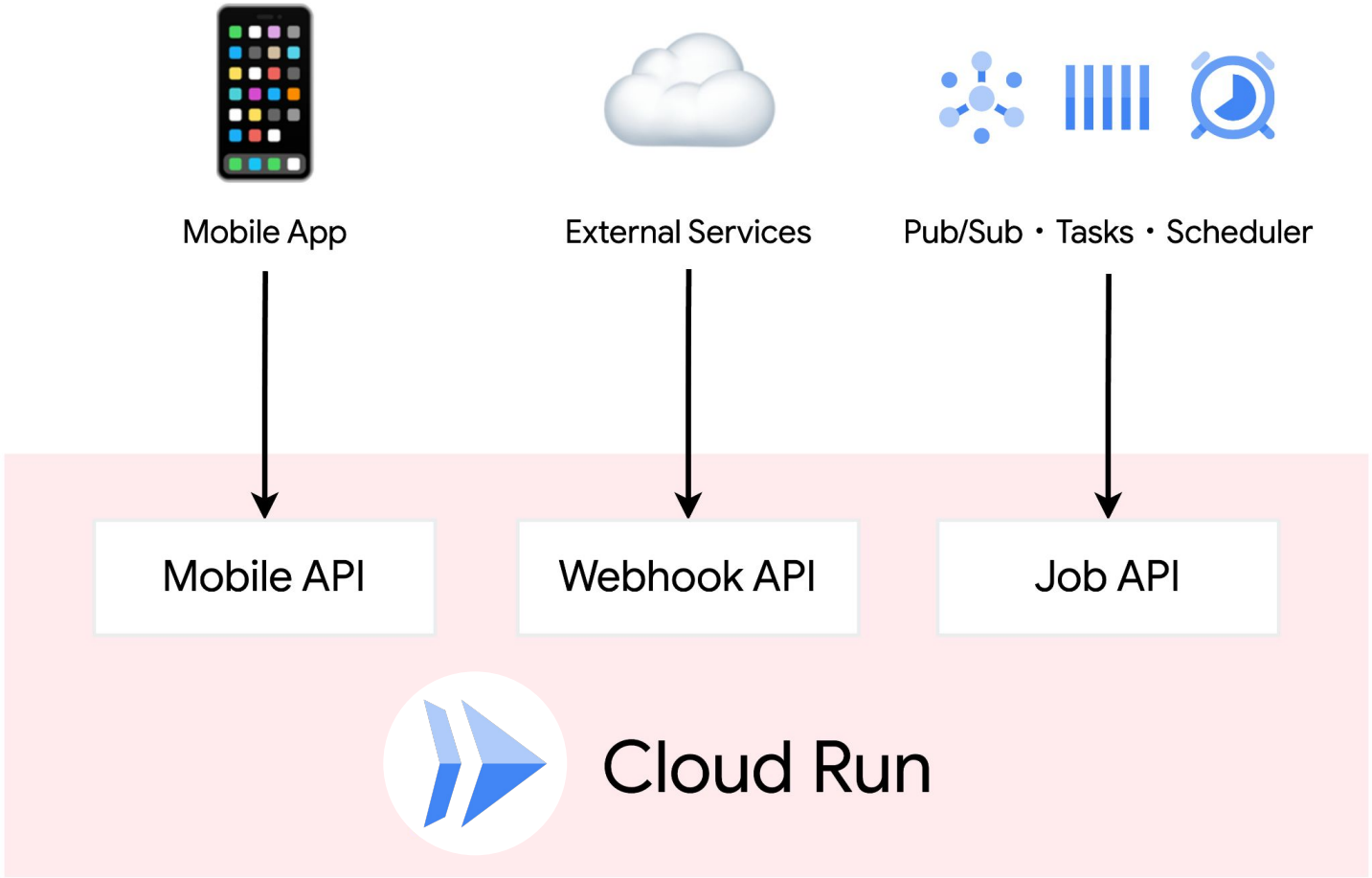
Cloud Run is a managed compute platform that enables you to run containers that are invocable via requests or events.

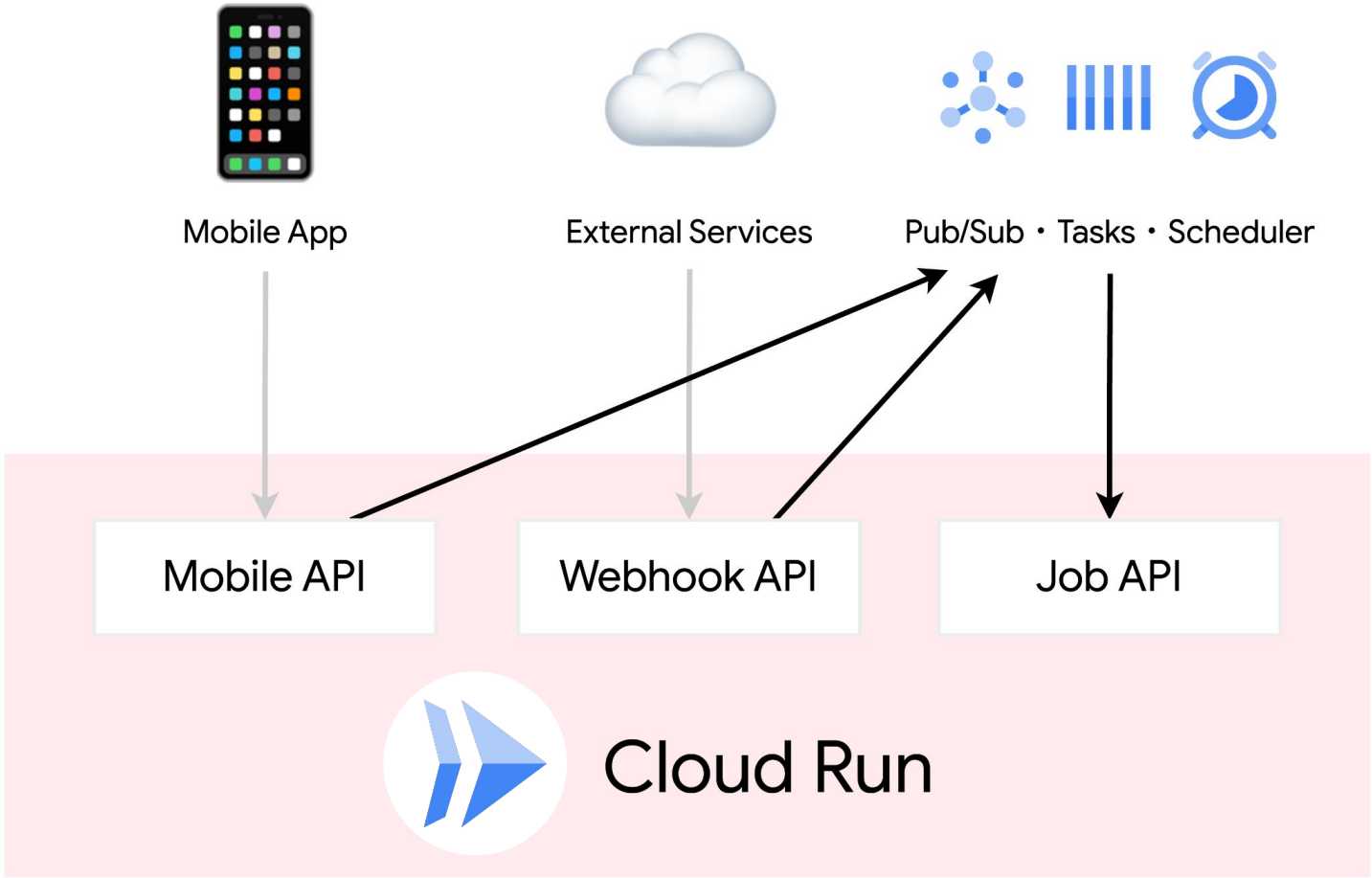
Cloud Run is serverless: it abstracts away all infrastructure management...

https://cloud.google.com/run/docs

Architecture - Key Principles -

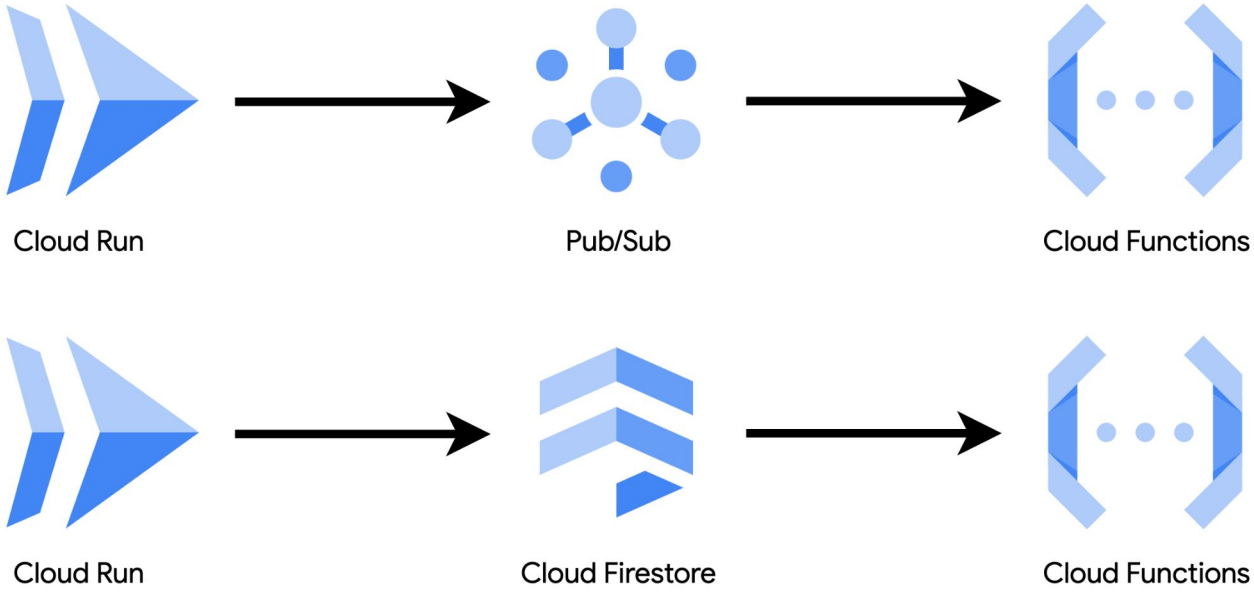
- Everything runs on Cloud Run
- Everything runs as an API

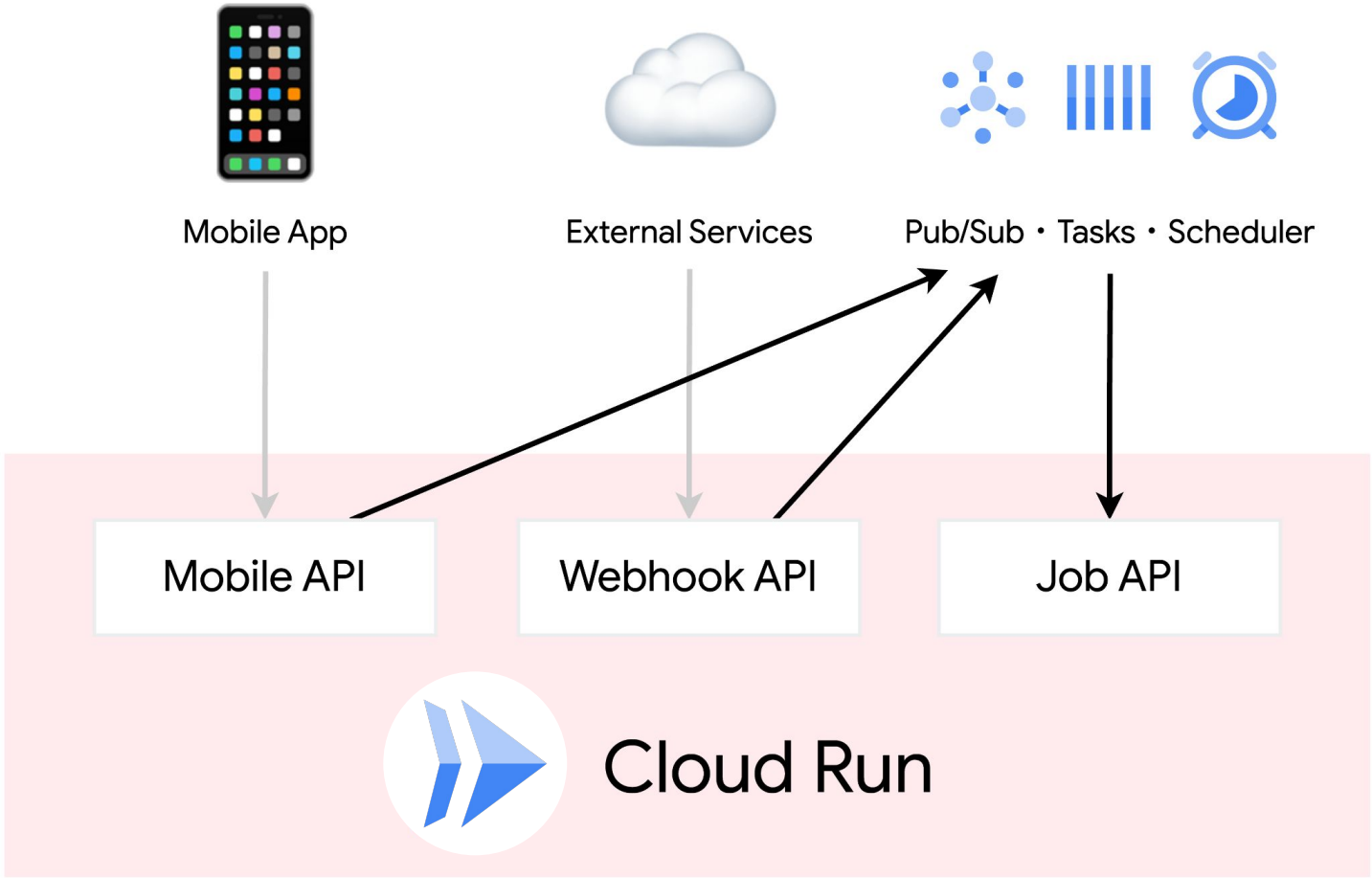

Mobile App
External Services
Pub/Sub • Tasks • Scheduler
Mobile API
Webhook API
Job API
Cloud Run


Mobile App
External Services
Pub/Sub • Tasks • Scheduler
Mobile API
Webhook API
Job API
Cloud Run

Architecture - Key Principles -

e.g.) VS. Cloud Functions Trigger




Mobile App
External Services
Pub/Sub • Tasks • Scheduler
Mobile API
Webhook API
Job API
Cloud Run

Architecture - Key Principles -

✅ Everything is Managed as API Definitions

✅ Reuse same implementation logic as APIs

✅ Use same Monitoring environments

Architecture - Key Principles -

✅ Everything is Managed as API Definitions

✅ Reuse same implementation logic as APIs

✅ Use same Monitoring environments

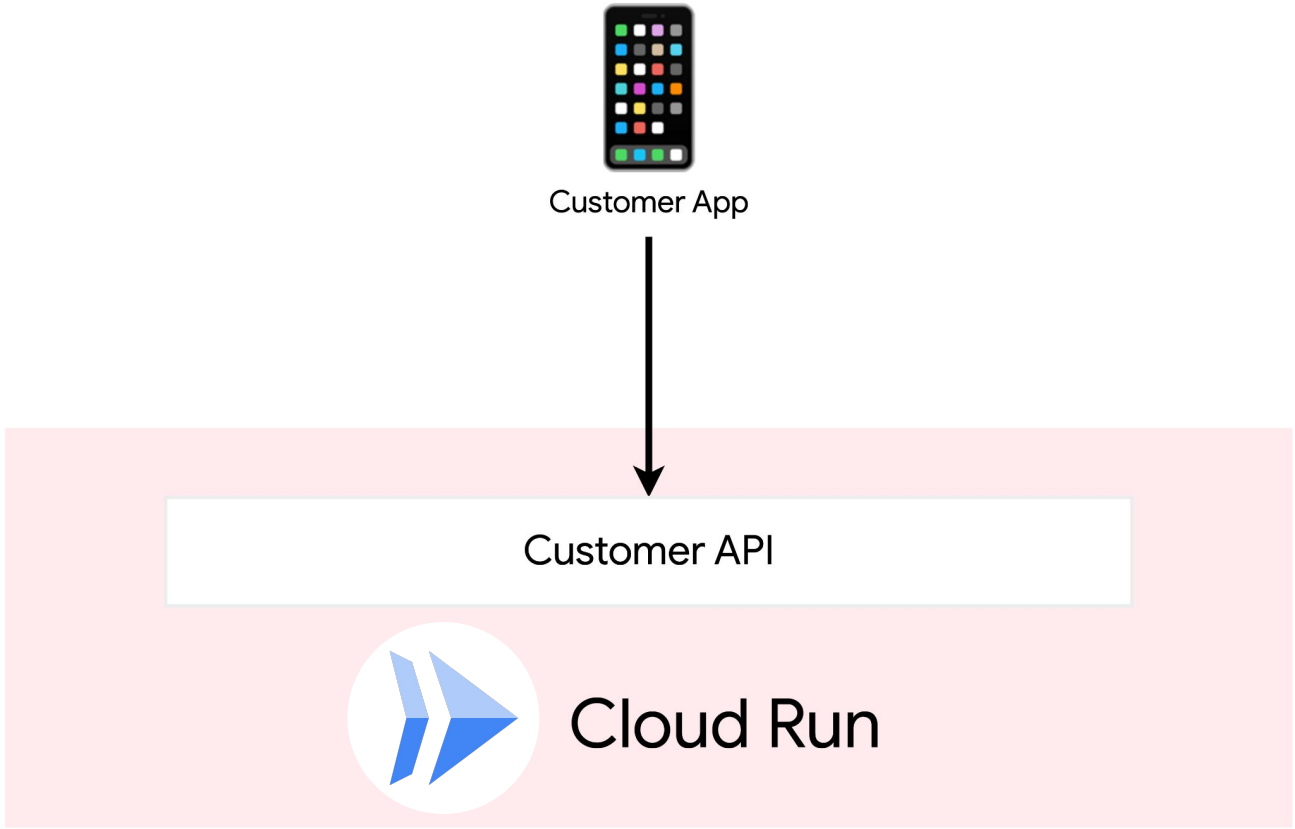
gRPC

gRPC is a modern open source high performance Remote Procedure Call (RPC) framework that can run in any environment.

