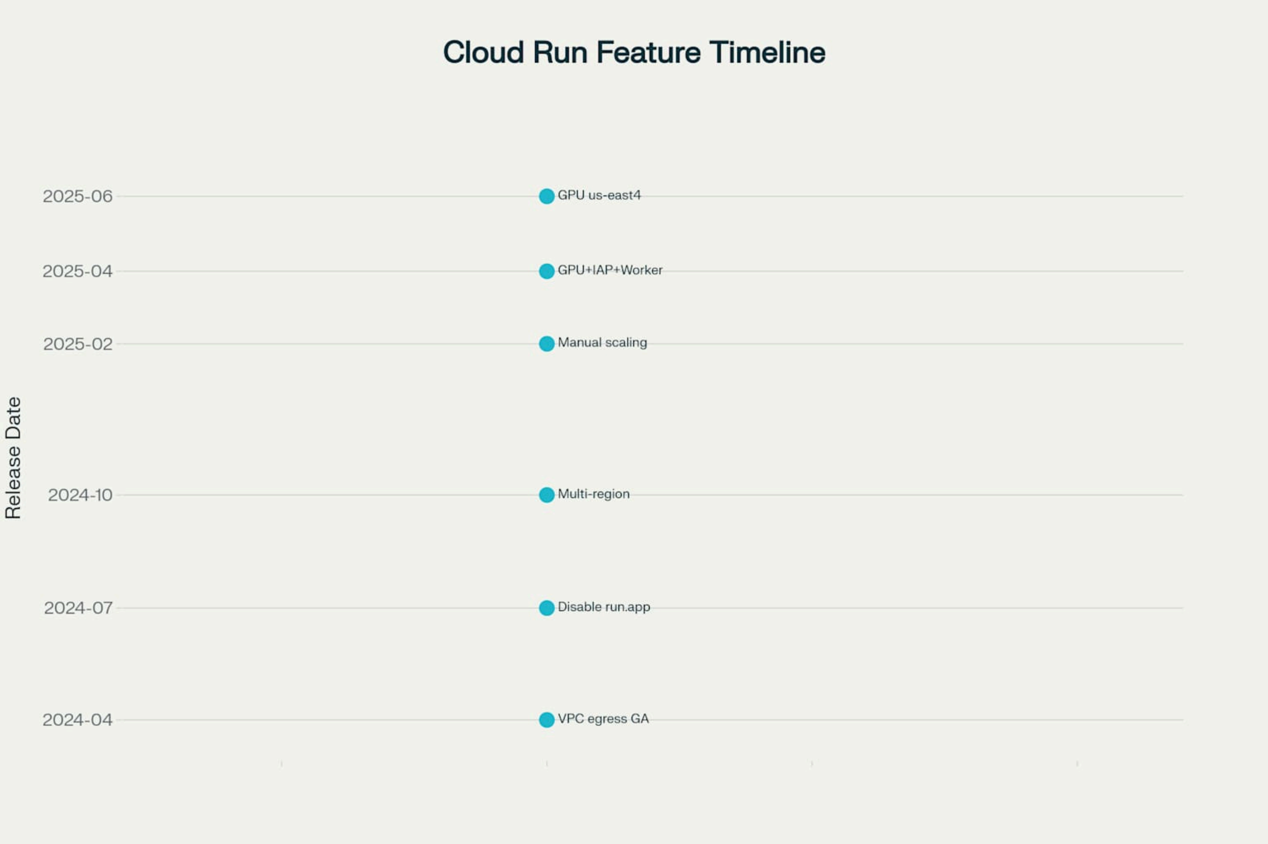
Mastering Google Cloud Run (June 2025 Edition)

Google Cloud Run is Google's fully managed, container-native, serverless platform that scales stateless workloads from zero to planet-scale without servers to manage[1] [2]. Since its GA launch in 2019, Cloud Run has added GPUs, multi-region deployment, manual scaling, direct VPC egress, worker pools, and dozens of developer-productivity improvements up to 26 June 2025[1] [3].

Below is a deep-dive tutorial covering every feature and use case, cross-verified with official docs, blog posts, and community best practices.



Timeline of major Cloud Run feature releases (2024–2025)

Clickable Table of Contents

Service Basics & Architecture

What Is Cloud Run?

