



GCloud 15 Service Definition for Developer Services

Agenda

- 1 Cloudflare Connectivity Cloud Overview
- 2 Cloudflare Developer Platform Overview
- 3 Cloudflare for Developers
- 4 Core components of the platform
- 5 Customer Success Resources

Connectivity cloud architecture & key differentiators

**One network,
one control
plane**

**Composable,
programmable
services**



**Global scale and
resilience**

**Simplified
management
interface**

One network, one control plane on a global scale



335+ cities

in 125+ countries, including mainland China



w/180+ cities

for AI inference powered by GPUs



~20%

of web properties sit behind Cloudflare



227 billion

cyber threats blocked every day



348 Tbps

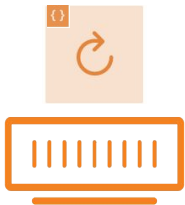
of network capacity (and growing)





Cloudflare Developer Platform Overview

The first serverless platform that truly scales down to zero

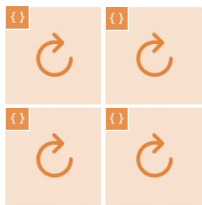


Generation One

One box;
One application

Weeks to deploy

OS, runtime,
libraries and
application

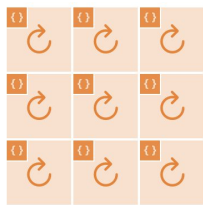


Generation Two

No metal box;
Virtual machines

Hours to deploy

OS, runtime,
libraries and
application

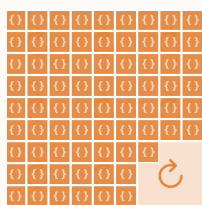


Generation Three

No virtual machine;
Containers/Kubernetes

Minutes to deploy

Runtime, all
libraries, and
applications



Generation Four

No containers;
Serverless V8 isolates

Seconds to deploy

Uncommon
libraries and
applications

0ms

Cold starts (no container
runtime) for auto-scaling

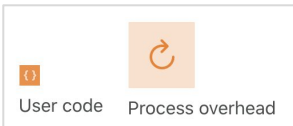
10x

More cost-efficient than
similar serverless
platforms

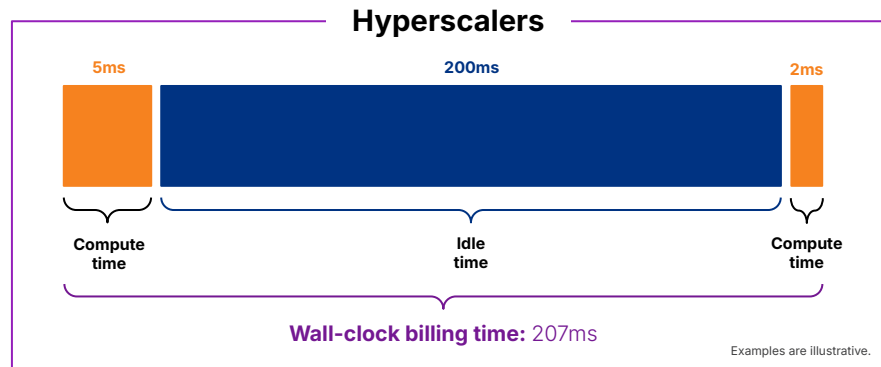
0

Scales down to zero

No “DevOps” or
infrastructure
management (scaling up
happens automatically)