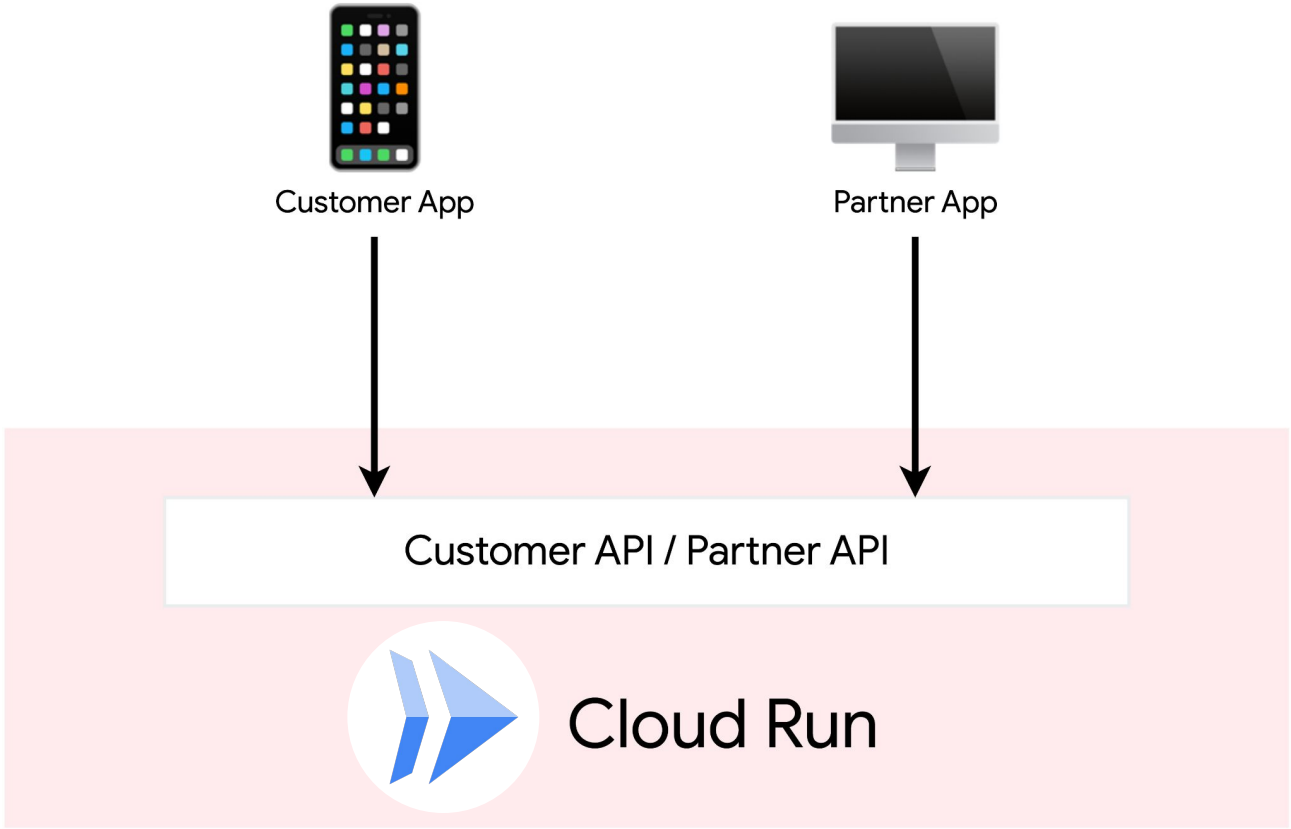
https://grpc.io/


gRPC

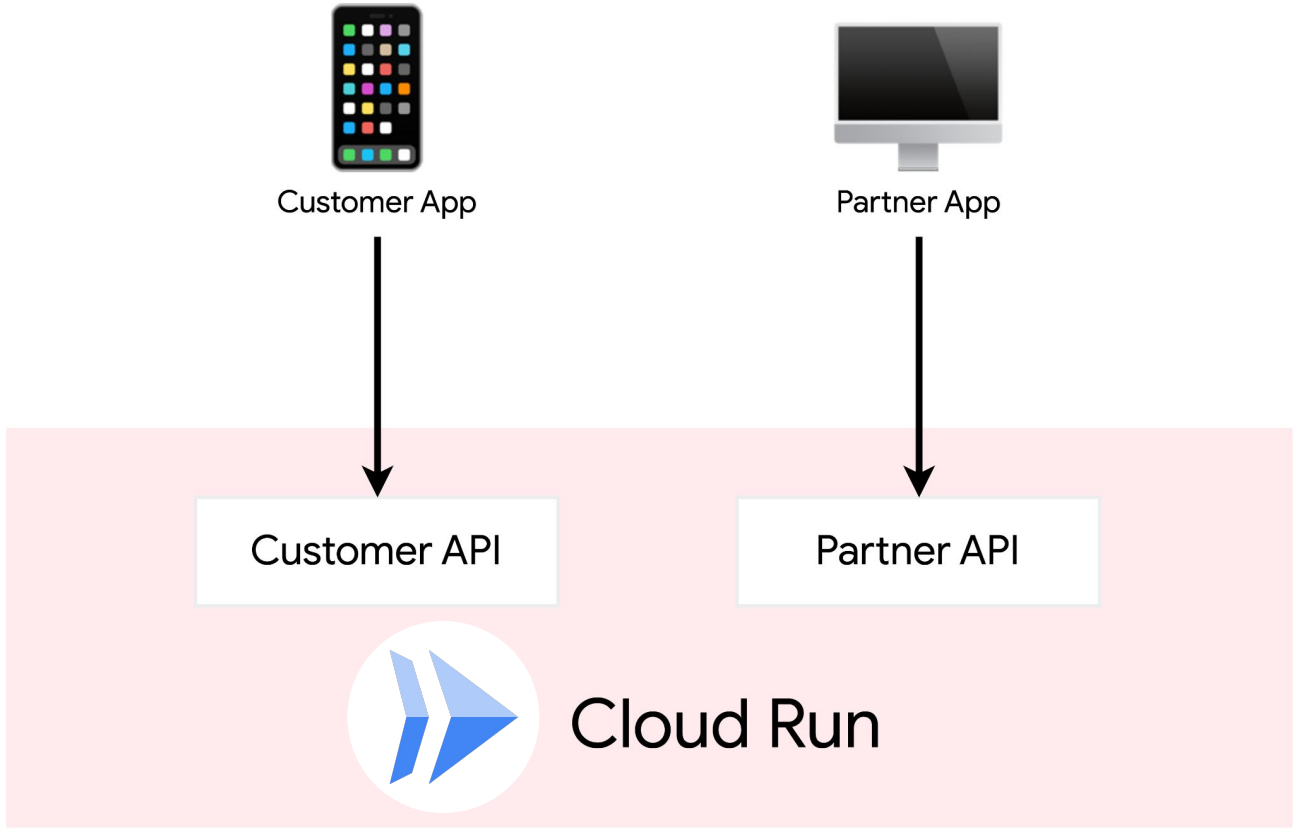
Architecture: 2020 ~



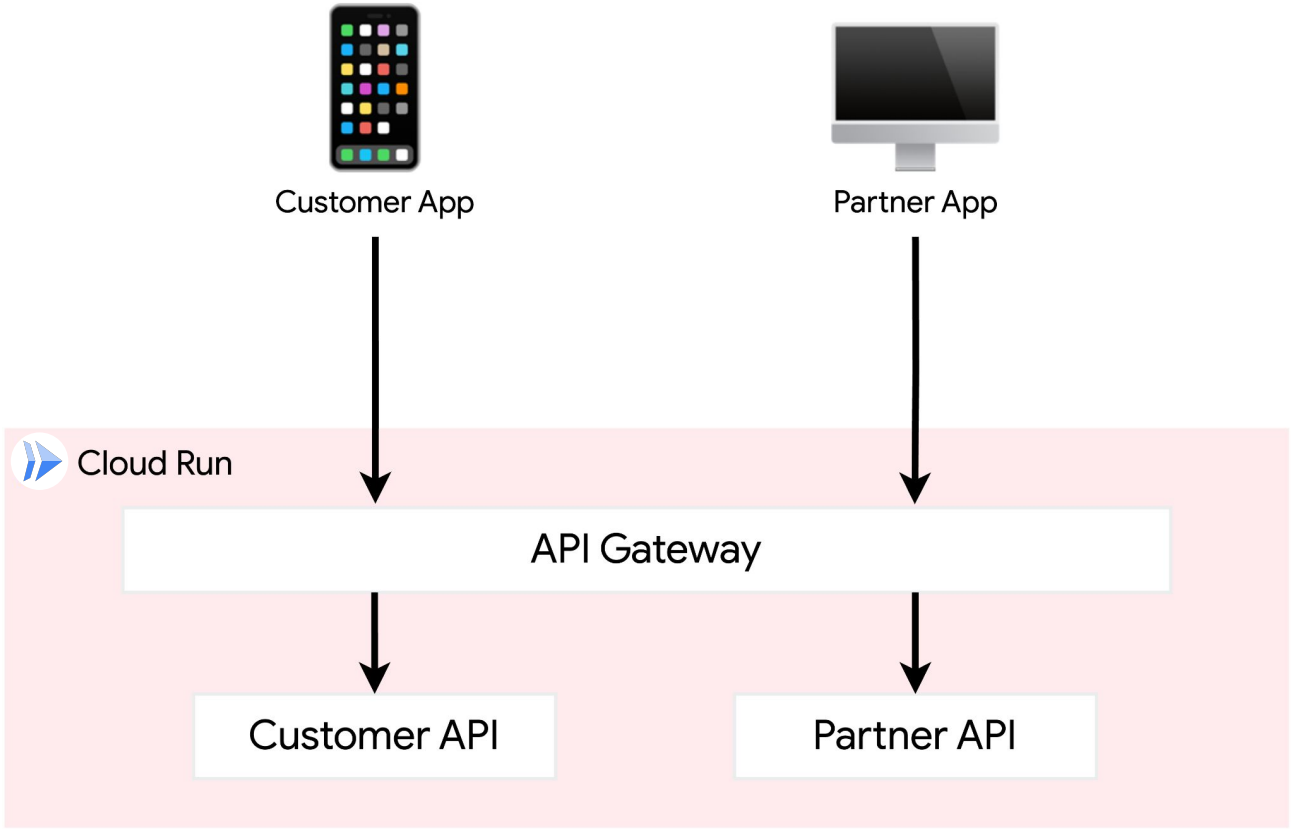
Architecture: 2021 ~



Architecture: 2022 ~



Architecture: 2022 ~



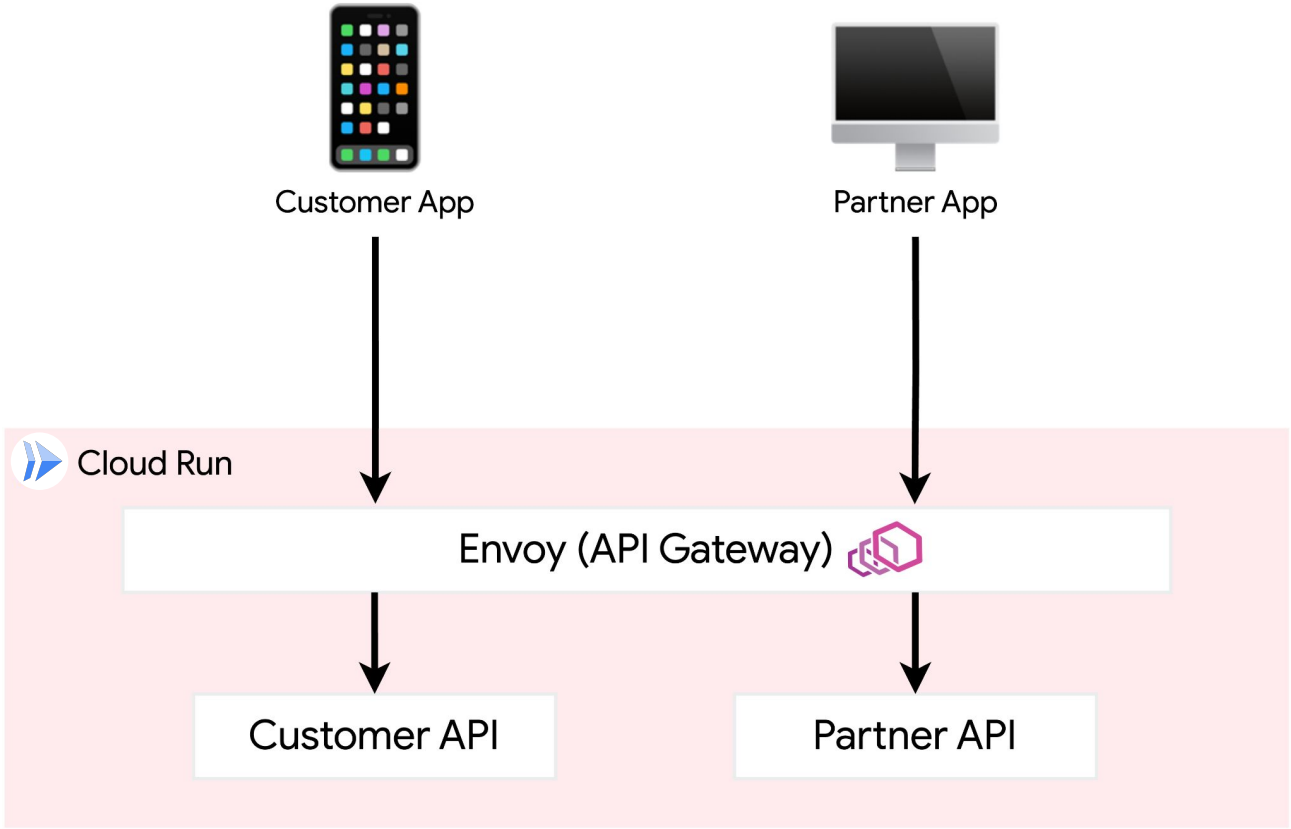
Offloading Cross-Cutting Concerns to the API Gateway

✅ Authentication / Authorization

✅ Transcoding

✅ TLS / Domain / CDN / IP ...

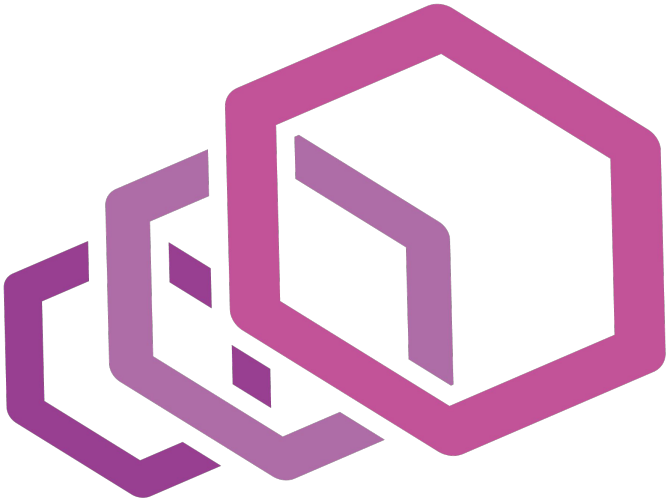
Architecture: 2022 ~



Envoy

Envoy is an L7 proxy and communication bus designed for large modern service oriented architectures. The project was born out of the belief that:

The network should be transparent to applications. When network and application problems do occur it should be easy to determine the source of the problem.



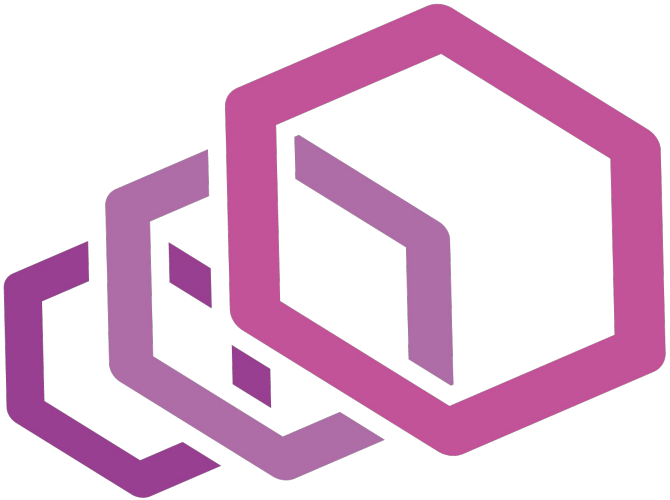
https://www.envoyproxy.io/docs/envoy/v1.26.2/intro/what_is_envoy