- Runs any OCI container that listens on `$PORT` over HTTP/2 or gRPC[2] [4].
- Per-request billing (CPU + memory + GPU) with scale-to-zero and instance-based billing modes[1] [5].
- Execution environments: first-gen sandbox, second-gen full-Linux (GA Dec 2022)[1] [6].
- Resources: Services (long-lived HTTPS), Jobs (batch/task), Worker Pools (preview Apr 2025)[1] [7].

High-Level Architecture

```
             +-----------+     HTTPS/GCS/PubSub
requests --> |  L7 LB    | --------------+
             +-----------+               |
                     | (Cloud Service Mesh optional)
             +-----------------+
             | Cloud Run fleet |
             +-----------------+
                     | (Direct VPC, VPC Connector, or Public)
             VPC / Internet / Private Services
```

- Each revision is an immutable container spec[1].
- Instances are short-lived VMs patched by Google; cold-start 5–30 s (L4 GPU ~5 s)[4].

Deployment Models (gcloud, YAML, Terraform)

gcloud CLI (source → buildpacks → deploy)

```
gcloud run deploy hello \
  --source . \
  --region=europe-west1 \
  --platform=managed \
  --allow-unauthenticated
```

- Cloud Build builds → Artifact Registry image → Cloud Run service[8].

Image-first

```
gcloud run deploy api \
  --image=europe-west4-docker.pkg.dev/$PROJ/app/api:1.4.3 \
  --region=europe-west4
```

Declarative YAML

```
apiVersion: serving.knative.dev/v1
kind: Service
metadata:
  name: analytics
spec:
  template:
    metadata:
      annotations:
        run.googleapis.com/min-instances: "1"
    spec:
      containers:
      - image: europe-west1-docker.pkg.dev/p/analytics:latest
        env:
        - name: DB_HOST
          value: 10.0.0.3
```

Deploy with:

```
gcloud run services replace analytics.yaml
```

Terraform

```
resource "google_cloud_run_v2_service" "cron" {
  name     = "cron"
  location = "us-central1"

  template {
    containers {
      image = "us-central1-docker.pkg.dev/${var.project}/jobs/cron:v0.9.0"
    }
    vpc_access {
      connector = google_vpc_access_connector.default.id
    }
    scaling {
      max_instance_count = 5
    }
  }
}
```

Module examples published by Google cover domain mapping and IAM[9].

Autoscaling, Concurrency & Instance Limits

Knob	Default	Range	Purpose
Concurrency	80 reqs	1–1000 per instance[10]	Throttle CPU-bound vs IO workloads
Min Instances	0	up to 1000	Keep warm to avoid cold starts[1]
Max Instances	quota-bounded	adjustable	Cap spend and limit fan-out[10]
CPU Boost	off	on/off	Extra CPU during startup (GA Apr 2023)[1]
Manual Scaling	preview Feb 2025	fixed N	Bypass autoscaler for streaming[1]

Performance formula: **QPS = (min(instances)+autoscaled) × concurrency**[11].

Traffic Splitting & Rollbacks

- Cloud Run keeps all revisions; assign % weights or tags[12].
- Example canary 10/90 then promote:

```
gcloud run services update-traffic api \
  --to-revisions rev-2=10,rev-1=90
```

- Instant rollback:

```
gcloud run services update-traffic api --to-latest
```

- Tags give stable URLs per revision for smoke tests[1].

Custom Domains & SSL

1. Map DNS A/AAAA to Google front-ends[13] [14].
2. Managed certs auto-provision; limit 15 certs per project, use wildcard to bypass[15].
3. Disable default `*.run.app` URL (July 2024)[1]:

```
gcloud run services update web --no-default-url
```

VPC Connectors & Serverless VPC Access

Options

Mode	Path	Use Cases
Public (default)	Direct to internet	Simplicity
Serverless VPC Connector	NAT via connector VM	Private DB, Cloud SQL[16]
Direct VPC Egress (GA Apr 2024)	No connector, lower latency, uses subnet[17]	Private NAT, Secure Web Proxy[6]

Connector creation

```
gcloud compute networks vpc-access connectors create svc \
  --region=us-central1 --range=10.8.0.0/28
```