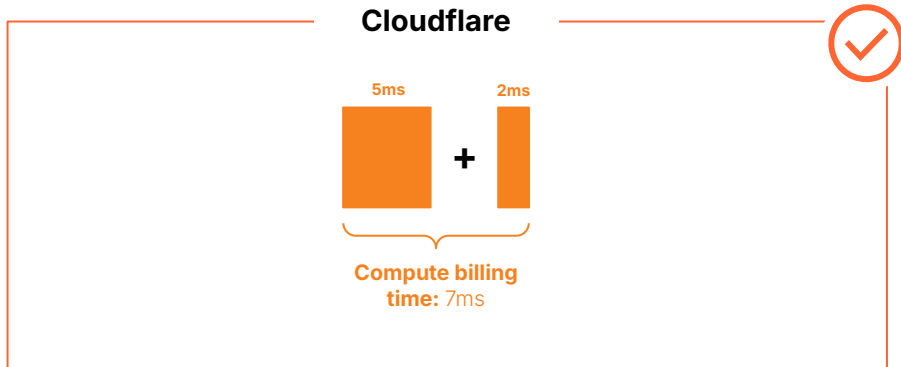
Serverless billing that scales down to zero



Most compute platforms today — including those that are considered to be “serverless” — are based on containers or virtual machines under the hood.

The programming model of containers requires them to stay running while waiting on external tasks, meaning they cannot truly scale to zero.

Regardless of whether code is actively executing or just waiting on slow external resources or even LLMs, the cost of running the container (wall-clock time) is passed to the user.

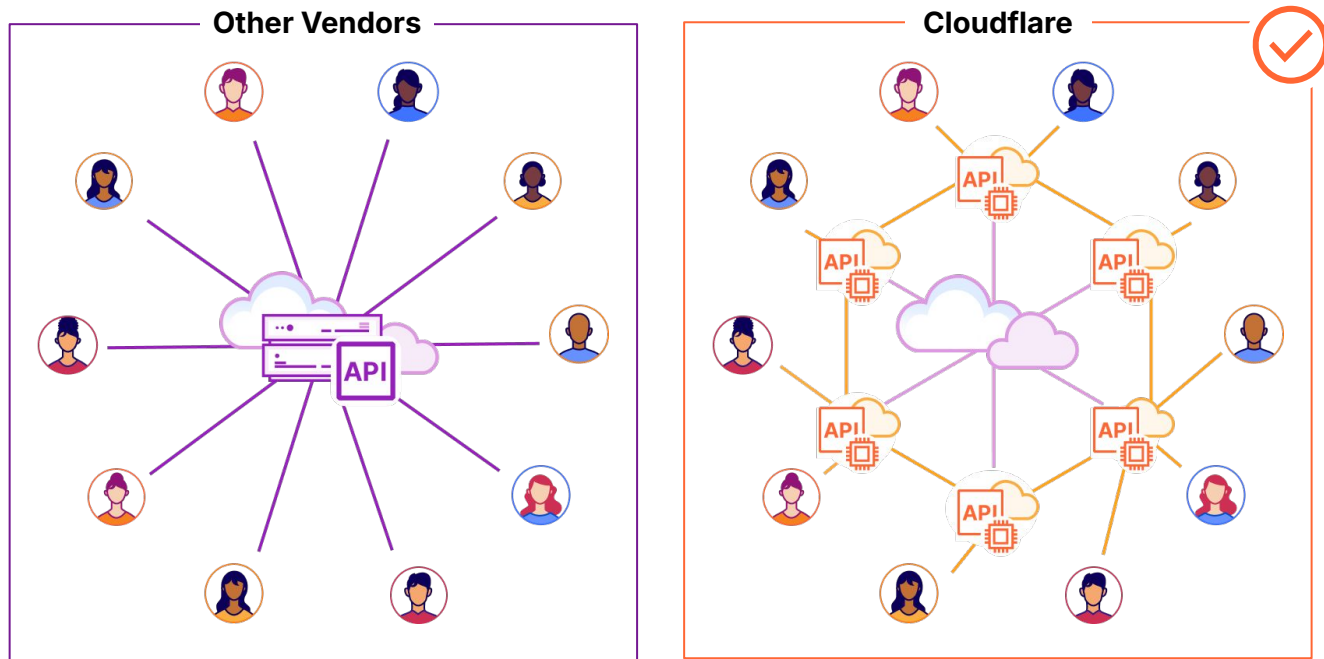


By utilizing V8 isolates, Workers can effectively scale to zero when there are no requests — meaning you never have to keep instances of your application sitting idle, waiting for something to happen.