Envoy

✅ Extensibility with WebAssembly

✅ Dynamic Configuration

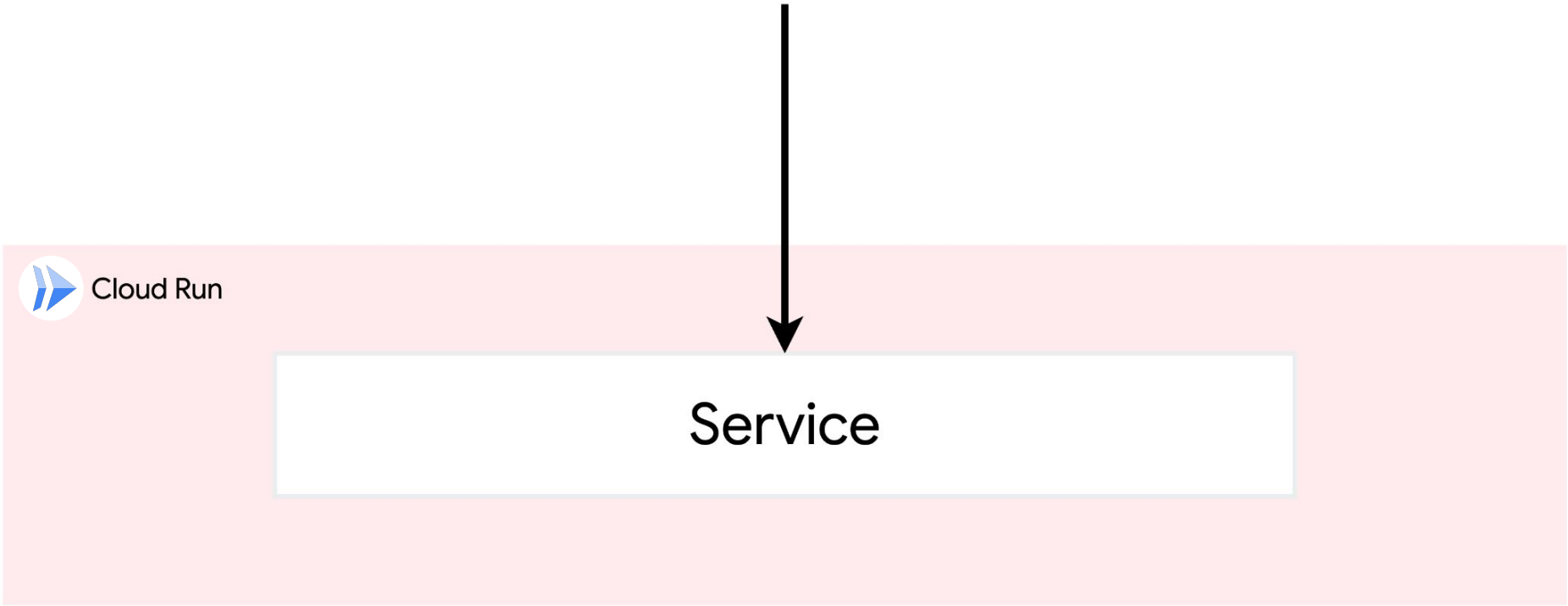
✅ Widely used in the Cloud Native World



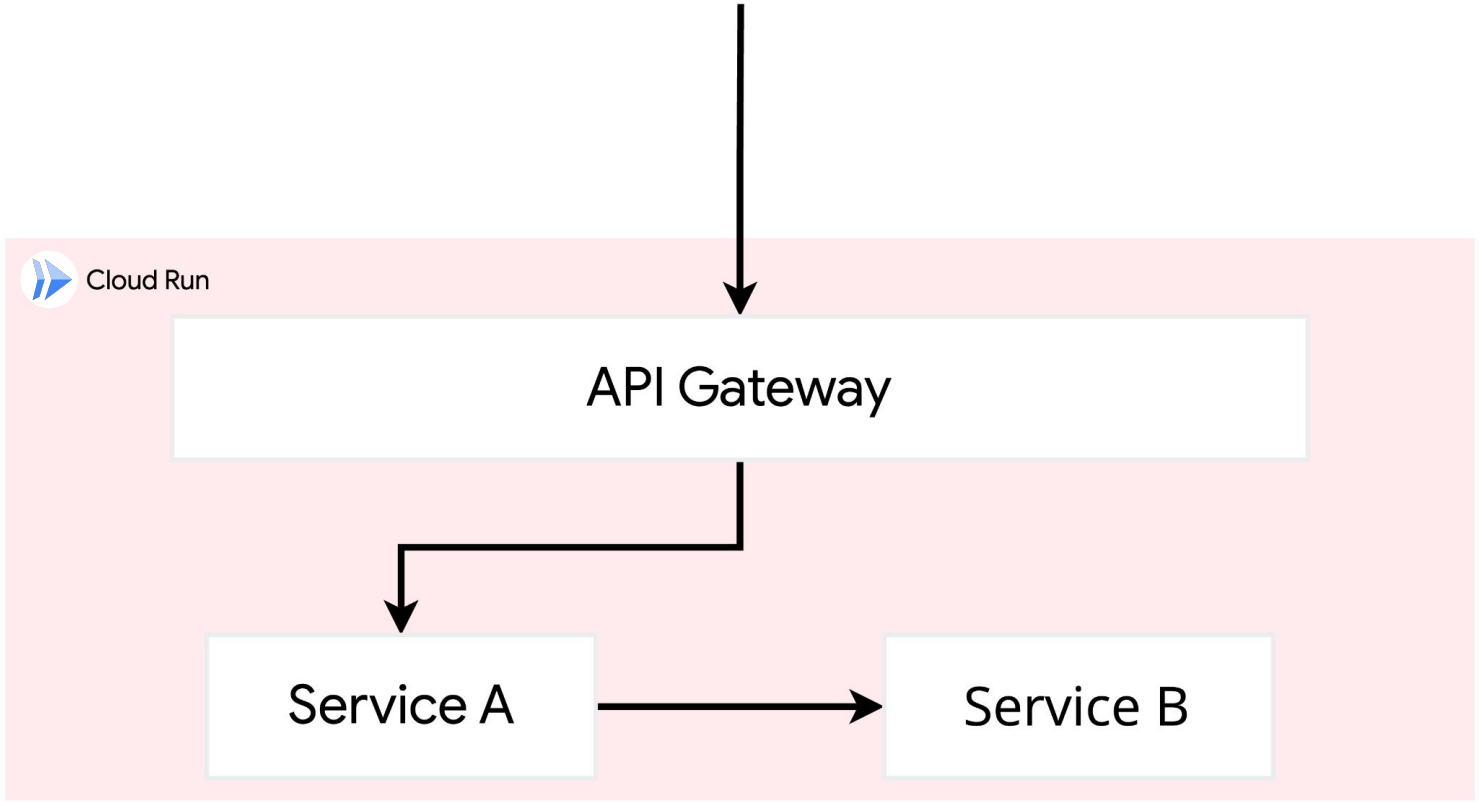
02

Access Control / Network

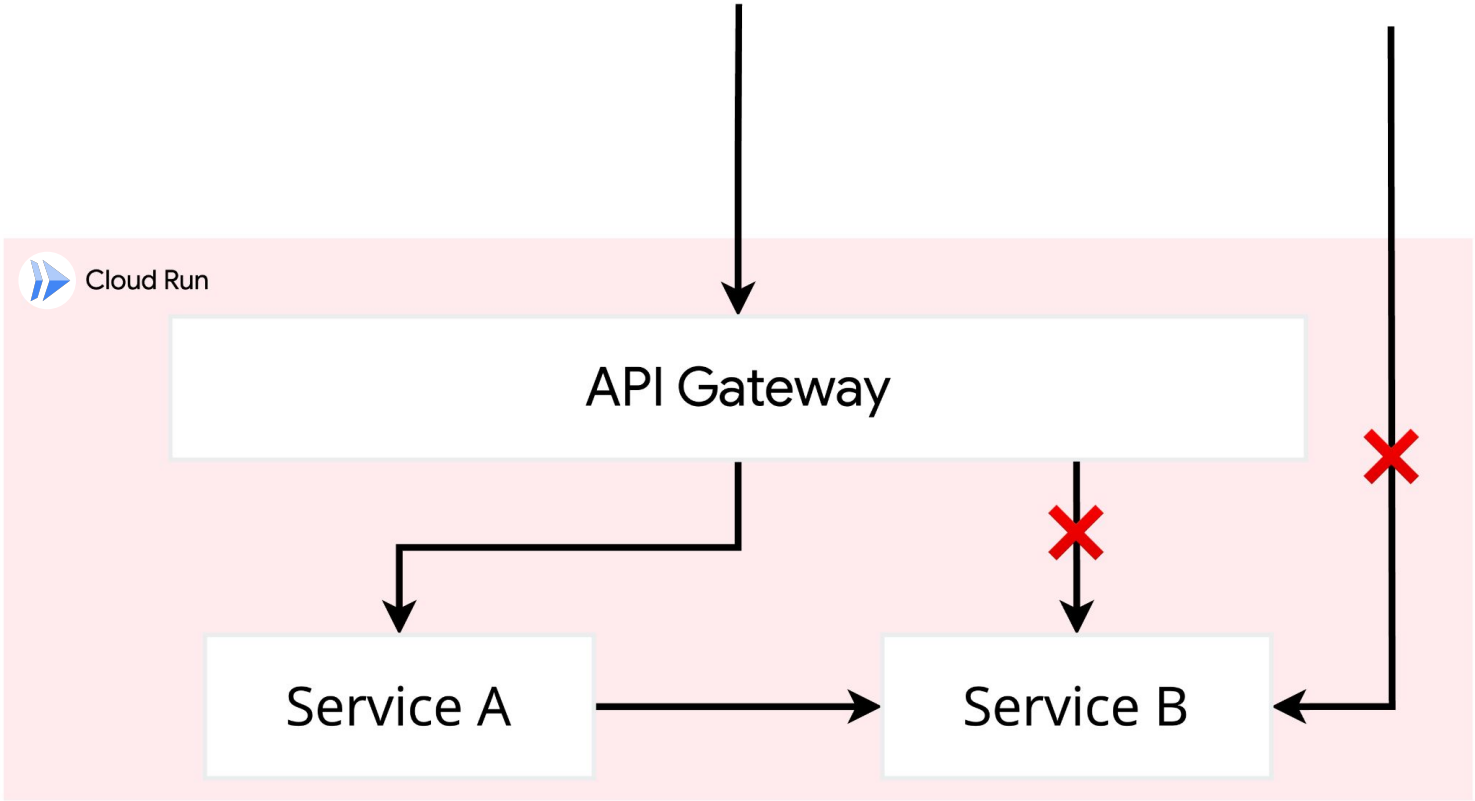
Single Service



Service-to-Service Communication



Access Control



Access Control

✅ Access Control with IAM

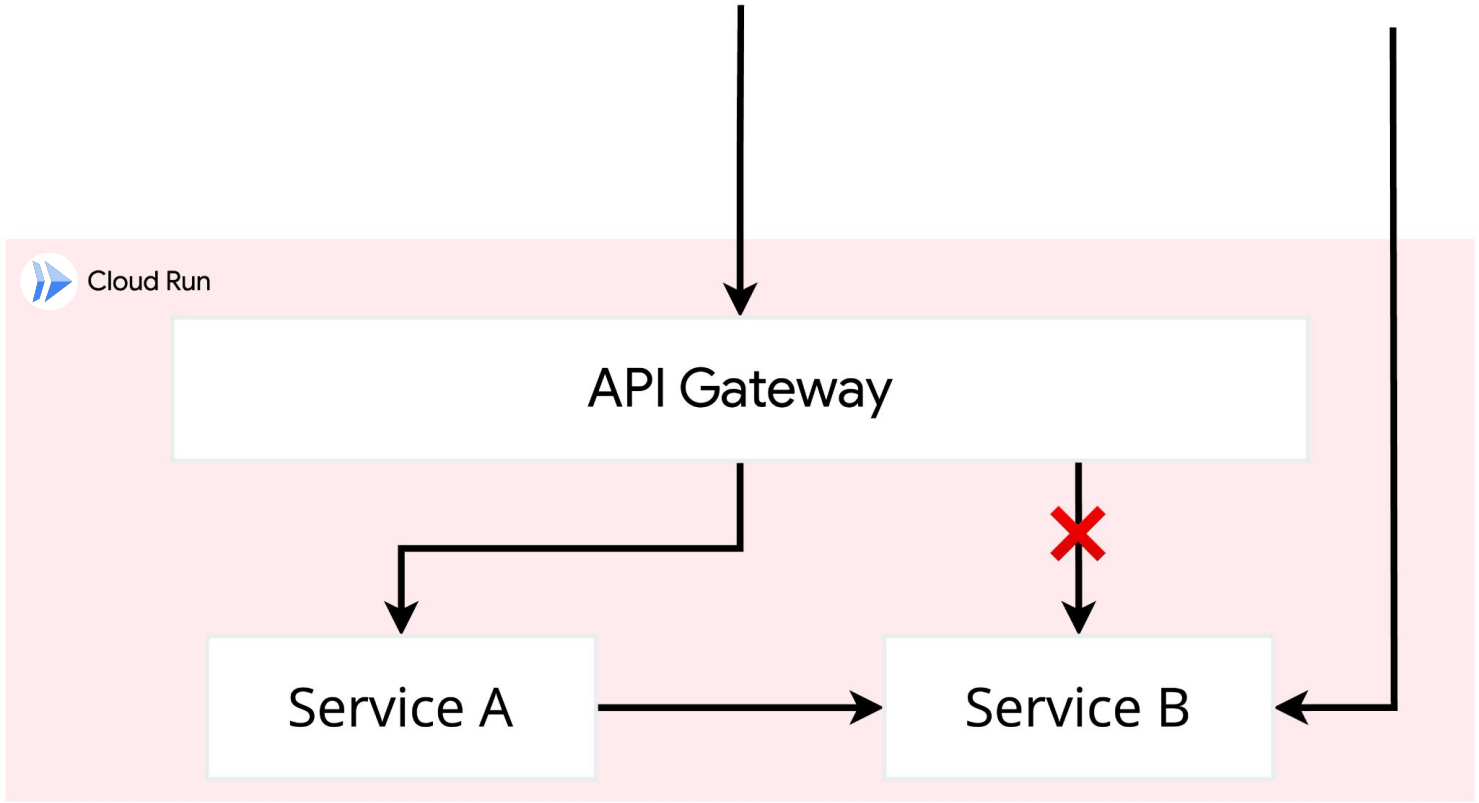
✅ Restricting Ingress

Access Control

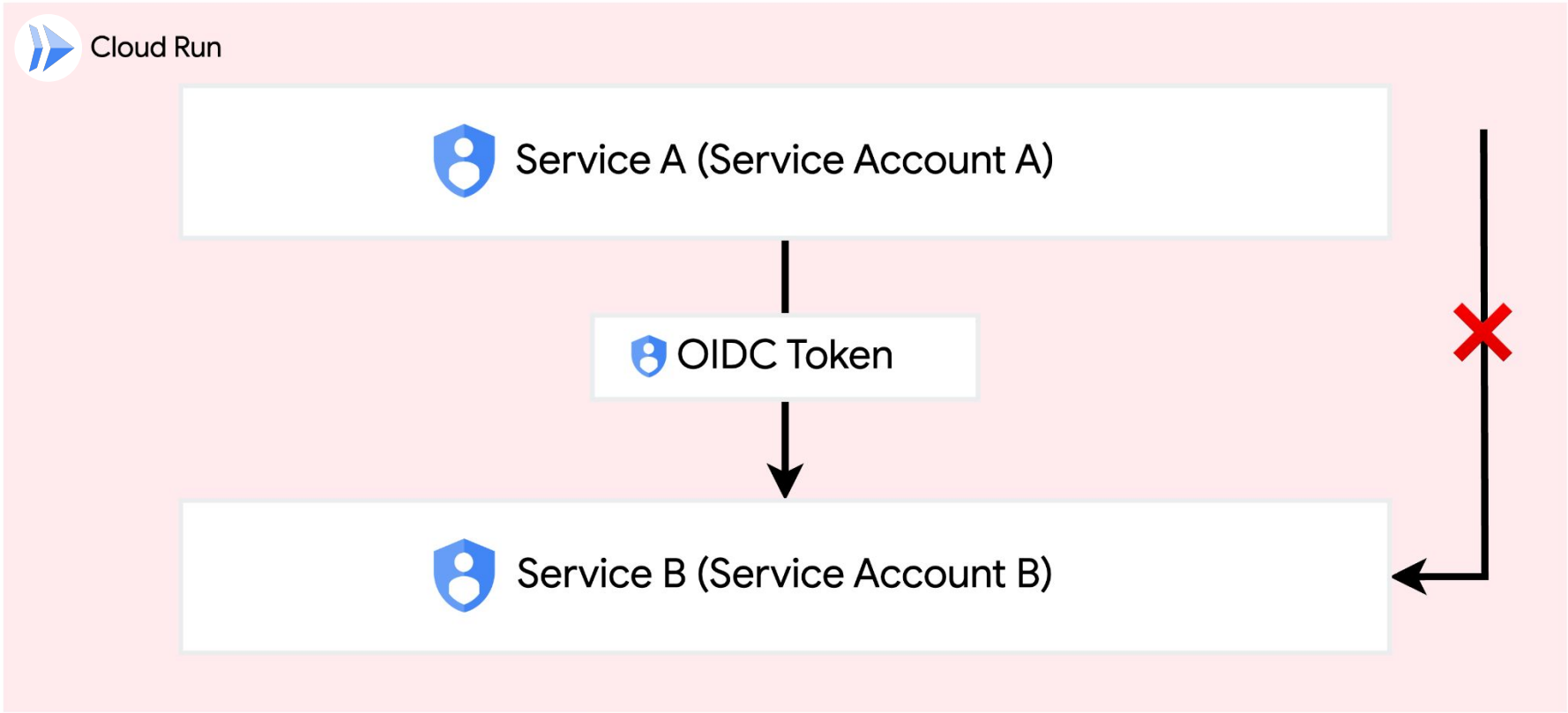
✅ Access Control with IAM

✅ Restricting Ingress

Access Control with IAM



Access Control with IAM

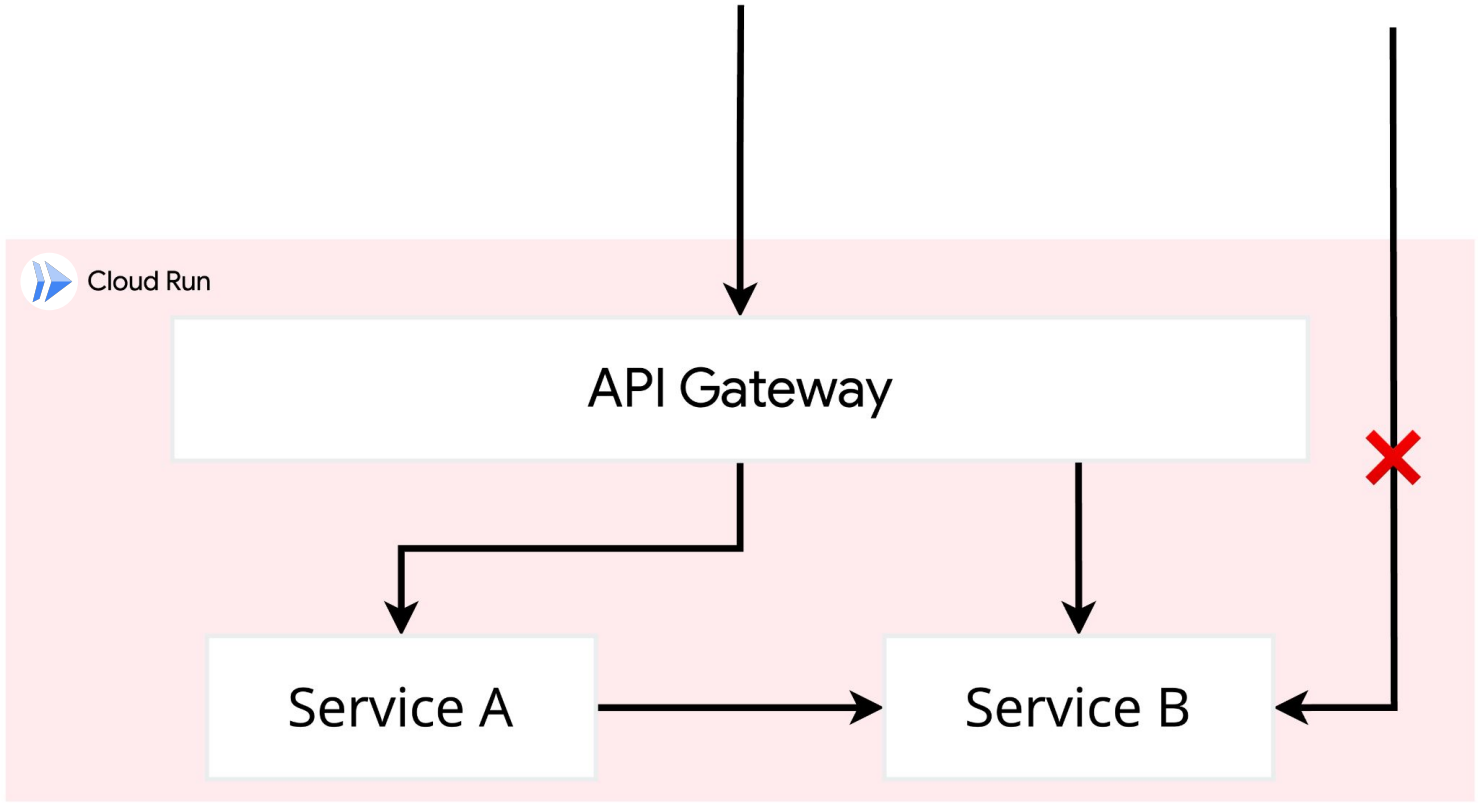


Access Control

✅ Access Control with IAM

✅ Restricting Ingress

Restricting Ingress



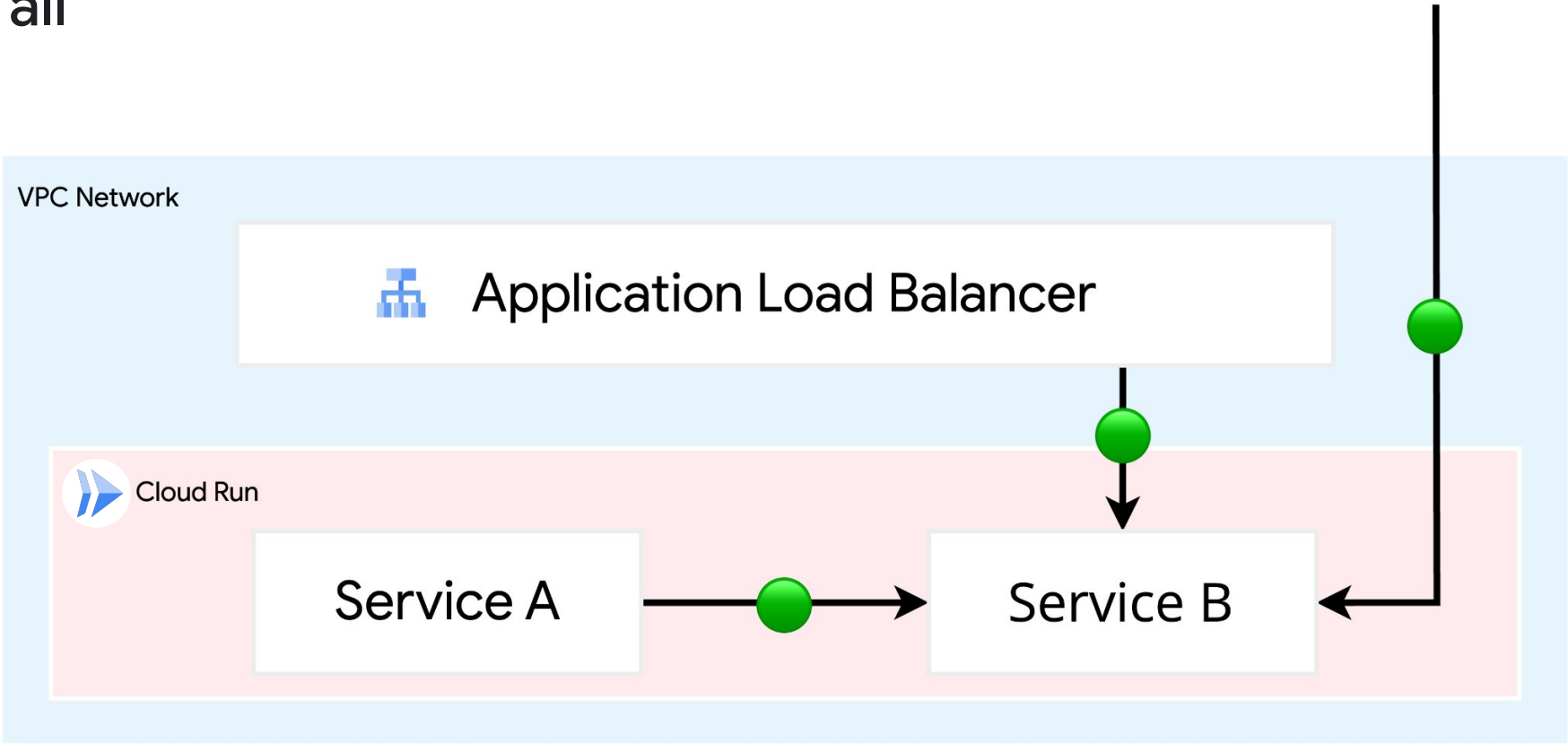
Ingress Settings

✅ all

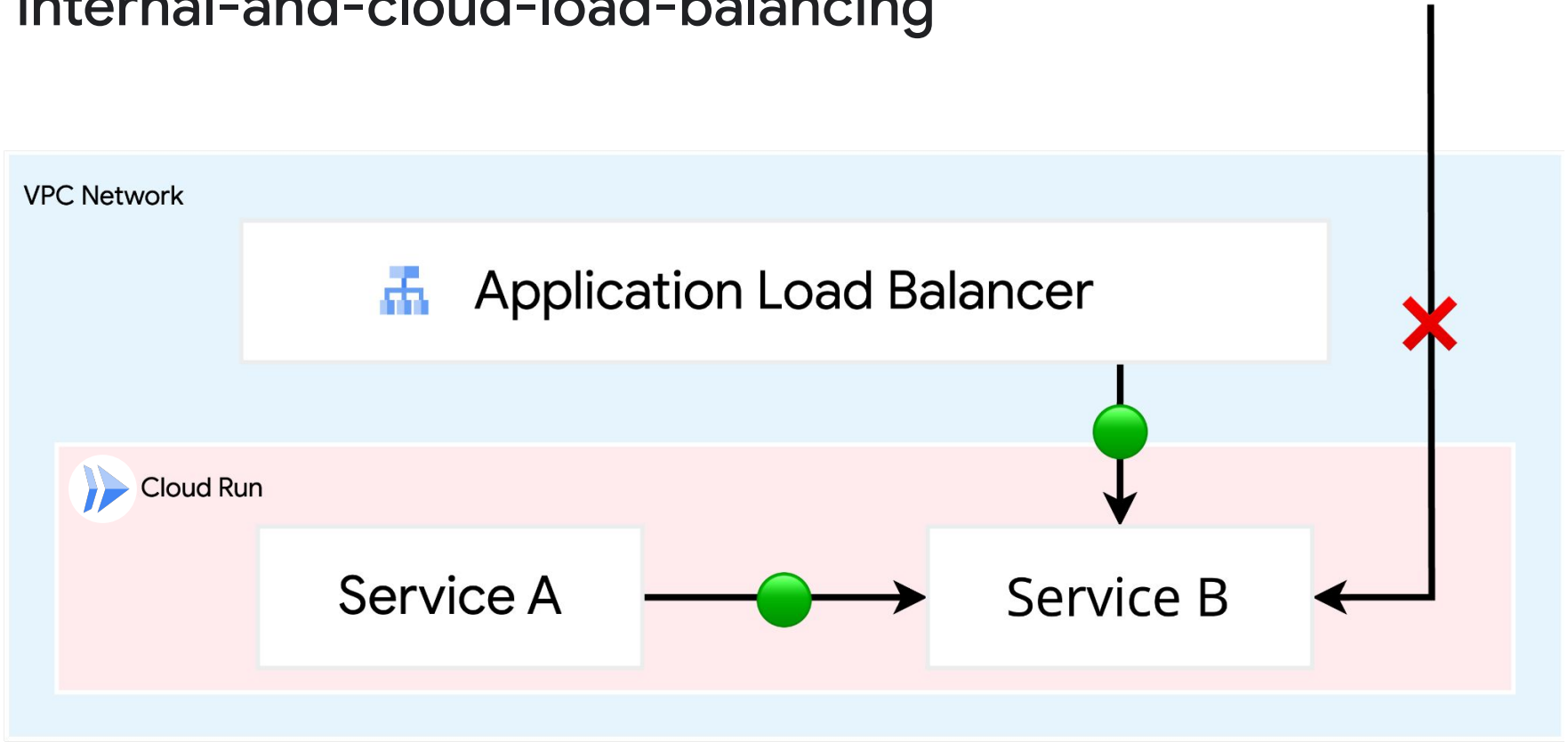
✅ internal-and-cloud-load-balancing

✅ internal

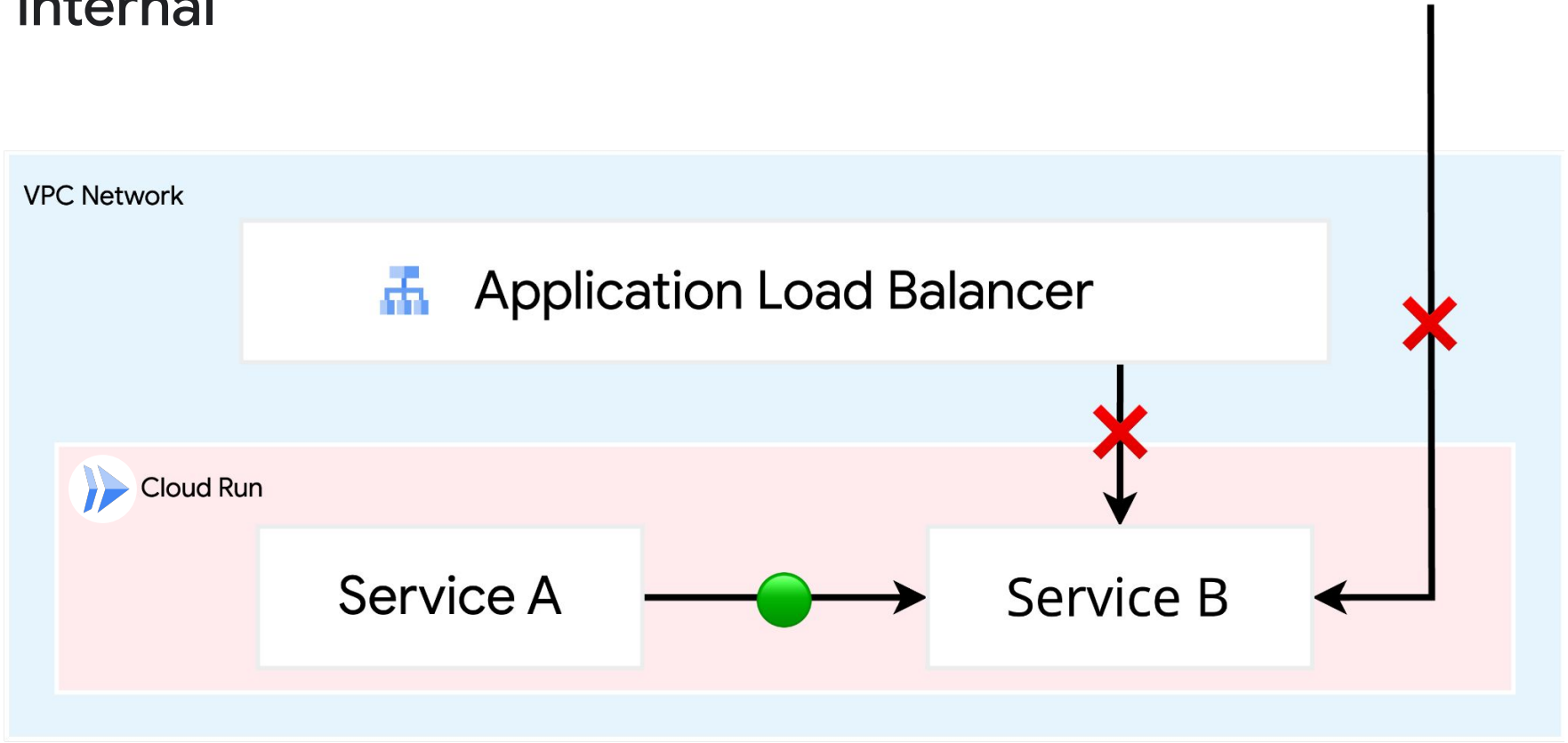
all


VPC Network
Application Load Balancer
Cloud Run
Service A
Service B

internal-and-cloud-load-balancing



internal


VPC Network
Application Load Balancer
Cloud Run
Service A
Service B

Ingress Settings

```yaml
---
apiVersion: serving.knative.dev/v1
kind: Service
metadata:
  annotations:
    run.googleapis.com/ingress: internal
spec:
  # ...
```

Ingress Settings

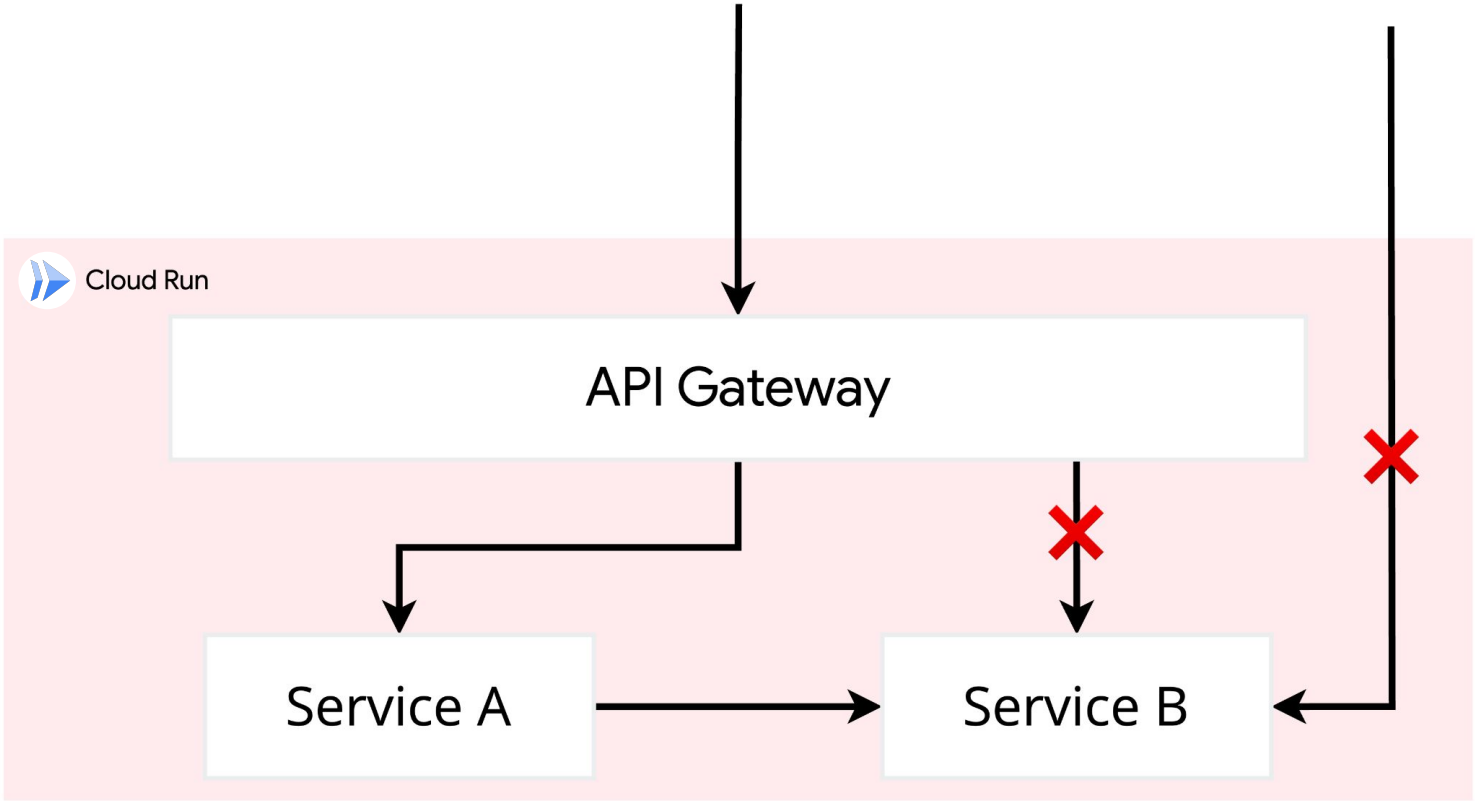
```
---
apiVersion: serving.knative.dev/v1
kind: Service
metadata:
  annotations:
    run.googleapis.com/ingress: internal
spec:
  # ...
```

Access Control

Ingress settings and IAM authentication methods are two ways of managing access to a service. They are independent of each other. For a layered approach to managing access, use both.

https://cloud.google.com/run/docs/securing/ingress

Access Control



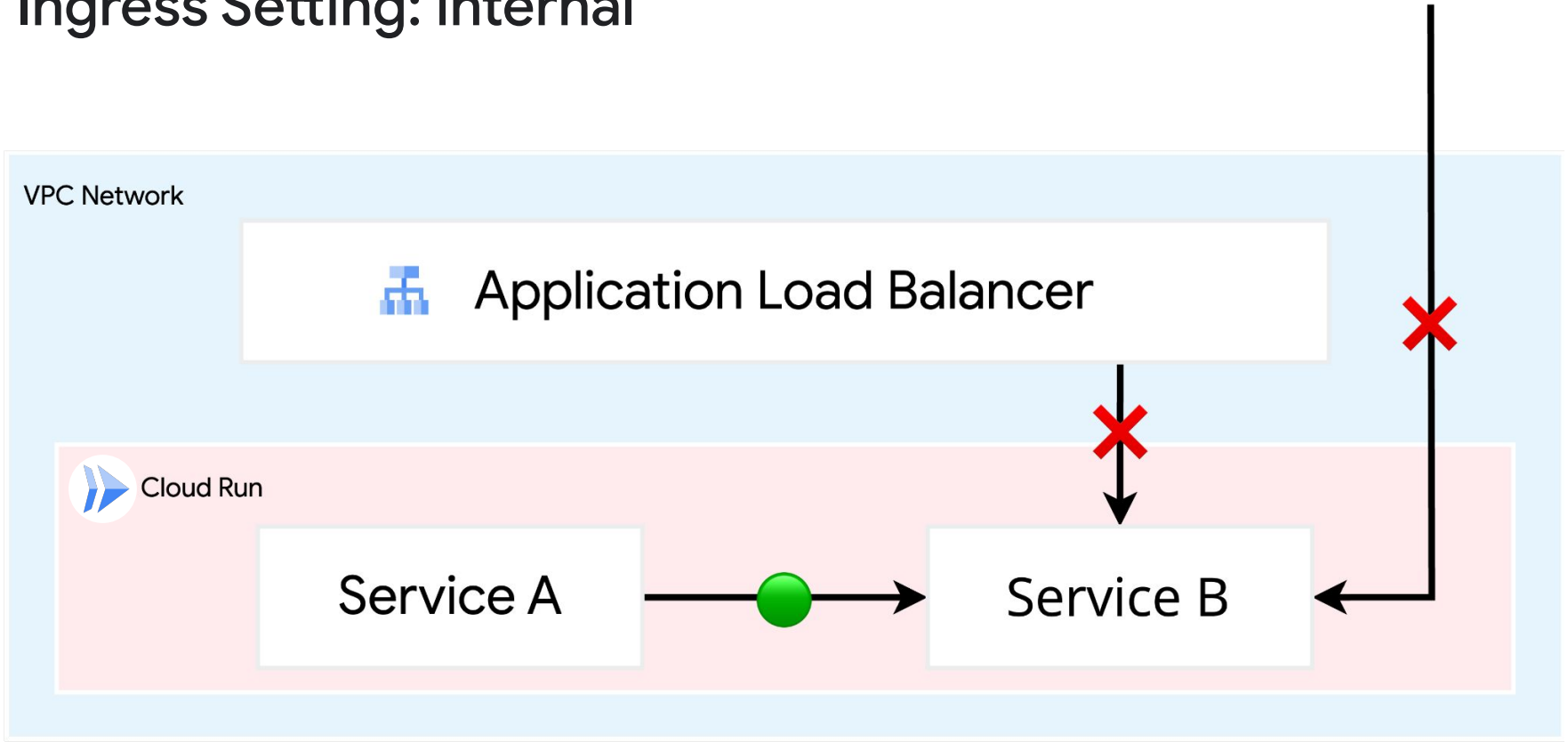
Network

- Serverless VPC Access Connector