Attach with:

```
gcloud run deploy api --image $IMG --vpc-connector svc \
  --vpc-egress=all-traffic
```

Direct VPC YAML snippet:

```
annotations:
  run.googleapis.com/network-interfaces: |
    [{"network":"default","subnetwork":"subnet-us","tags":"proxy-routed"}]
  run.googleapis.com/vpc-access-egress: all-traffic
```

IAM Roles & Security Best Practices

Pre-defined Roles

Role	Purpose
`roles/run.admin`	Full control[18]
`roles/run.developer`	Deploy but no IAM
`roles/run.invoker`	HTTPS invoke
`roles/run.builder` (preview 2025-01-22)	Build from source[1]

- Principle of least privilege: separate runtime SA vs build SA[8].
- Enable workload identity federation to avoid long-lived keys[18].
- IAP single-click secure ingress (preview Apr 2025)[1].
- Binary Authorization GA Sep 2021 for supply-chain policy[1].

Observability (Logging, Monitoring, Tracing)

- Cloud Logging streams stdout/stderr; tail with `gcloud run services logs tail` (GA Nov 2022)[1].
- Metrics dashboard shows request latency, container start, billable time[10].
- Automatic traces captured; integrate Cloud Trace & Managed Service for Prometheus sidecar (Dec 2023)[1].
- Error Reporting groups 5xx and custom exceptions[19].

CI/CD Integrations (Cloud Build, GitHub Actions)

Cloud Build trigger (`cloudbuild.yaml`)

```
steps:
- name: gcr.io/cloud-builders/docker
  args: ['build','-t','${_IMG}','.']
- name: gcr.io/cloud-builders/docker
  args: ['push','${_IMG}']
- name: gcr.io/google.com/cloudsdktool/cloud-sdk
  args: ['run','deploy','api','--image','${_IMG}','--region','us-central1','--quiet']
images: ['${_IMG}']
substitutions:
  _IMG: us-central1-docker.pkg.dev/$PROJECT_ID/app/api:$COMMIT_SHA
```

- Requires roles: Cloud Run Developer, Artifact Registry Writer, SA User[8].

GitHub Actions reusable workflow

```
jobs:
  deploy:
    permissions:
      contents: read
      id-token: write
    runs-on: ubuntu-latest
    steps:
    - uses: actions/checkout@v4
    - uses: google-github-actions/auth@v2
      with:
        workload_identity_provider: ${{ secrets.WIF }}
        service_account: cicd@$PROJECT.iam.gserviceaccount.com
    - uses: google-github-actions/deploy-cloudrun@v2
      with:
        service: api
        image: ${{ env.IMAGE }}
```

- Official action supports YAML-based services and multiple environments[20] [21].

Event-Driven Patterns (Pub/Sub, Cloud Events, Knative)

- Create Eventarc trigger → Cloud Run service[22].
- Cloud Run services autoconvert HTTP to CloudEvents[23].
- Knative Eventing underpins Cloud Run; Anthos “Events for Cloud Run” simplifies on-prem[23].
- Jobs can be invoked on schedules via Cloud Scheduler hitting HTTPS endpoint or Pub/Sub topic[1].

Hybrid & Multi-Cloud Scenarios (Anthos, GKE)

- Cloud Run for Anthos (GA) runs serverless workloads on GKE on-prem or any cloud[24].
- Multi-region deployment command (preview Oct 2024)[25]:

```
gcloud beta run deploy web --image $IMG \
  --regions=europe-west1,us-east4,asia-northeast1
```

- Anthos Service Mesh can route traffic between Cloud Run, GKE, and Compute Engine[6].

Advanced Networking (Cloud NAT, Ingress/Egress Settings)

- Ingress modes: All, Internal & LB, Internal only[26].
- Configure Cloud NAT for outbound static IP when using connector or direct VPC[26].

```
resource "google_compute_router_nat" "run_nat" {
  name   = "run-nat"
  router = google_compute_router.edge.name
  nat_ip_allocate_option = "AUTO_ONLY"
  source_subnetwork_ip_ranges_to_nat = "LIST_OF_SUBNETWORKS"
}
```

- Secure Web Proxy supported with Direct VPC (Sep 2024)[6].
- Private NAT preview May 2025 for direct VPC egress[1].