Workers' pricing model extends the concept of “scale to zero” even further, billing only if an application is actually performing work (compute-time).

The CPU or GPU can be used for other tasks, which improves the utilization of and, therefore, return on our invested capital.

Orchestrating AI + APIs with Cloudflare Workers



Cloudflare can schedule workloads near users, GPUs, and/or APIs — **maximizing throughput and minimizing latency** — to enable agents to make real-time decisions.

Orchestration & execution

Workers are able to interface with APIs, LLMs, external or internal services, or even spin up browsers.

With compute time pricing, users do not have to pay while waiting on execution.

Durable Objects and **Workflows** can guarantee execution and retry as needed.

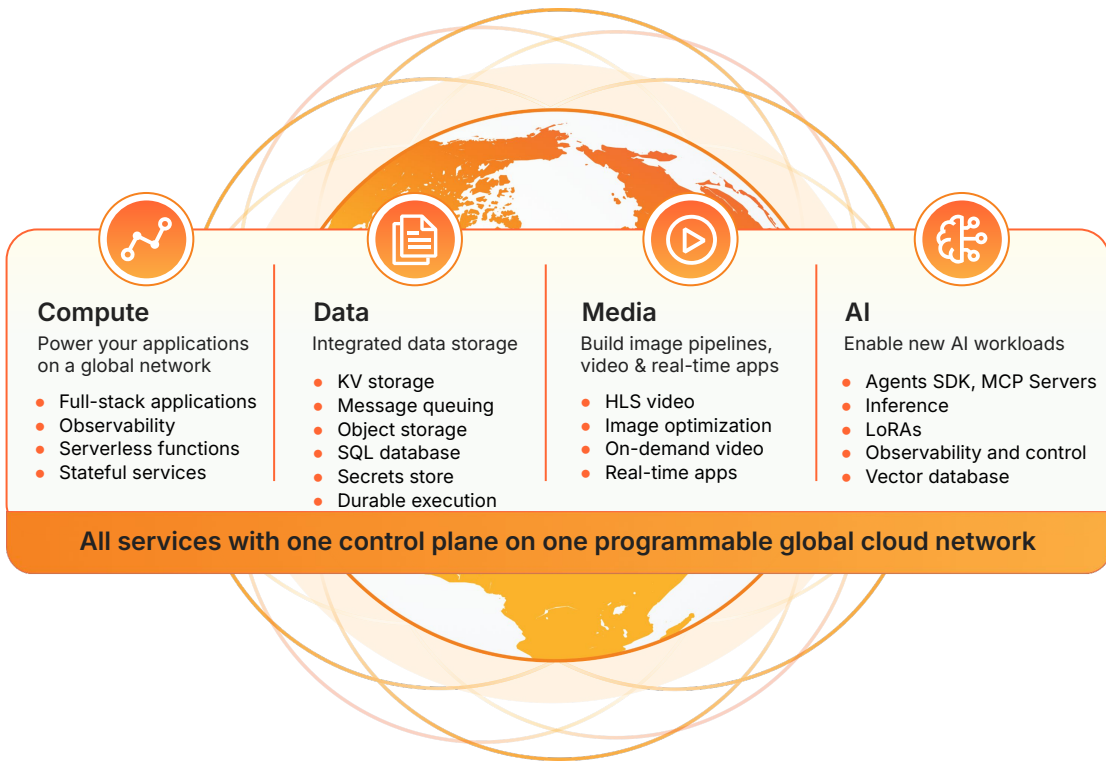
Cloudflare for Developers

Deploy applications to region: Earth

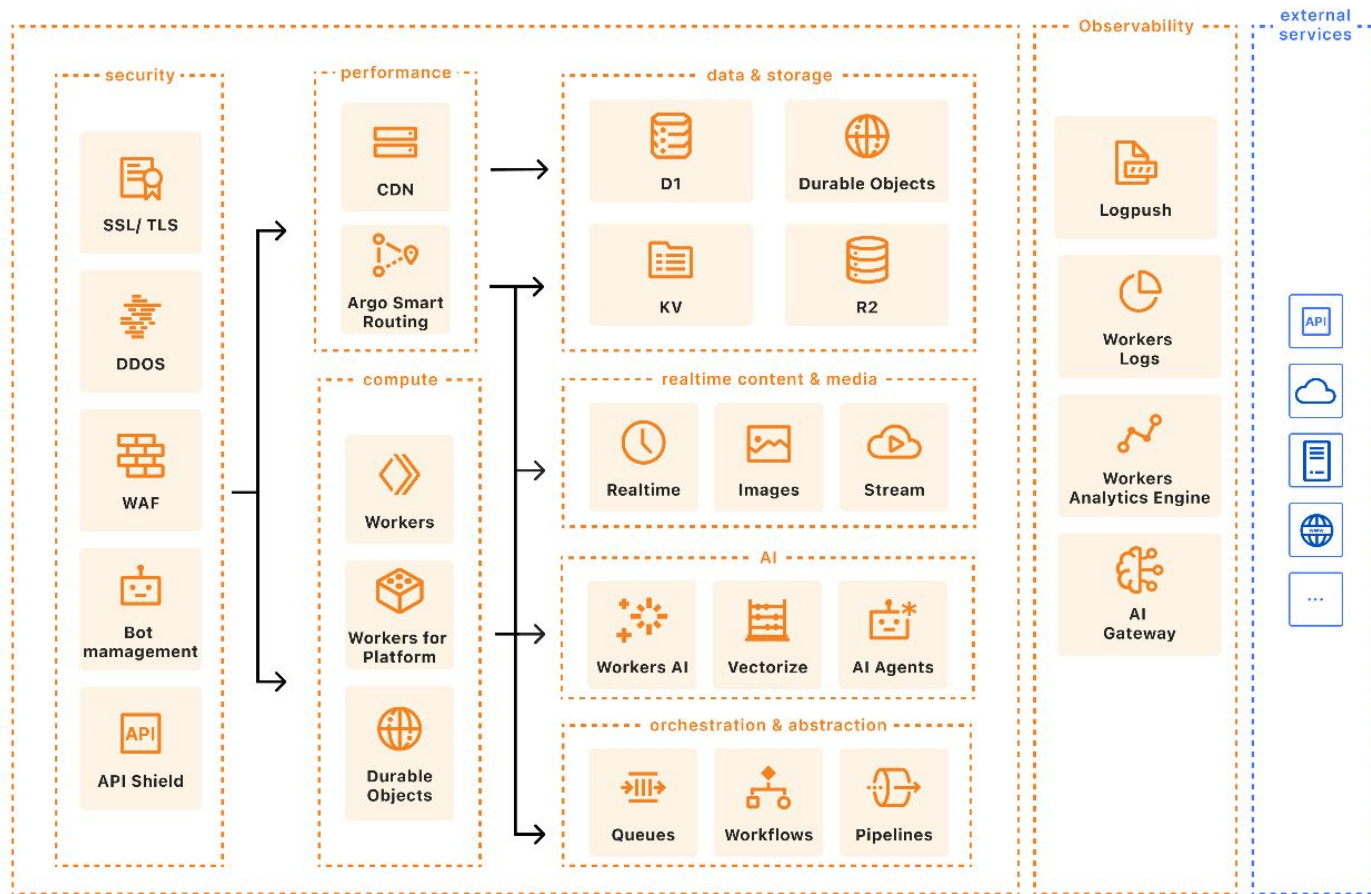
One unified platform for building full-stack and AI applications uniquely powered by an intelligent, programmable global cloud network.