- Shared VPC Network

Network

- Serverless VPC Access Connector
- Shared VPC Network

Ingress Setting: internal



What does “internal” mean …?

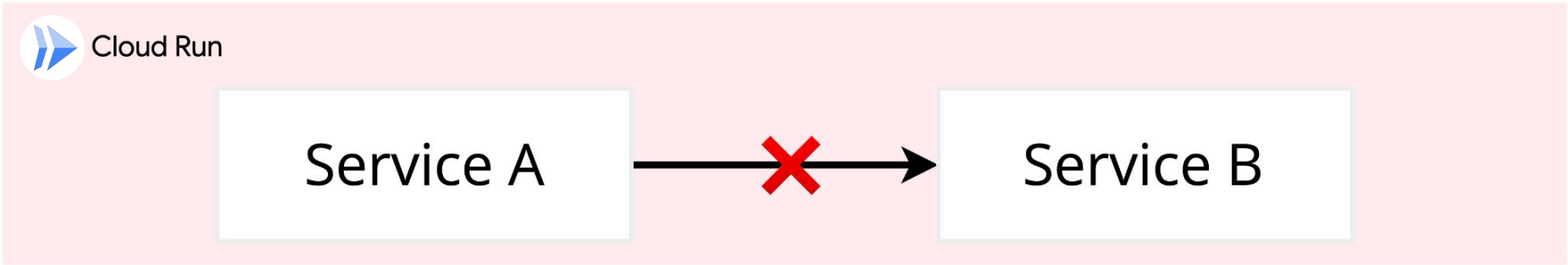
Setting	Description

Internal	Most restrictive. Allows requests from the following sources:
	• Internal Application Load Balancer, including requests from Shared VPC networks when routed through the internal Application Load Balancer
	• Resources allowed by any VPC Service Controls perimeter that contains your Cloud Run service
	• VPC networks that are in the same project or VPC Service Controls perimeter as your Cloud Run service
	• Shared VPC ingress (Preview): The Shared VPC network that your revision is configured to send traffic to. For information about when Shared VPC traffic is recognized as "internal", see Special considerations for Shared VPC.
	• The following Google Cloud products, if they are in the same project or VPC Service Controls perimeter as your Cloud Run service:
	  • Cloud Scheduler (Preview)
	  • Cloud Tasks
	  • Eventarc
	  • Pub/Sub
	  • Workflows
	  • BigQuery
	Requests from these sources stay within the Google network, even if they access your service at the `run.app` URL. Requests from other sources, including the internet, cannot reach your service at the `run.app` URL or custom domains.
	There is no support for multi-tenancy, that is, multiple trust domains within the same project.

https://cloud.google.com/run/docs/securing/ingress#settings

Serverless VPC Access Connector

By default, requests from other Cloud Run Services are not treated as "internal"

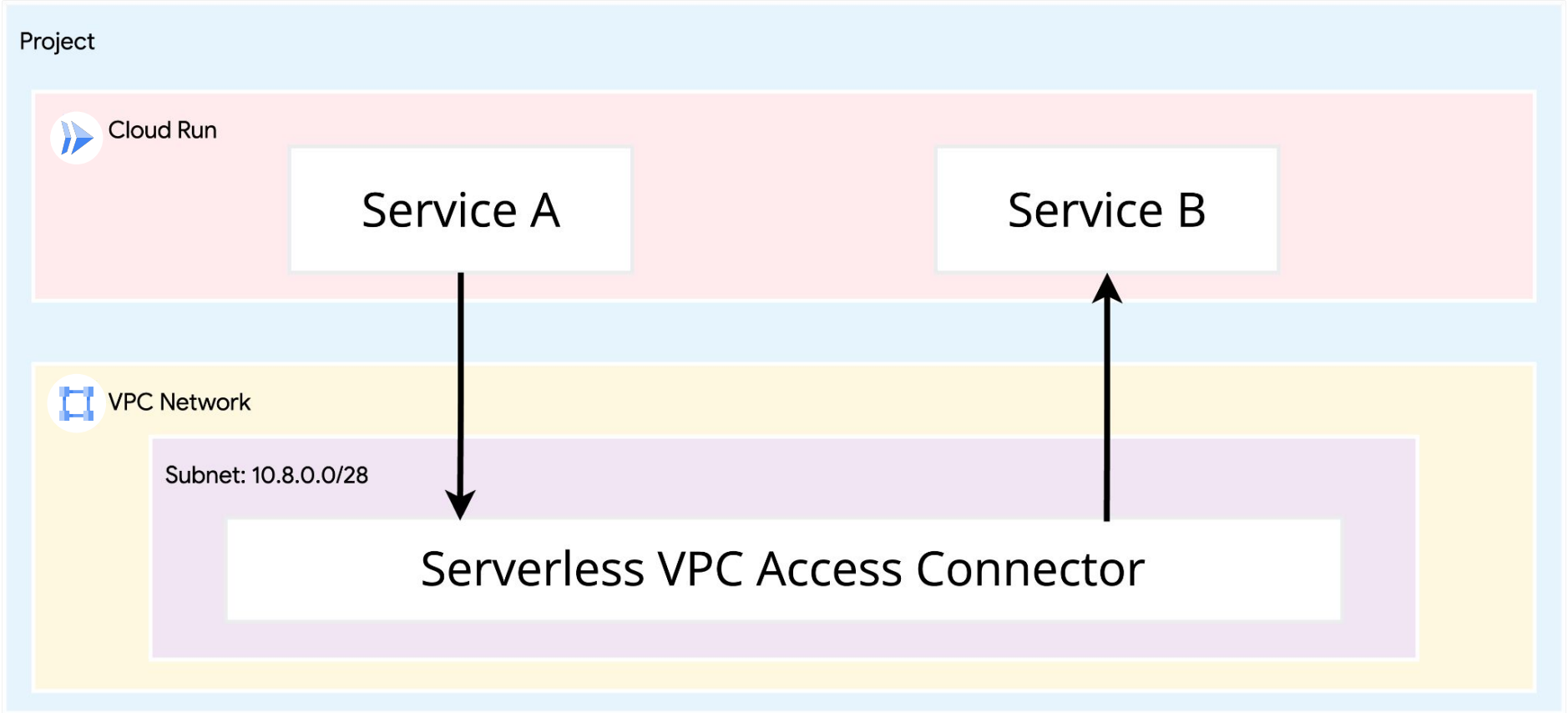


Serverless VPC Access Connector

For requests from other Cloud Run services in the same project, connect the service ... to a VPC network and route all egress through the connector