Cost Optimisation & Pricing Calculator

- Free tier - 180 k vCPU-s & 360 k GiB-s per month plus 2 M requests[5].
- Use request-based billing for bursty workloads; switch to instance-based if WebSockets or always-on[1].
- Committed use discounts share with GKE/Compute (July 2024)[1].
- Pricing calculator now lists Cloud Run (May 2024)[5].
- GPU pricing per-second, zonal redundancy adds surcharge; preview non-redundant discount for batch jobs[7].

Best Practices & Gotchas

- **Use min instances=1** for low-latency APIs; combine with CPU boost to cut P99 by >50%[1] [10].
- **Cap max instances** to protect backend databases and cost[11].
- **Prefer Direct VPC egress** for lower latency and simpler ops; only use connectors when Shared VPC in another project[17].
- **Shift Traffic Gradually**; tag revisions and run probes before 100% rollout[12].
- **Secure defaults**: disable default URL, enforce IAP, rotate runtime SA keys, enable CMEK for sensitive data[27] [15].
- **Observability first**: set explicit timeouts, instrument OpenTelemetry, and alert on out-of-memory kills[19].
- **Parallel jobs**: watch GPU job parallelism quota; non-zonal redundancy saves cost but is best-effort[7].
- **Regional strategy**: co-locate with data stores; for global apps deploy multi-region + Cloud Armor to reduce latency and improve DR[25].
- **CI storage**: cache Docker layers in Artifact Registry to speed Cloud Build and Actions, avoiding repeated pulls[8].

Last updated: 26 June 2025.

⁂

1. https://cloud.google.com/run/docs/release-notes
2. https://cloud.google.com/blog/products/serverless/cloud-run-bringing-serverless-to-containers
3. https://www.infoq.com/news/2025/06/google-cloud-run-nvidia-gpu/
4. https://cloud.google.com/run/docs/configuring/max-instances-limits
5. https://www.linkedin.com/pulse/demystifying-google-cloud-run-pricing-untangling-cpu-memory-zakaria
6. https://cloud.google.com/run/docs/configuring/networking-best-practices
7. https://cloud.google.com/run/docs/multiple-regions
8. https://cloud.google.com/build/docs/deploying-builds/deploy-cloud-run
9. https://github.com/GoogleCloudPlatform/terraform-google-cloud-run
10. https://www.semanticscholar.org/paper/4fa2777d0f67dda8a897405b996bdb6dc547cd3f
11. https://stackoverflow.com/questions/65753023/google-cloud-run-concurrency-limits-autoscaling-clarifications
12. https://www.pluralsight.com/labs/gcp/cloud-run-revisions-and-traffic-routing
13. https://cloud.google.com/appengine/docs/flexible/securing-custom-domains-with-ssl
14. https://www.googlecloudcommunity.com/gc/Infrastructure-Compute-Storage/Cloud-Run-gt-Custom-Domains-gt-At-most-15-SSL-certificates/m-p/797264
15. https://www.pulumi.com/answers/securing-custom-domains-on-google-cloud-run/
16. https://cloud.google.com/run/docs/configuring/vpc-connectors

17. https://cloud.google.com/run/docs/configuring/vpc-direct-vpc
18. https://trendmicro.com/cloudoneconformity/knowledge-base/gcp/CloudIAM/
19. https://cloud.google.com/run/docs/monitoring-overview
20. https://cloud.google.com/blog/products/devops-sre/deploy-to-cloud-run-with-github-actions/
21. https://github.com/google-github-actions/deploy-cloudrun
22. https://cloud.google.com/eventarc/standard/docs/run/route-trigger-cloud-pubsub
23. https://atamel.dev/posts/2020/10-09_events_cloud_run_anthos_knative_eventing/
24. https://www.devoteam.com/expert-view/anthos-1-year-of-hybrid-and-multi-cloud-application-modernisation/
25. https://cloud.google.com/run/docs
26. https://www.googlecloudcommunity.com/gc/Serverless/Cloud-run-egress-traffic-to-internet/m-p/502062
27. https://www.mdpi.com/1424-8220/24/9/2781