This enables organizations to ...

- ✓ **innovate faster,**
- ✓ **lower costs, and**
- ✓ **scale effortlessly.**



Full-stack application reference architecture



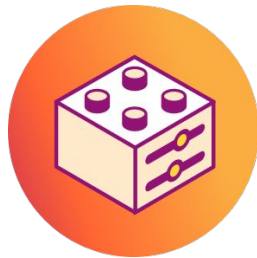
Goals for Developers



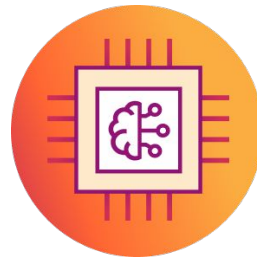
**Remove friction &
improve developer
velocity**



**Provide
primitives
to build on**



**Enterprise grade
solutions and
scale**



**Enable new
AI
workloads**

Containers

Overall platform:

Managed container service to run heavy workloads beyond Workers

Globally distributed infrastructure meant for full stack applications.

Containers

Simple, scalable, global containers

Use Cases / Benefits



Run new heavy workloads
beyond Workers



Stateful Containers (with DOs)



Backend tasks interfaced by
Workers



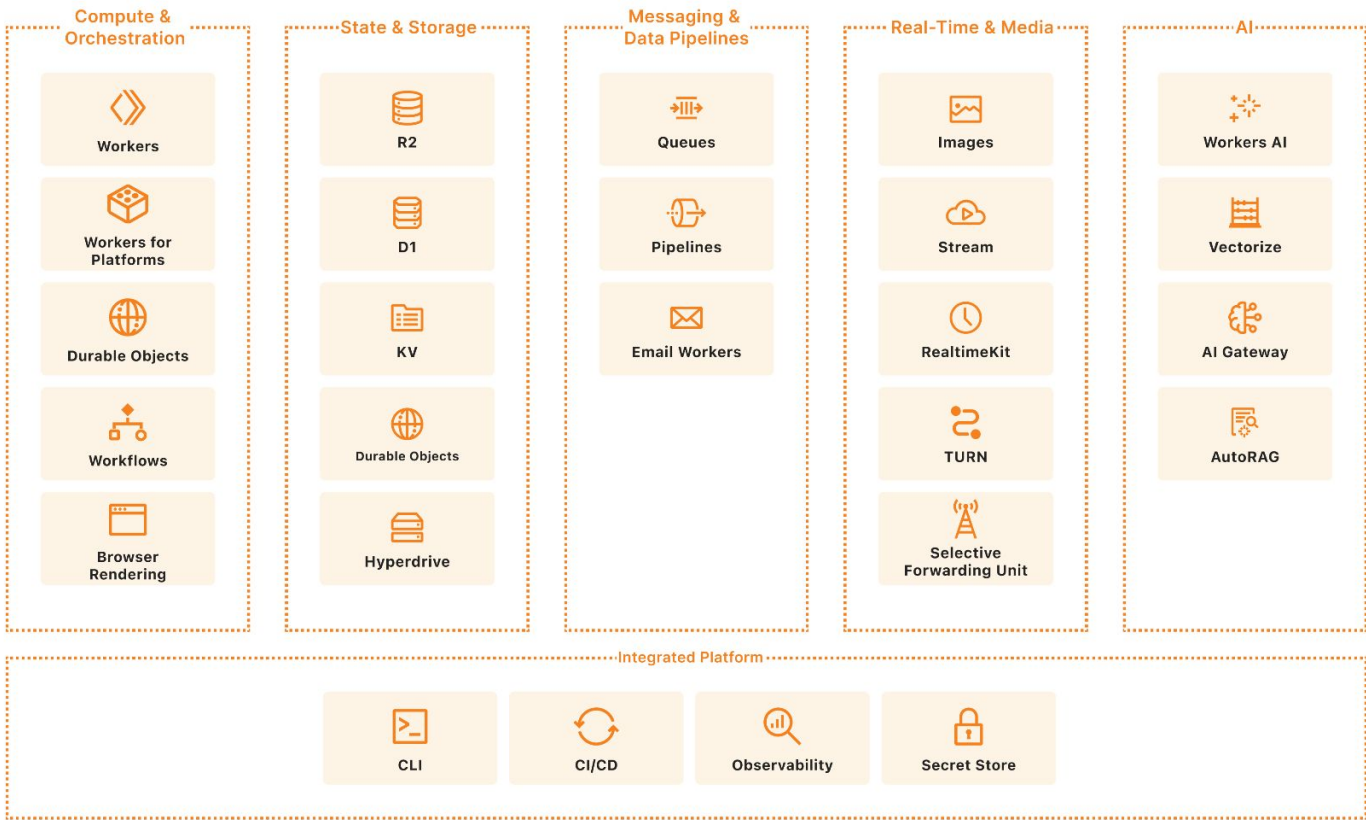
Cron Containers & Addressability



Global platform, close to users

Core components of the platform

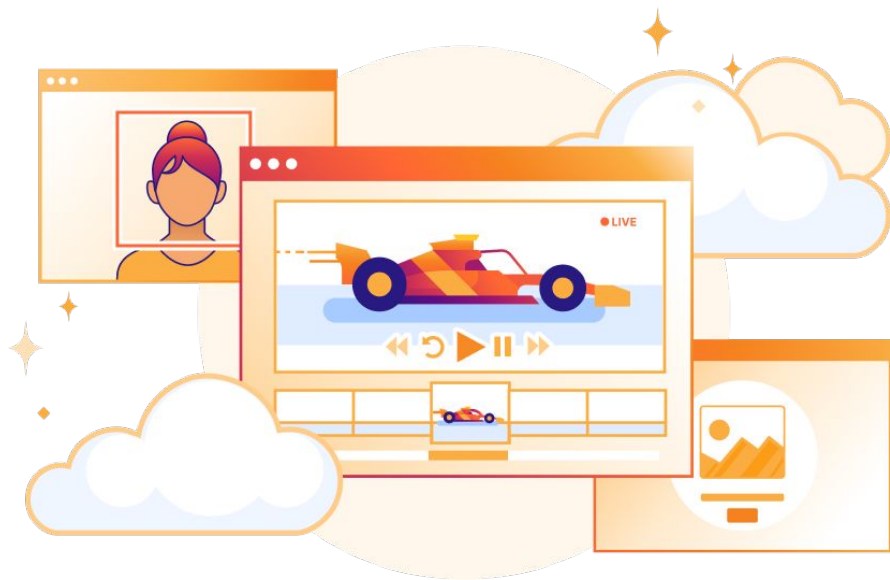
Building blocks to create full-stack applications