https://cloud.google.com/run/docs/securing/ingress#internal-services

Serverless VPC Access Connector



Serverless VPC Access Connector

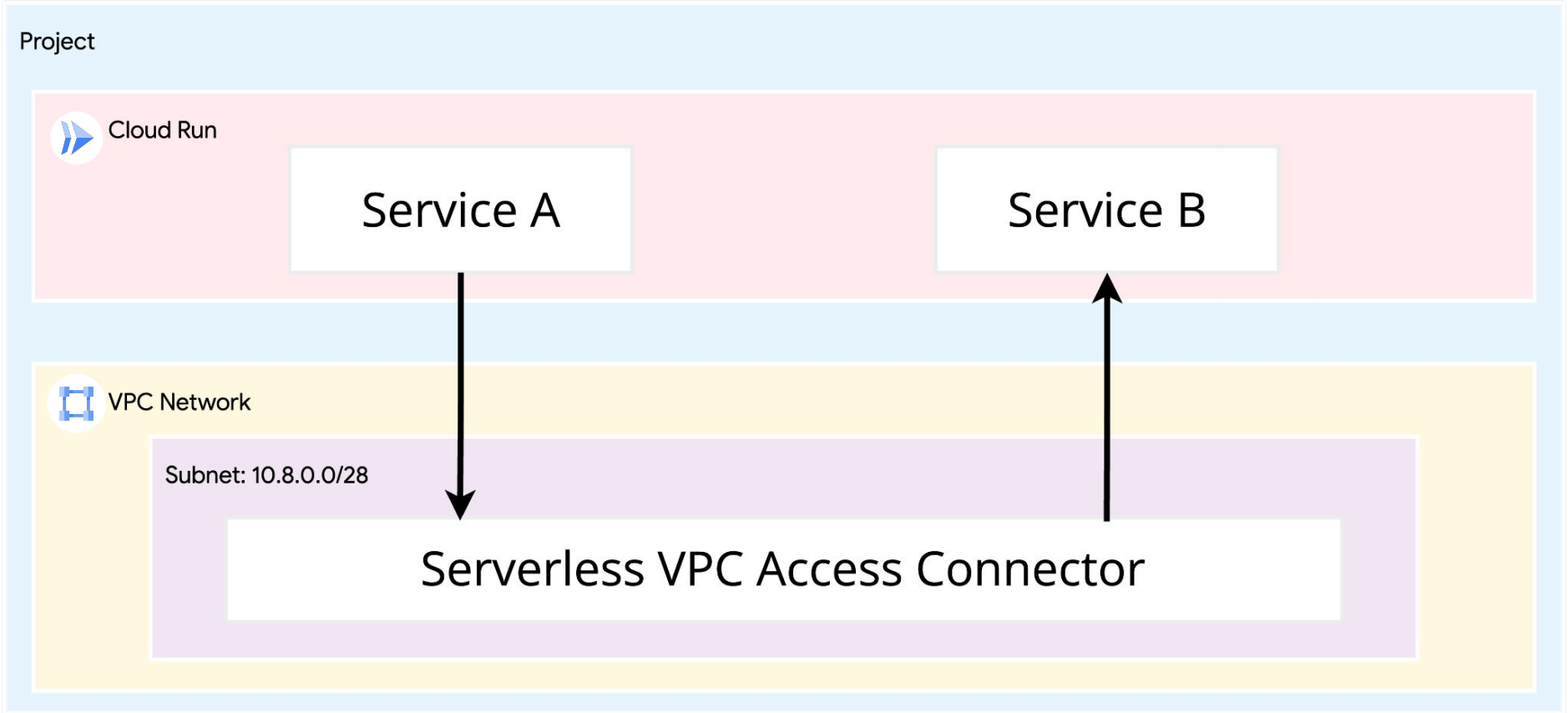
Serverless VPC Access makes it possible for you to connect directly to your Virtual Private Cloud (VPC) network from serverless environments such as Cloud Run ...

https://cloud.google.com/vpc/docs/serverless-vpc-access

Serverless VPC Access Connector

```
---
apiVersion: serving.knative.dev/v1
kind: Service
metadata:
  name: service-a
spec:
  template:
    metadata:
      annotations:
        run.googleapis.com/vpc-access-egress: all-traffic
        run.googleapis.com/vpc-access-connector: projects/<project>/locations/<location>/connectors/<connector name>
```

Serverless VPC Access Connector



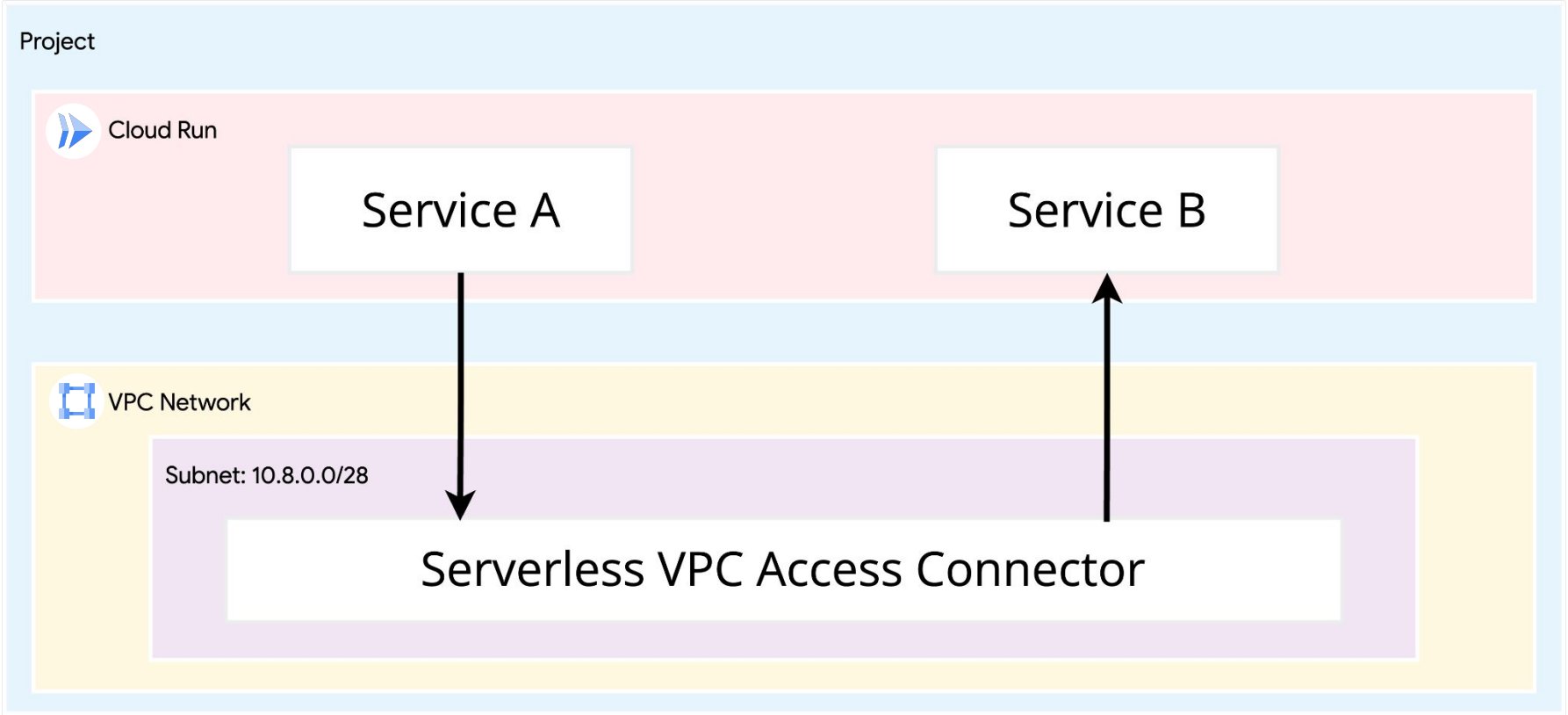
Network

- Serverless VPC Access Connector
- Shared VPC Network

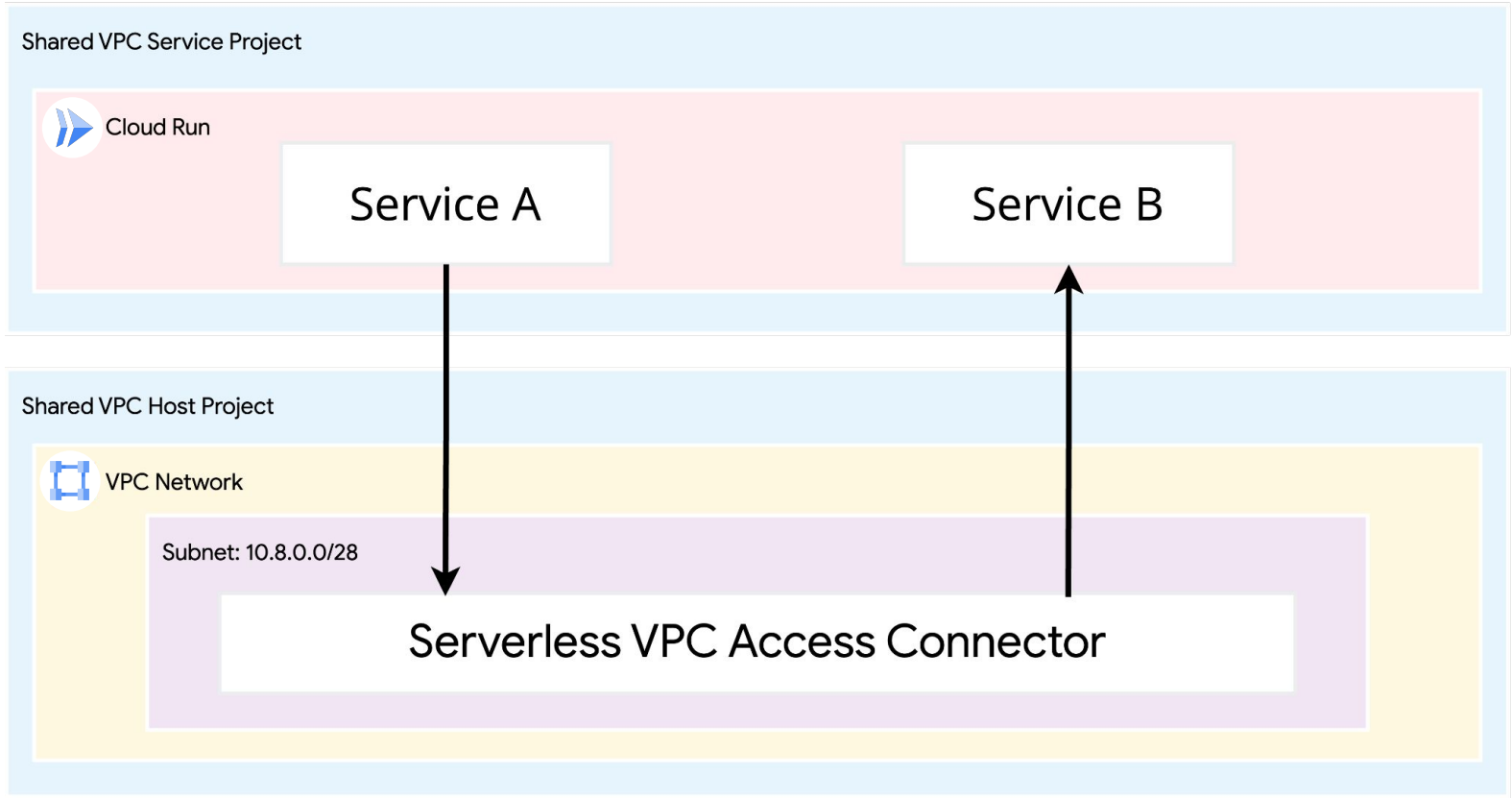
Network

- Serverless VPC Access Connector
- Shared VPC Network

Single VPC Network



Shared VPC Network



Shared VPC Network

Shared VPC allows an organization to connect resources from multiple projects to a common Virtual Private Cloud (VPC) network, so that they can communicate with each other securely and efficiently using internal IPs from that network

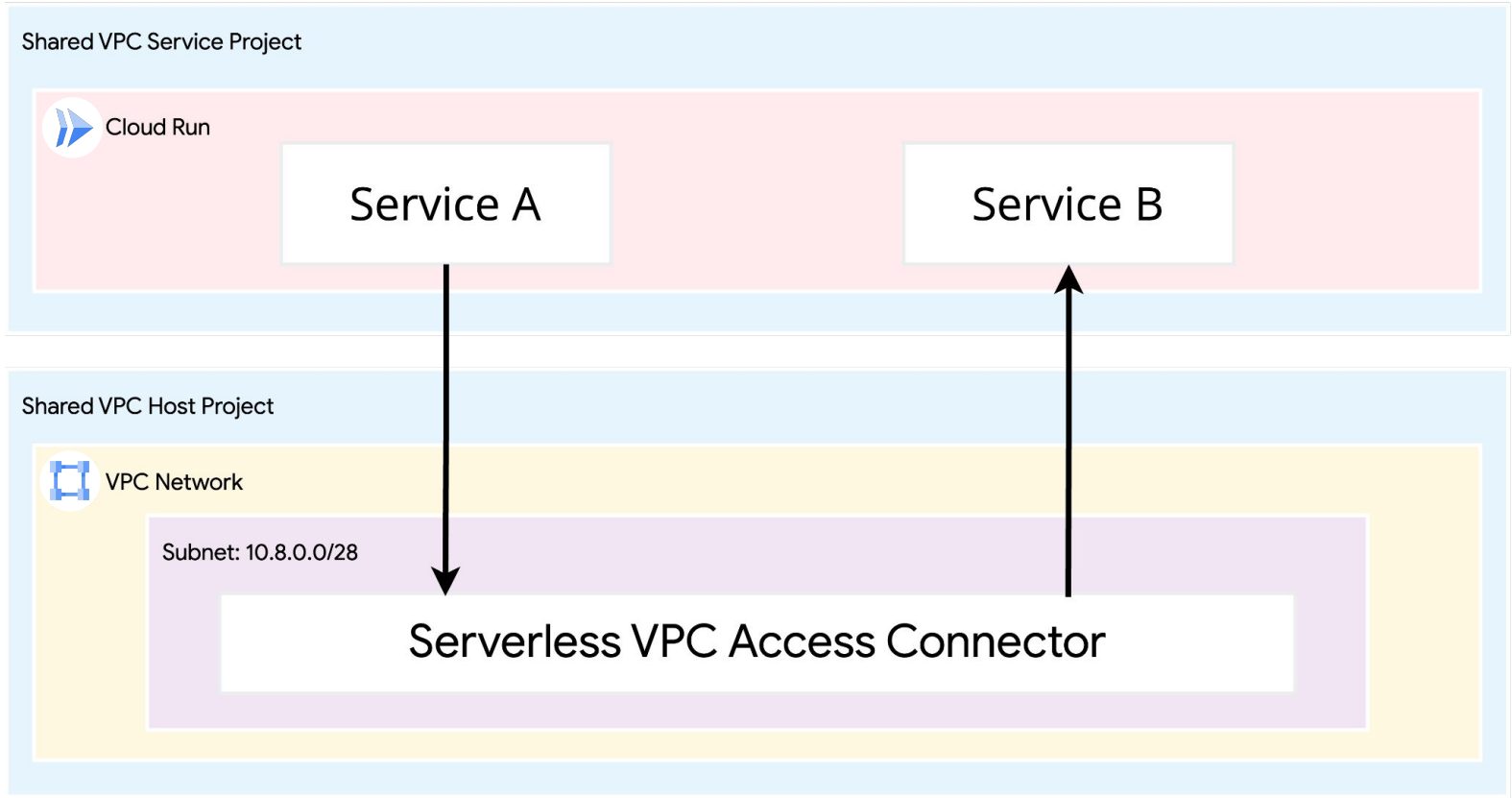
https://cloud.google.com/vpc/docs/shared-vpc

Shared VPC Network

✅ Delegating network responsibilities to administrators

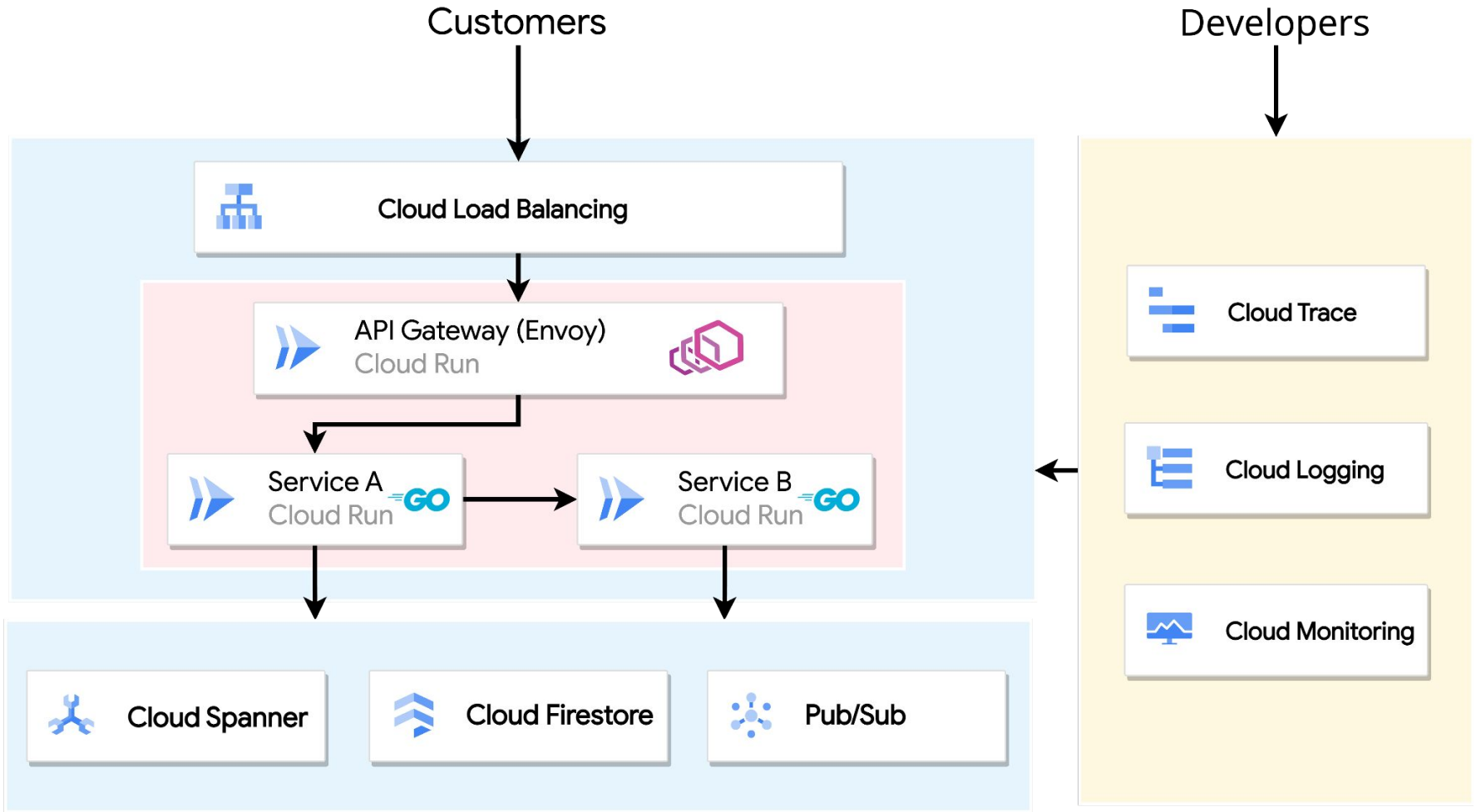
✅ Centralized control over network resources

Shared VPC Network



03

Reliability / Observability

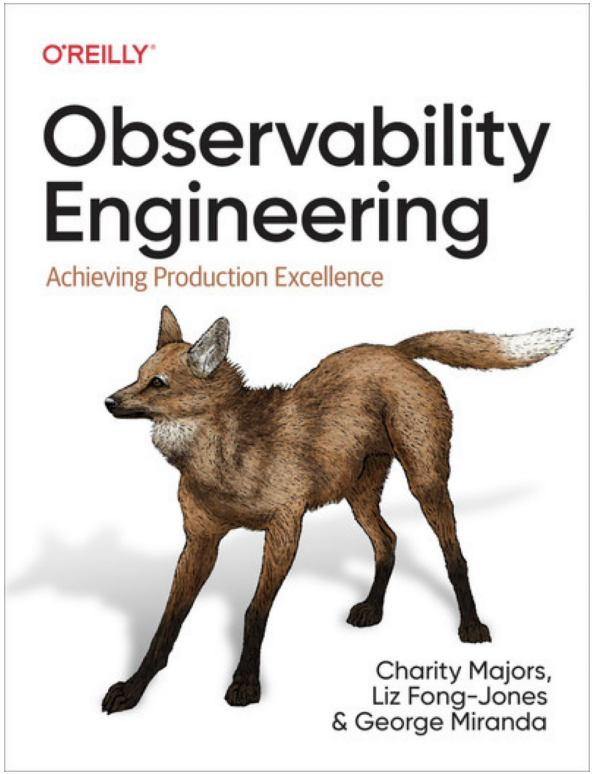

Customers
Developers
Cloud Load Balancing
API Gateway (Envoy)
Cloud Run
Service A
Cloud Run
GO
Service B
Cloud Run
GO
Cloud Trace
Cloud Logging
Cloud Monitoring
Cloud Spanner
Cloud Firestore
Pub/Sub

Observability

our definition of “observability” for software systems is a measure of how well you can understand and explain any state your system can get into, no matter how novel or bizarre.

...

If you can understand any bizarre or novel state without needing to ship new code, you have observability.



https://www.oreilly.com/library/view/observability-engineering/9781492076438/

Reliability / Observability

✅ Enable new members

to understand system states on Day 1.

Approaches

- Logs
- Traces
- Metrics
- SLI / SLO

Approaches

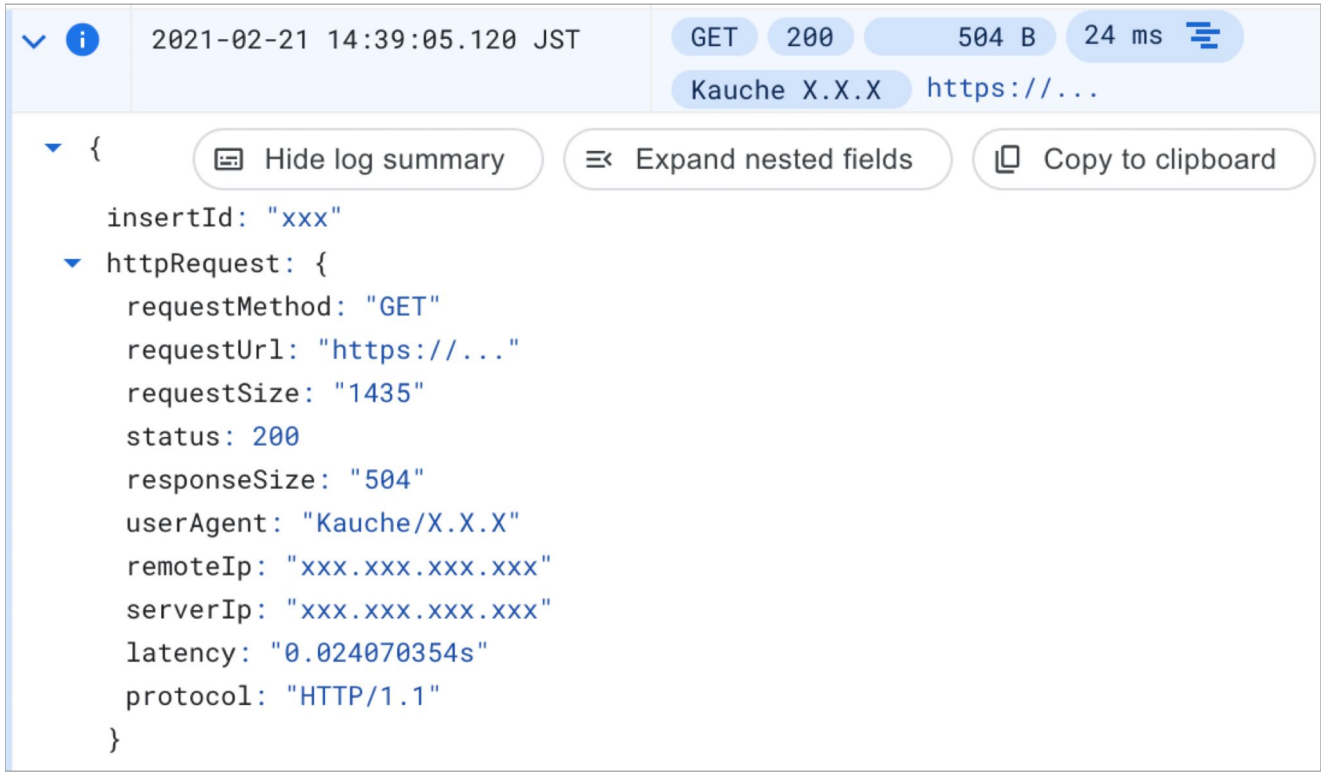
- Logs
- Traces
- Metrics
- SLI / SLO

Logs

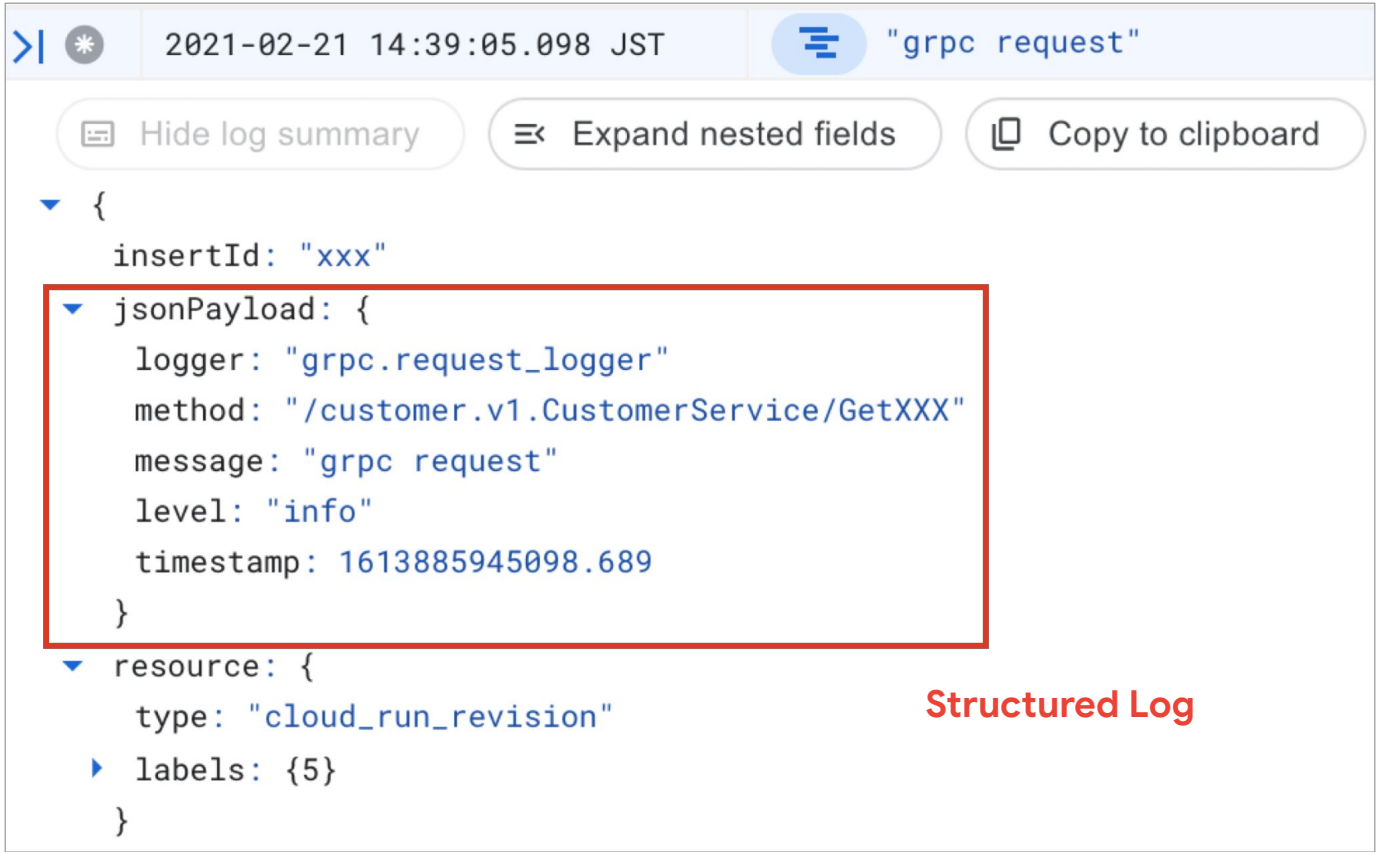
- Request Logs
- Container Logs

https://cloud.google.com/run/docs/logging

Request Logs



Container (Application) Logs



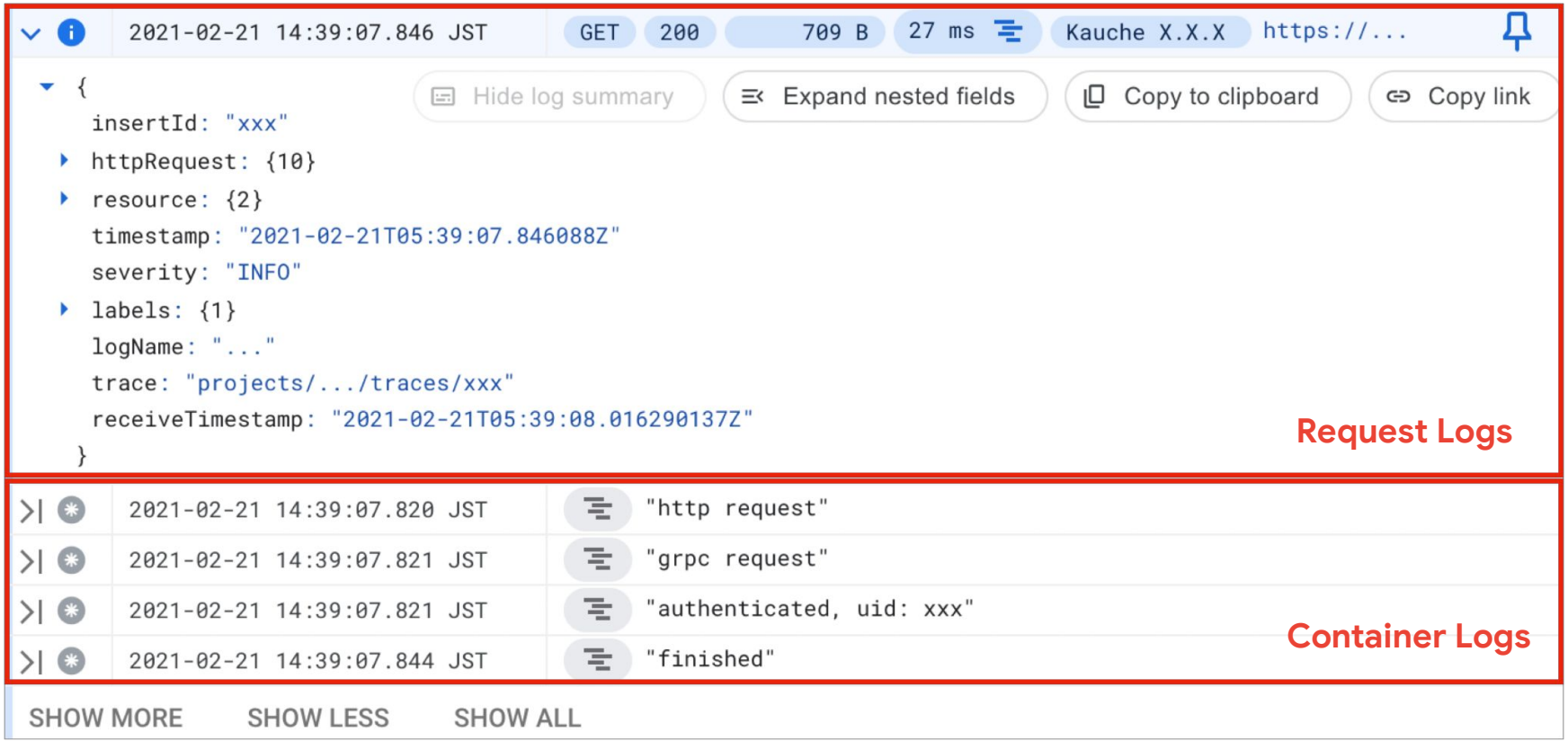
Correlate Logs

Request Logs

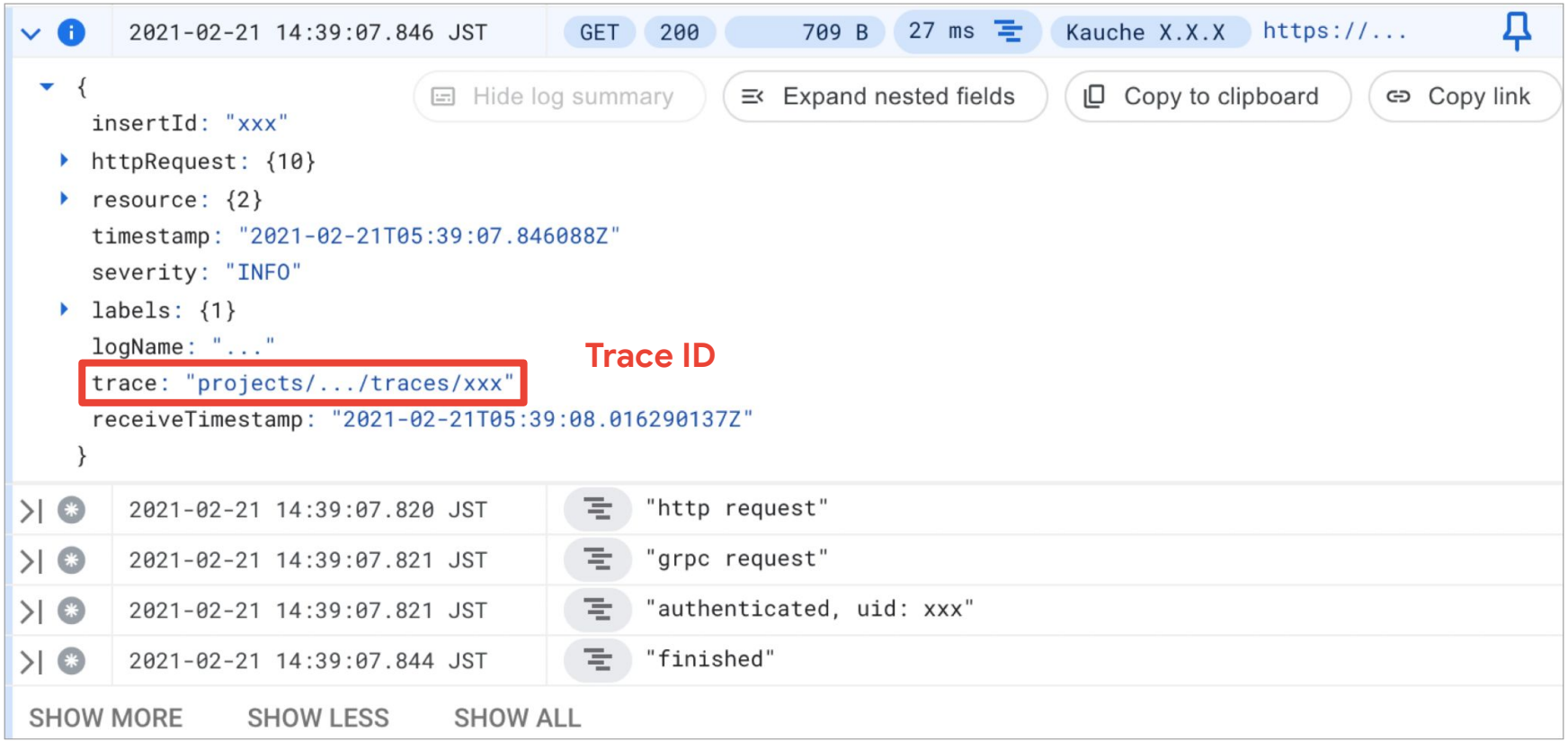
+

Container Logs

Correlate Logs



Correlate Logs



Correlate Logs

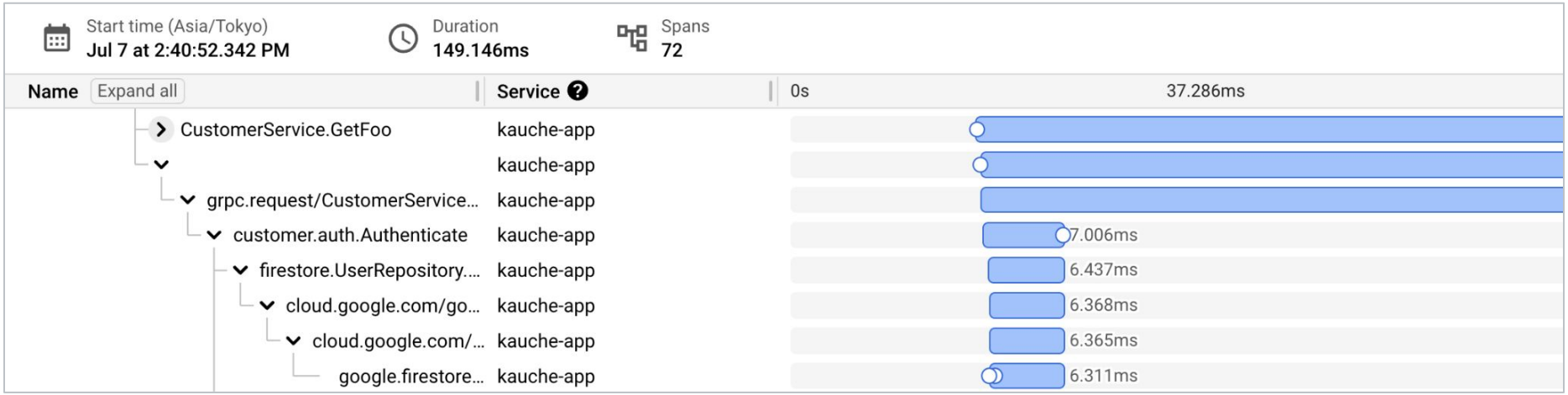
```
{
  "message":    "grpc request",
  "logger":     "grpc.request_logger",
  "method":     "/customer.v1.CustomerService/GetXXX",
  "level":      "info",
  "timestamp": 1613885945098.689
  "logging.googleapis.com/trace": "projects/.../traces/xxx",
}
```

https://cloud.google.com/logging/docs/structured-logging

Approaches

- Logs
- Traces
- Metrics
- SLI / SLO

Cloud Trace

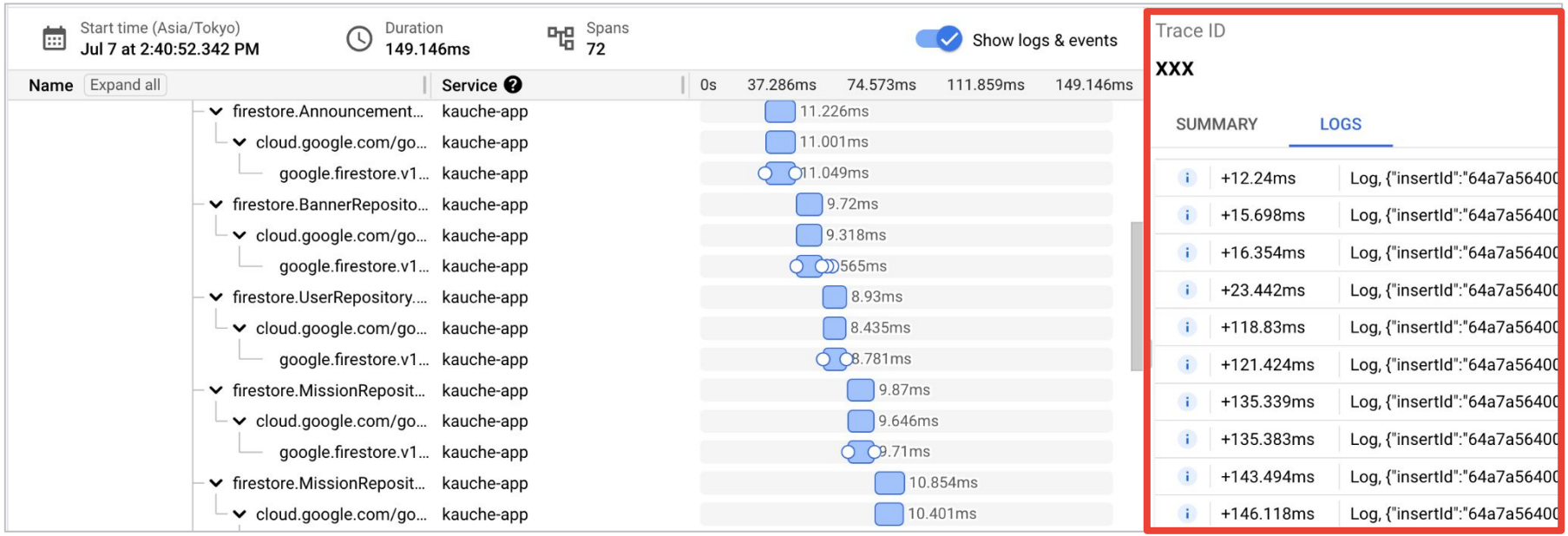

Start time (Asia/Tokyo)
Jul 7 at 2:40:52.342 PM
Duration
149.146ms
Spans
72
Name
Expand all
Service
0s
37.286ms
CustomerService.GetFoo
kauche-app
kauche-app
grpc.request/CustomerService...
kauche-app
customer.auth.Authenticate
kauche-app
7.006ms
firestore.UserRepository....
kauche-app
6.437ms
cloud.google.com/go...
kauche-app
6.368ms
cloud.google.com/...
kauche-app
6.365ms
google.firestore...
kauche-app
6.311ms