Comprehensive data solutions for modern development

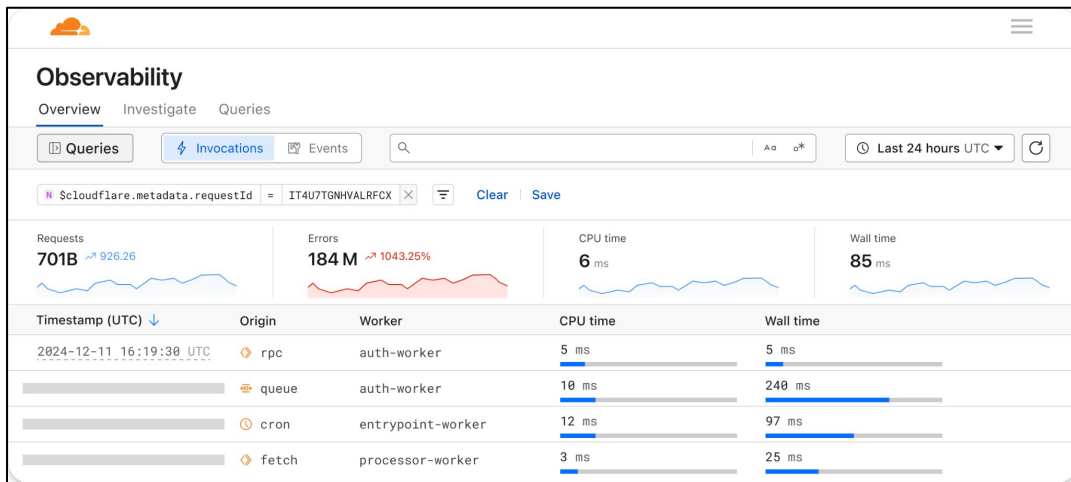
Category	Capability	Key offerings
Storage	Zero-egress object storage & key value storage	• R2 (with Data Catalog for Iceberg tables) • Workers KV
Databases	Serverless SQL & state management	• D1 (Outerbase) • Durable Objects
Streaming	Real-time ingestion & processing	• Pipelines (Arroyo) • Queues
Connectivity	Accelerated access to existing databases	• Hyperdrive
Real-time	Voice, video, and collaborative experiences	• RealtimeKit (Dyte) • Durable Objects
AI & ML	Vector search & AI workflows	• Vectorize • Workers AI

Toolkit to build real-time audio, video experiences



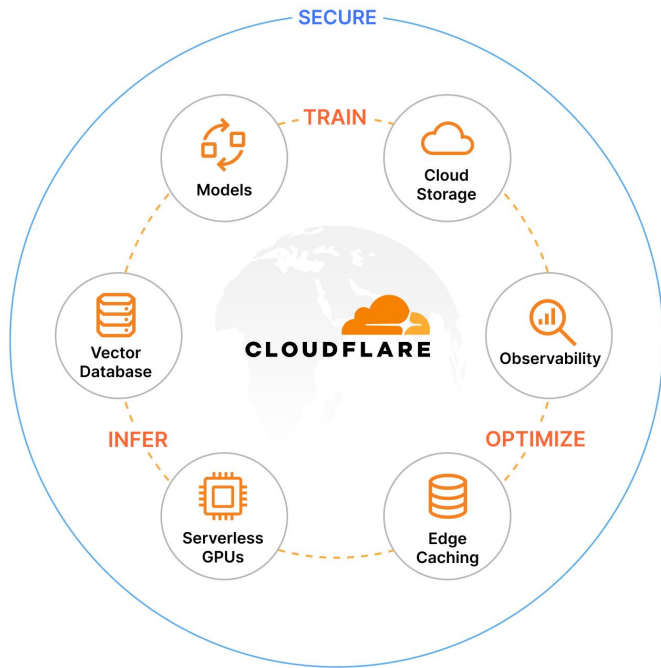
- Deliver low-latency, high-fidelity voice/video at scale
- Add media capabilities, customize UI with minimal configuration
- High observability to debug and optimize performance

Observability



- Capture high cardinality and high dimensionality telemetry data
- Compare metrics across Workers
- Save and share queries

A holistic, end-to-end AI solution



1) TRAIN

Zero Egress Cloud Storage with **R2**

2) INFER

Generate, store, and search embeddings with **Workers AI, Agents SDK, & AutoRAG**

3) OPTIMIZE

Gain visibility, control costs and improve performance via **AI Gateway**

4) SECURE

Protect and Defend with **AI Gateway Guardrails, Zero Trust** and **API Security**

Customer Success Resources

Ensuring your success



[Community](#)



[Docs](#)



[Resource Hub](#)



[Help Center](#)



[Discord](#)



[Blog](#)



[theNET](#)



[Cloudflare Radar](#)



[CloudflareTV](#)



[YouTube](#)

Thank you