OpenTelemetry

OpenTelemetry is a collection of tools, APIs, and SDKs. Use it to instrument, generate, collect, and export telemetry data (metrics, logs, and traces) to help you analyze your software's performance and behavior.

https://opentelemetry.io/

Trace + Logs

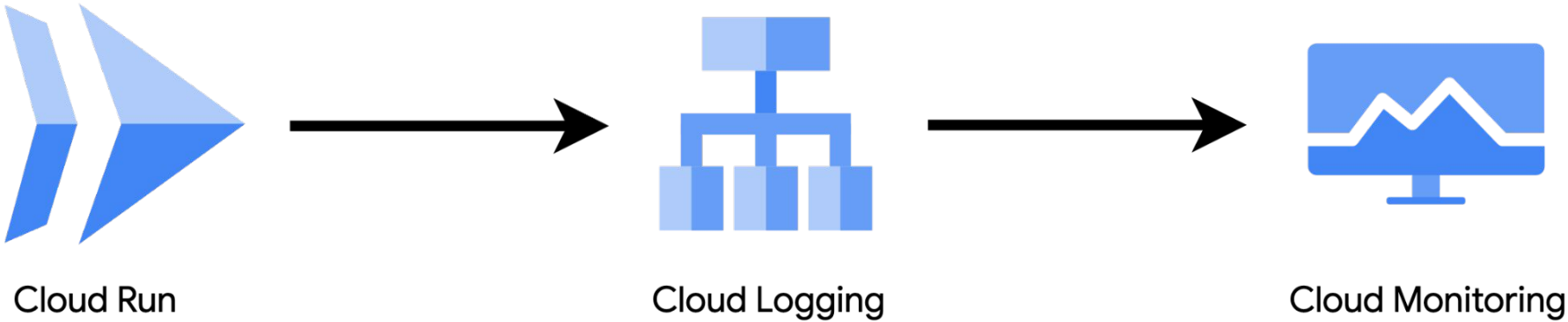


https://opentelemetry.io/

Approaches

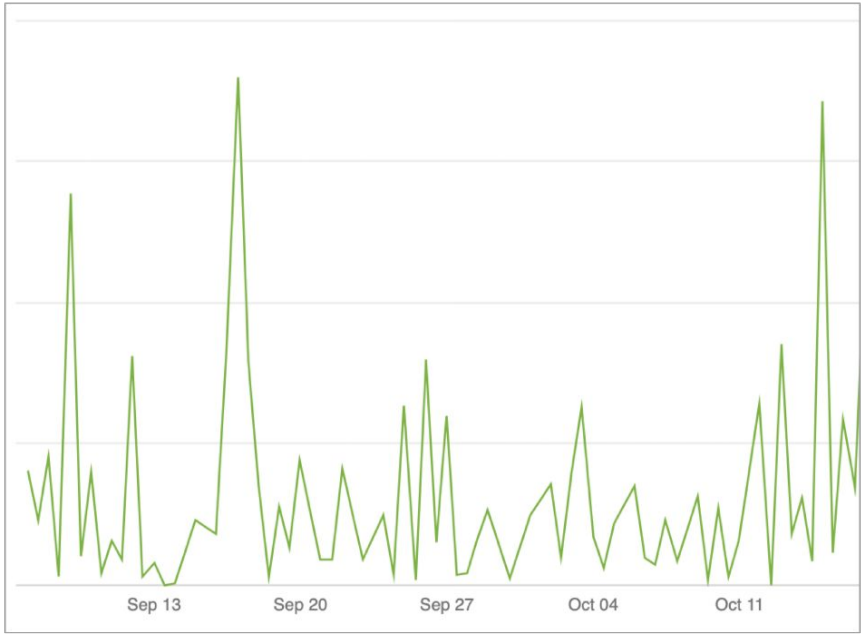
- Logs
- Traces
- Metrics
- SLI / SLO

Log Based Metrics

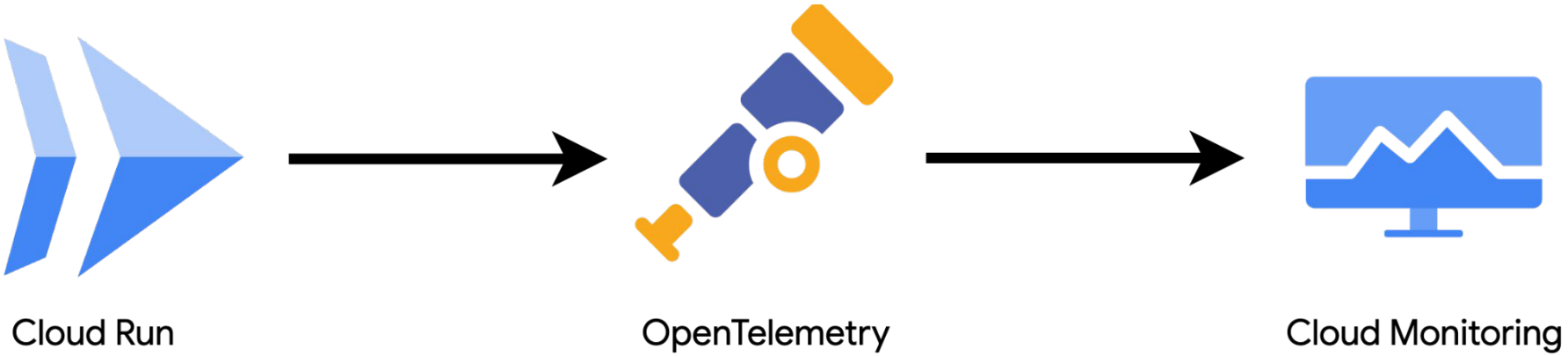


Log Based Metrics

user/CloudRunCustomerAPIErrorLogs	Counter	72 B	344 B	httpRequest.requestUrl=~"https://..." httpRequest.status!="200" httpRequest.status!="302" httpRequest.status!="404" resource.type="cloud_run_revision" resource.labels.project_id="xxx"



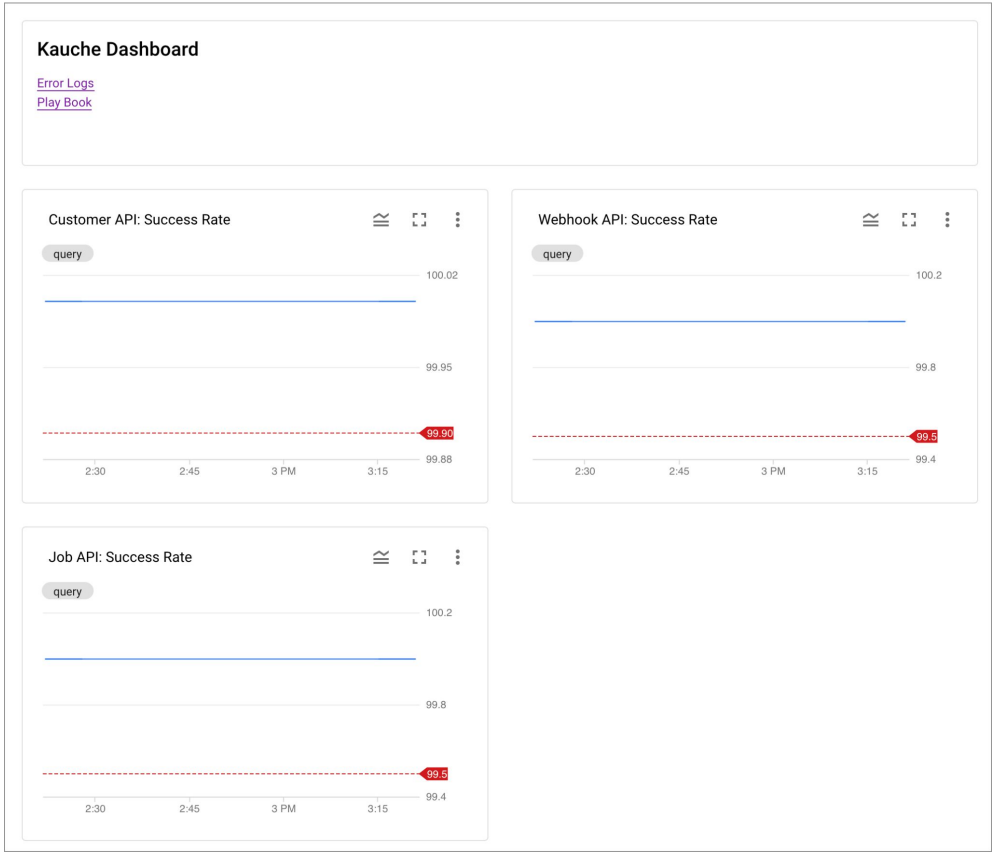
OpenTelemetry



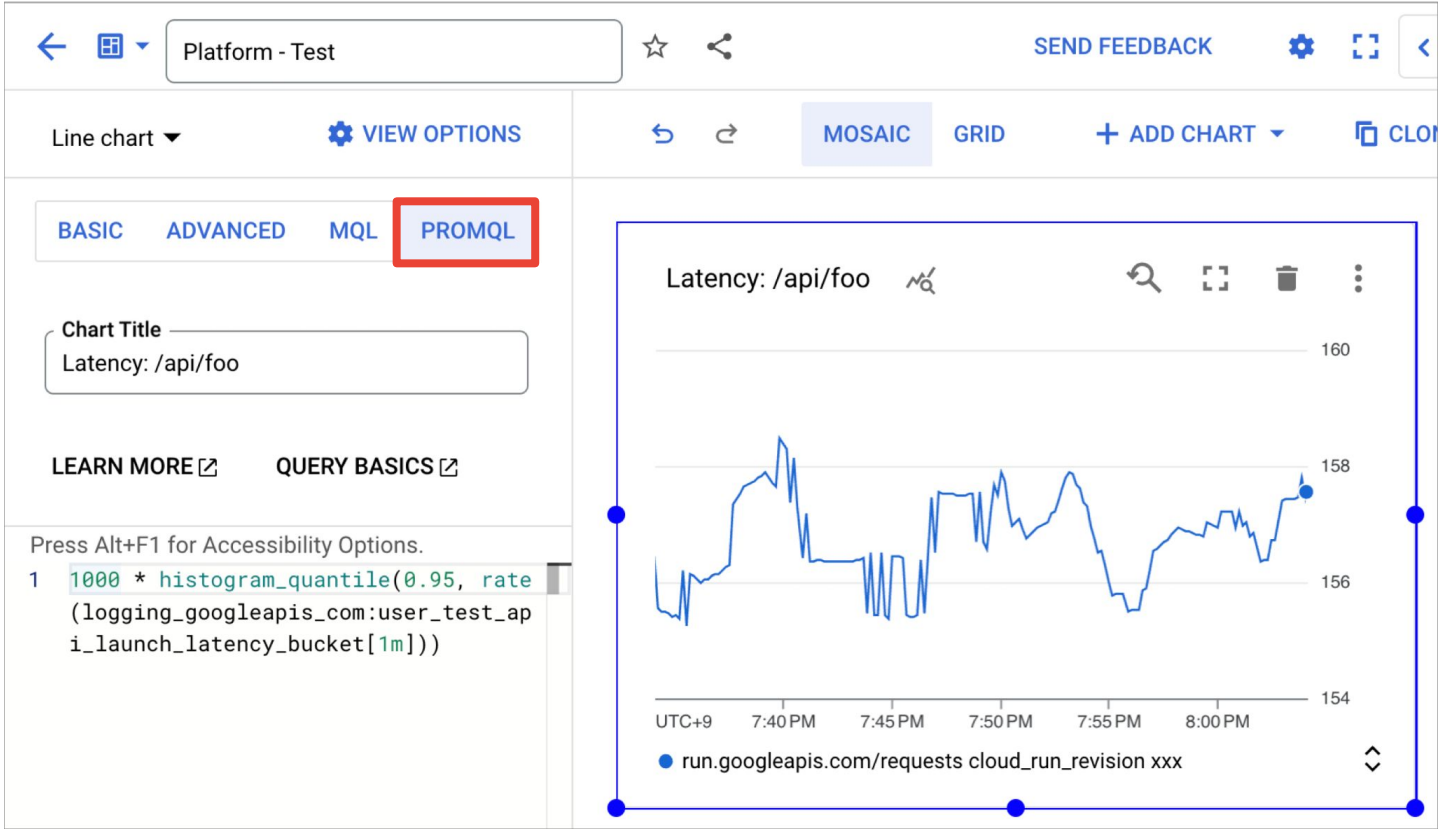
Approaches

- Logs
- Traces
- Metrics
- SLI / SLO

Cloud Monitoring



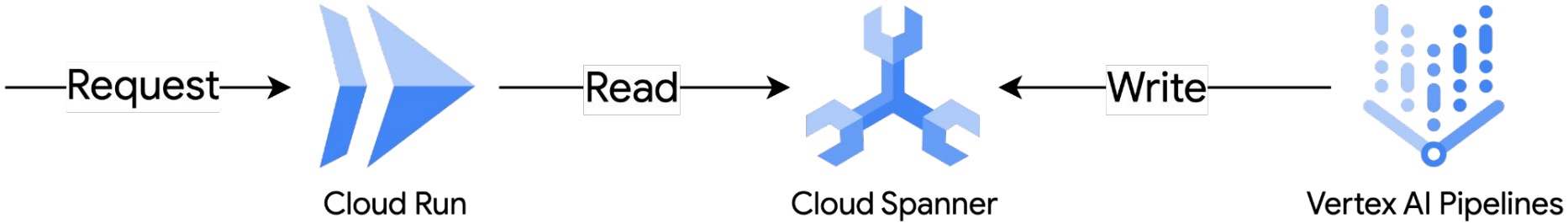
PromQL


Platform - Test
SEND FEEDBACK
Line chart
VIEW OPTIONS
MOSAIC
GRID
ADD CHART
BASIC
ADVANCED
MQL
PROMQL
Chart Title
Latency: /api/foo
LEARN MORE
QUERY BASICS
Press Alt+F1 for Accessibility Options.
1 1000 * histogram_quantile(0.95, rate
(logging_googleapis_com:user_test_ap
i_launch_latency_bucket[1m]))
Latency: /api/foo
160
158
156
154
UTC+9
7:40 PM
7:45 PM
7:50 PM
7:55 PM
8:00 PM
run.googleapis.com/requests cloud_run_revision xxx

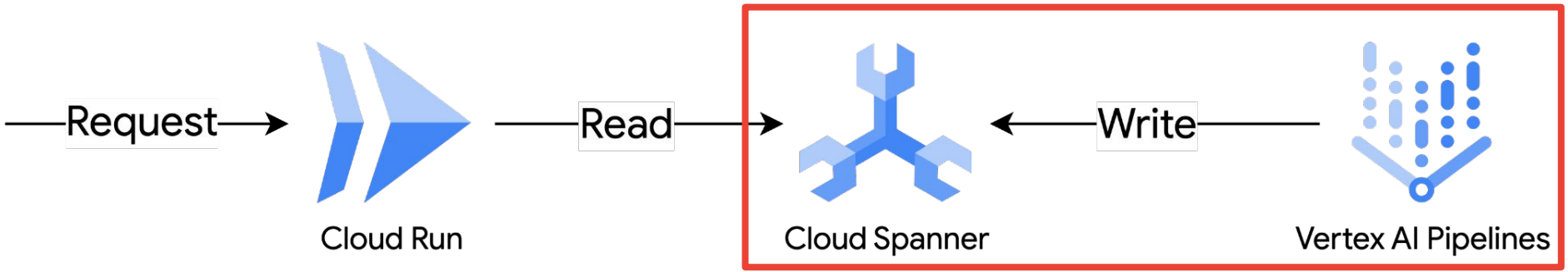
04

AI / ML

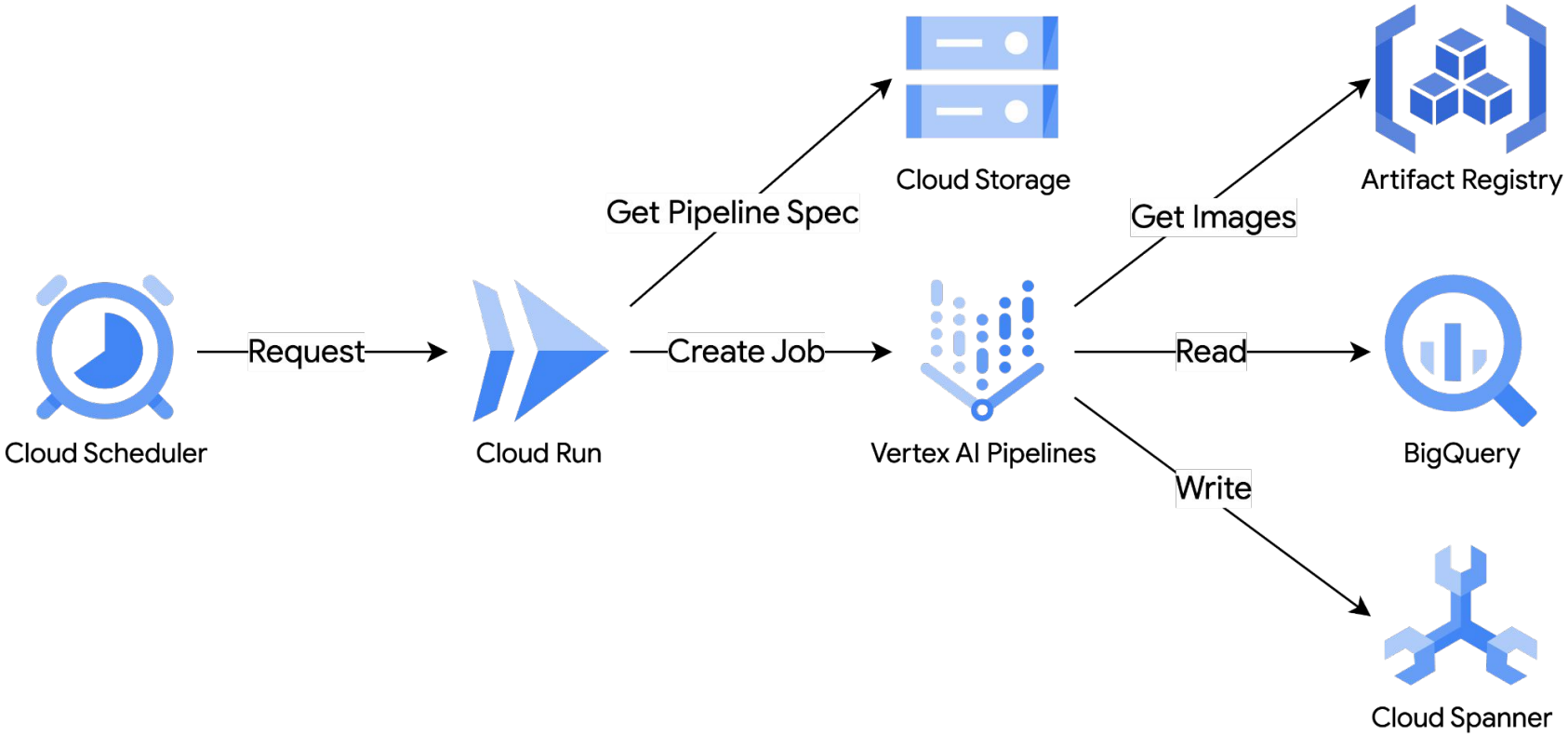
Architecture - Recommendation



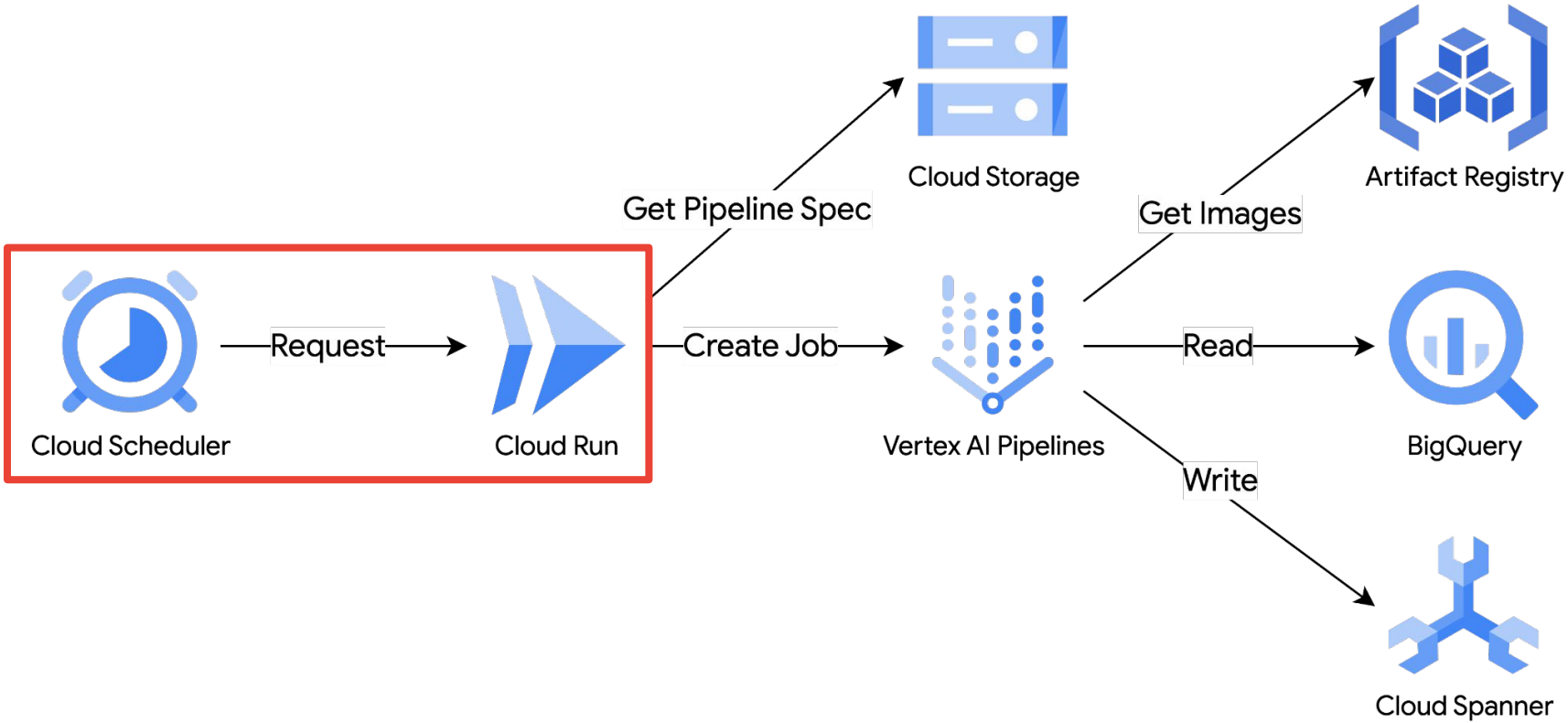
Architecture - Recommendation



Scheduling Vertex AI Pipelines



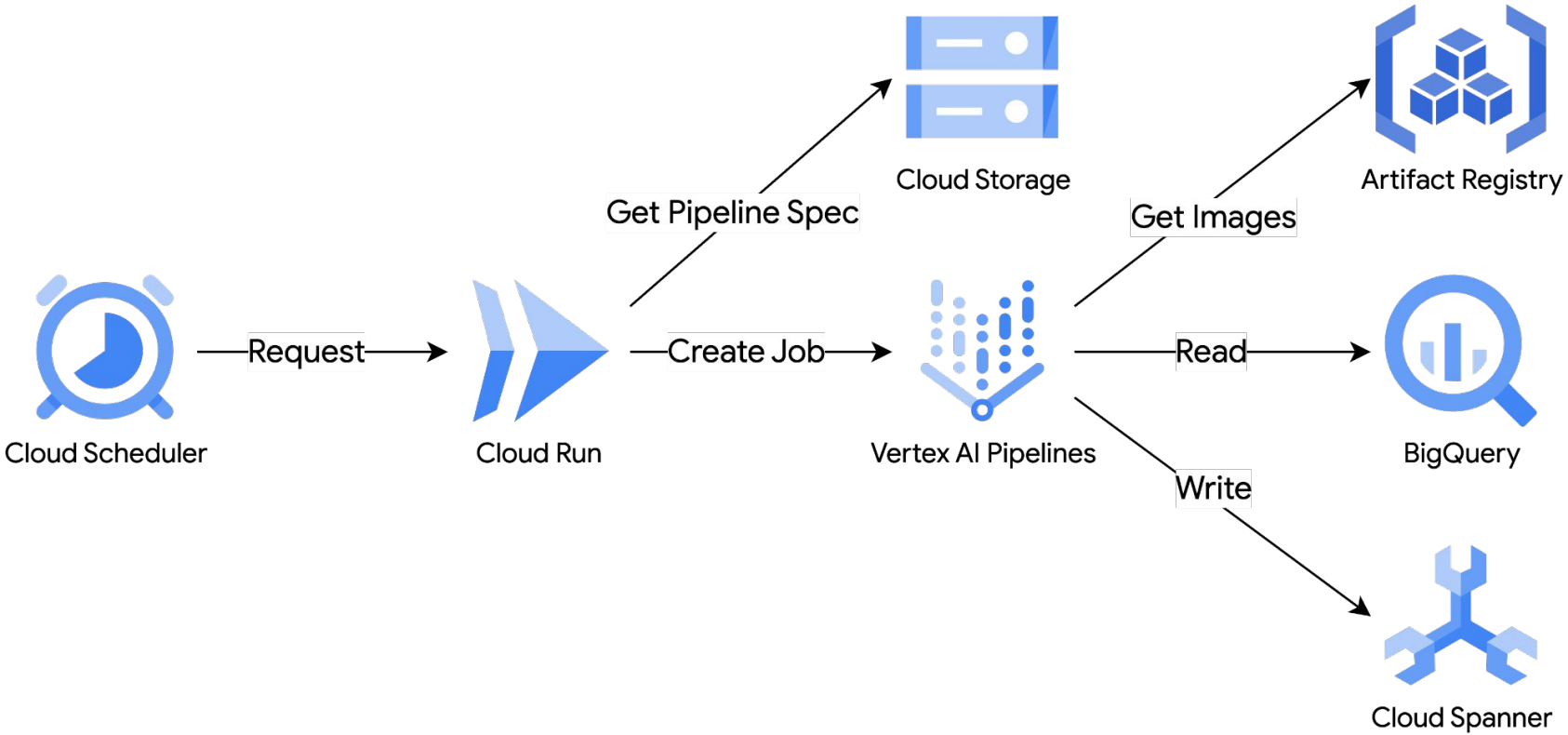
Scheduling Vertex AI Pipelines



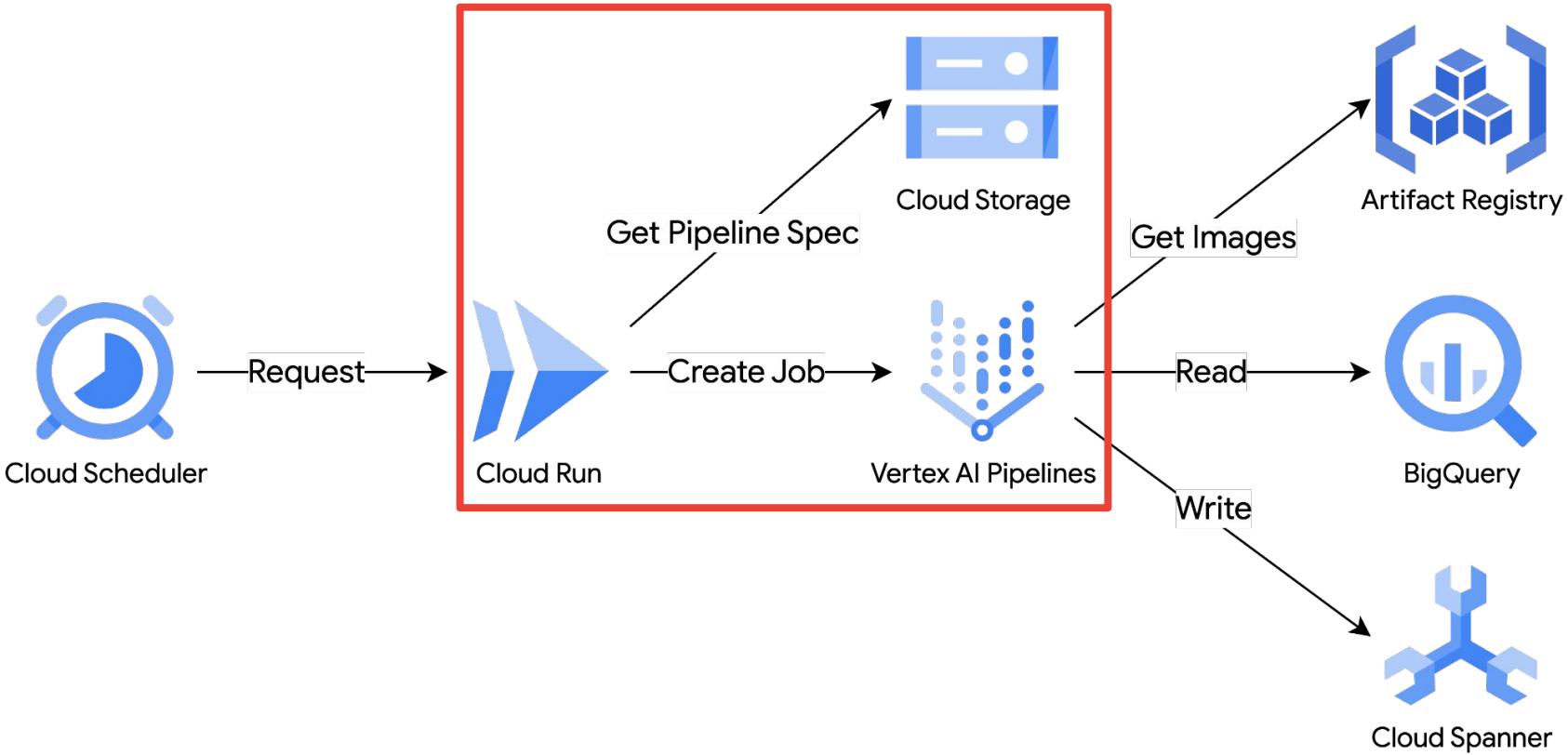
Scheduling Vertex AI Pipelines

```
resource "google_cloud_scheduler_job" "trigger-ai-pipeline" {
  schedule  = "12 0 * * *"
  time_zone = "Asia/Tokyo"
  http_target {
    http_method = "POST"
    uri         = "https://.../execute_pipeline"
    body        = base64encode("{\"pipeline_spec_path\":\"<gcs bucket name>/path/to/pipeline_spec.json\"}")
    oidc_token {
      service_account_email = "..."
      audience              = "..."
    }
  }
}
```

Scheduling Vertex AI Pipelines



Scheduling Vertex AI Pipelines



Scheduling Vertex AI Pipelines

```
110
111   // Creates a PipelineJob. A PipelineJob will run immediately when created.
112   rpc CreatePipelineJob(CreatePipelineJobRequest) returns (PipelineJob) {
113     option (google.api.http) = {
114       post: "/v1/{parent=projects/*/locations/*}/pipelineJobs"
115       body: "pipeline_job"
116     };
117     option (google.api.method_signature) =
118         "parent,pipeline_job,pipeline_job_id";
119   }
120
```

https://github.com/googleapis/googleapis/blob/ef2e2ea532248d6dc40a56bc6c95cea858ba31b6/google/cloud/aiplatform/v1/pipeline_service.proto#L111C1-L119

Scheduling Vertex AI Pipelines

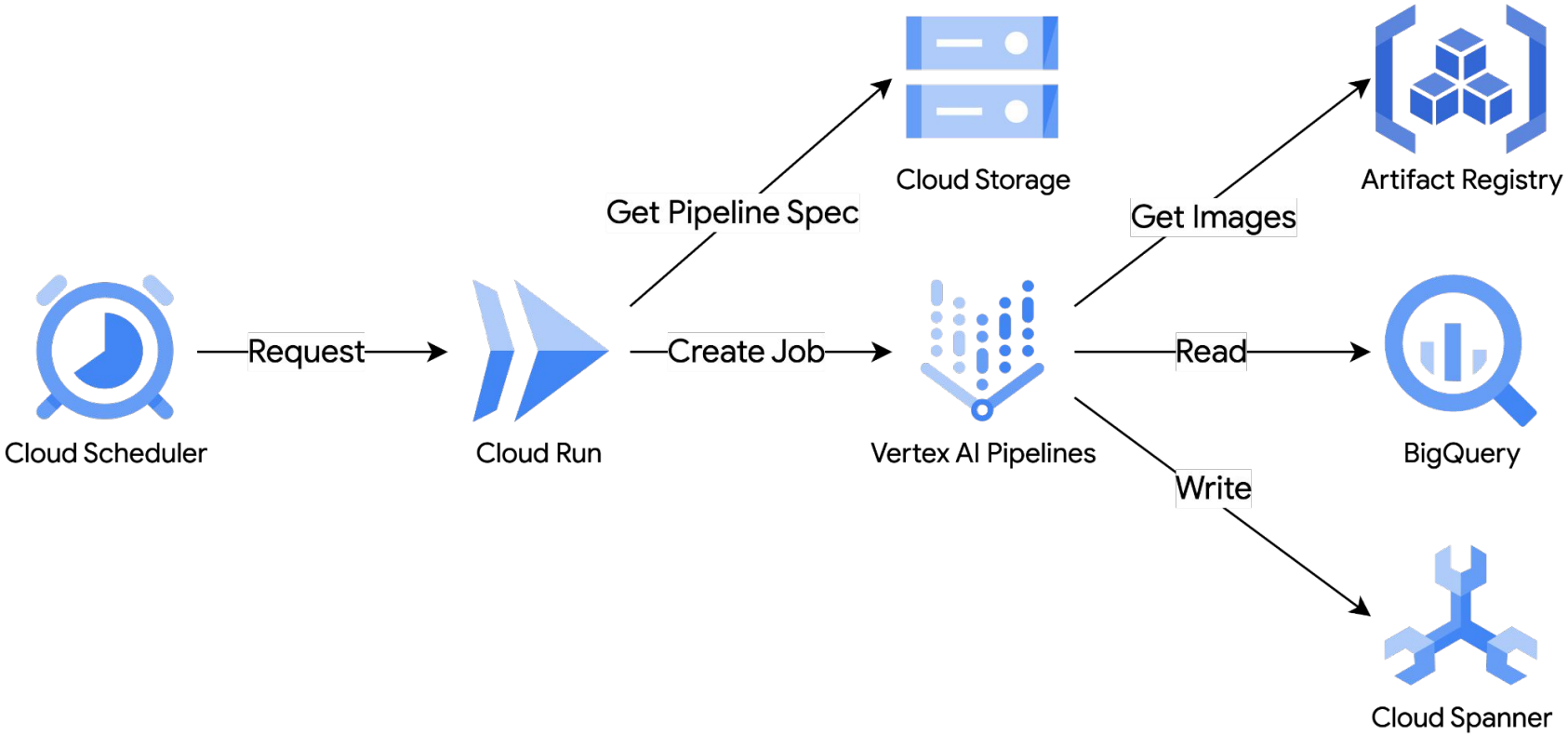
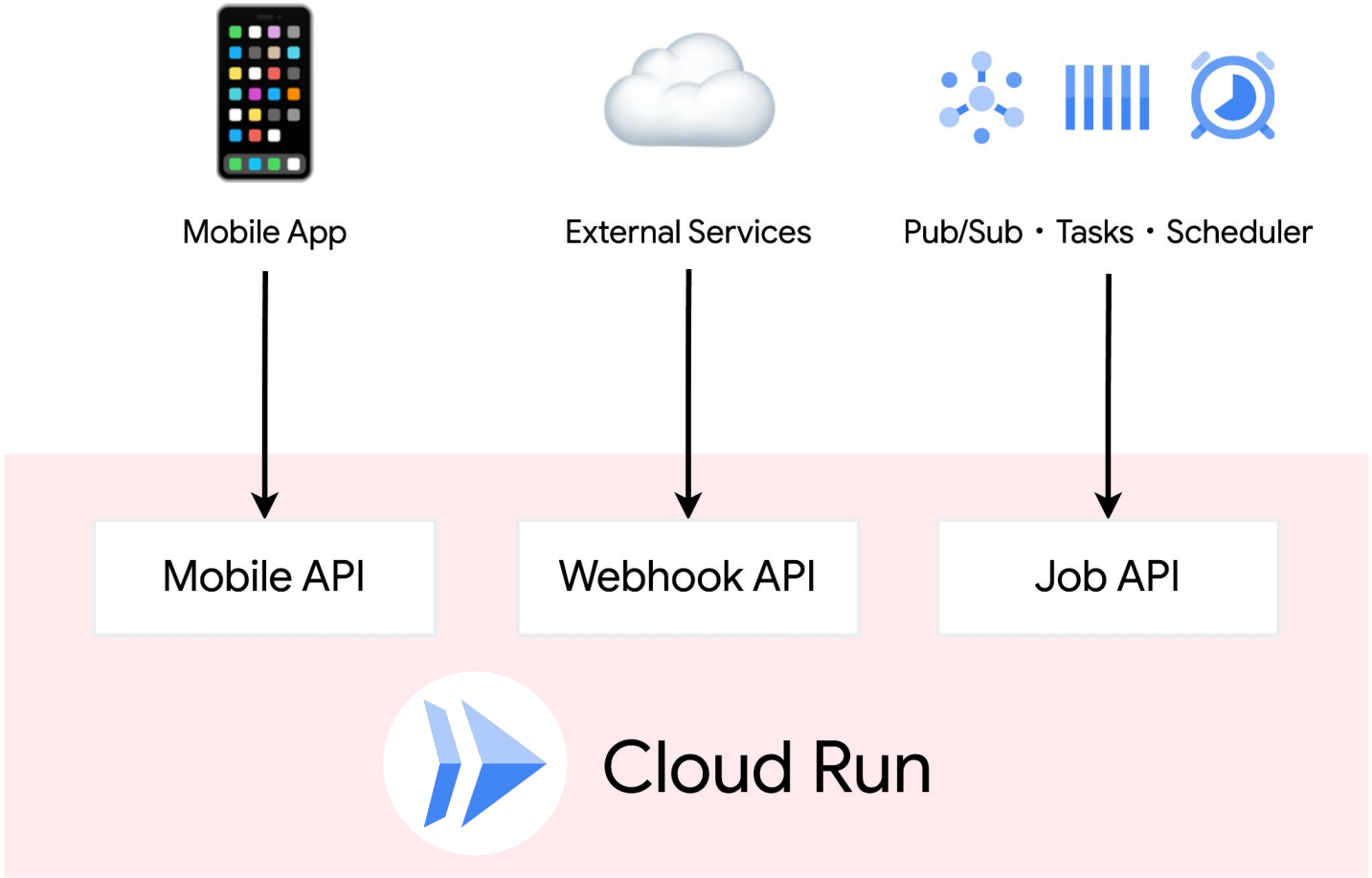


Table of Contents


Mobile App
External Services
Pub/Sub • Tasks • Scheduler
Mobile API
Webhook API
Job API
Cloud Run

Thank you.

Google Cloud