Bilko Observability (GCP-native) 2026-06-10

Status

LIVE and Proveo-verified as of 2026-06-10. GCP project **tribal-sign-487920-k0**, region **europe-north1**. MC #103329 (FlowForge implementation) + MC #103331 (Proveo independent verification). Parent MC #103328.

Related: Bilko Sentinel — Tier-0 Self-Healing Agent 2026-06-10

Environment Topology

The naming is deliberately confusing due to legacy reasons — read carefully:

Logical role	Cloud Run services	Cloud SQL instance	URLs	Notes
PROD (customer trial)	`bilko-api-demo`, `bilko-web-demo`	`bilko-demo-db`	app.bilko.cloud / bilko-demo.alai.no	Named "-demo" for legacy reasons. This is the **functionally live surface** — real customer self-serve trial traffic.
STAGE (internal CI/E2E)	`bilko-api-stage`, `bilko-web-stage`	`bilko-staging-db`	bilko-*-stage.run.app	Internal only. Used for CI validation and E2E test runs. Not customer-facing.
Reserved shell (dormant)	`bilko-web` (rev 00001)	`bilko-db`	N/A	Dormant. Excluded from all alerting. Do not SLO-bind until activated.

Uptime Checks (4 active)

#	Display name	Host / Path	Period	Regions	Env

1	Bilko Web Prod (app.bilko.cloud)	app.bilko.cloud /	60s	EUROPE, USA_VIRGINIA, ASIA_PACIFIC	prod
2	Bilko API Prod (app-api.bilko.cloud/api/v1/health)	app-api.bilko.cloud /api/v1/health	60s	EUROPE, USA_VIRGINIA, ASIA_PACIFIC	prod
3	Bilko Web Stage	bilko-web-stage-dh4m46blja-lz.a.run.app /	300s	EUROPE, USA_VIRGINIA, ASIA_PACIFIC	stage
4	Bilko API Stage	bilko-api-stage-dh4m46blja-lz.a.run.app /api/v1/health	300s	EUROPE, USA_VIRGINIA, ASIA_PACIFIC	stage

Note on API health check: `app-api.bilko.cloud` (Cloudflare proxy) returns HTTP 405 on GET — this is expected. The actual Cloud Run service returns 200. The uptime check accepts both 200 and 405.

Alert Policies (7 active, MC #103329)

Policy name	Services / instances	Threshold	Policy ID
Bilko Prod — HTTP 5xx rate high on bilko-api-demo	bilko-api-demo	>1% 5xx rate over 5-min window (ALIGN_RATE, REDUCE_SUM)	11502345168057990272
Bilko Prod — HTTP 5xx rate high on bilko-web-demo	bilko-web-demo	>1% 5xx rate over 5-min window	13840551641108771864
Bilko Prod — Request latency P95 high on prod services	bilko-api-demo, bilko-web-demo	API P95 >3000ms; Web P95 >5000ms	13840551641108772022
Bilko Prod — Container restart/crash on prod services	bilko-api-demo, bilko-web-demo	starting-state instance count MEAN >3 in 5-min window (crash-loop indicator)	10038710534975650645
Bilko — Cloud SQL CPU utilization high (prod + stage)	bilko-demo-db, bilko-staging-db	bilko-demo-db >70% CPU for 5min; bilko-staging-db >85% for 5min	1002243302492516643
Bilko Prod — Cloud SQL connections near max on bilko-demo-db	bilko-demo-db	num_backends >20 (80% of max 25 for db-f1-micro)	606613461467816964
Bilko Prod — Uptime check failed	app.bilko.cloud, app-api.bilko.cloud	REDUCE_COUNT_FALSE >1 for 120s duration (2+ regions failing)	8433909893104140357

There is also one pre-existing legacy policy from MC #103245: **Bilko CIAM — High 429 rate on bilko-api-demo** (policy ID 4279915624784430014), kept and already had Slack+email attached.

Notification Channels

Channel	Type	GCP channel ID	Attached to
Slack **#ceo** (ALAI workspace T0AELHU0E13)	Slack (GCP-native OAuth)	17620748118296880307	All 7 MC#103329 policies + legacy CIAM policy
alem@alai.no	Email	16578157527237754053	All 7 MC#103329 policies
dev@alai.no	Email (pre-existing)	2103834221134748174	All 7 MC#103329 policies

Dashboard

Display name: **Bilko Observability — Prod + Stage (MC #103329)**
Dashboard ID: `070613fa-a0b6-41e1-8606-ccdf0e52a87a`

Open in GCP Console

Dashboard tiles:

- Prod API Request Rate by response class
- Prod API Latency P50/P95/P99
- Prod Container Instance Count
- Prod DB CPU Utilization (bilko-demo-db)
- Prod DB Active Connections
- Uptime Check Pass Rate (prod web + api)
- Stage API Request Rate
- Stage API Latency P95
- Stage DB CPU Utilization (bilko-staging-db)

Proveo Verification (End-to-End Alert Delivery)

Proveo (MC #103331) ran an independent end-to-end proof:

- Created a temporary uptime probe pointing at a non-existent URL guaranteed to return 404
- GCP confirmed REDUCE_COUNT_FALSE=3 (threshold breached) within ~90 seconds

- Slack #ceo received a native GCP alert message; incident ID `0.o8uwptg3xflh`, channel type confirmed as `channelType=slack`
- Email delivery structurally proven: GCP fires all attached channels from the same alert event; email channel is `enabled: true` and correctly attached
- Both test artifacts (probe + policy) deleted after verification; zero regression on prod services

Verdict: PASS.

IAM Note

No new IAM bindings were created. All setup used `gcloud monitoring` commands only. The existing `alai-cli-deployer` service account already held Monitoring Admin role.

Tuning and Maintenance

Adding or modifying an alert policy

```
# List all policies
gcloud monitoring policies list --project=tribal-sign-487920-k0

# Describe a specific policy (by ID)
gcloud monitoring policies describe POLICY_ID --project=tribal-sign-487920-k0

# Update a threshold (edit JSON/YAML and update)
gcloud monitoring policies update POLICY_ID --policy-from-file=policy.json --project=tribal-sign-487920-k0

# Create a new policy from file
gcloud beta monitoring policies create --policy-from-file=new-policy.json --project=tribal-sign-487920-k0
```

Known threshold that may need raising

The `bilko-demo-db` SQL connections threshold (20/25) was set at 80% of the `db-f1-micro` max_connections=25. After a few weeks of baseline data, consider whether to raise the instance tier (which raises max_connections) or adjust this threshold. Check current connection count:

```
gcloud monitoring time-series list \
  --filter='metric.type="cloudsql.googleapis.com/database/postgresql/num_backends" AND
resource.labels.database_id:"bilko-demo-db"' \
  --project=tribal-sign-487920-k0 \
  --freshness=5m
```

Supersedes

This page supersedes `docs/infrastructure/MONITORING.md` v1.0 (2026-02-25), which described the Railway/Vercel/Express era with PLANNED Sentry/BetterStack. That file has been updated with a superseded header pointing here. See also Bilko Sentinel — Tier-0 Self-Healing Agent 2026-06-10 for the detection and diagnosis agent built on top of this observability layer. Discussion note: `docs/infrastructure/OBSERVABILITY-DISCUSSION-2026-06-09.md`.

Revision #1
Created 2026-06-10 02:29:40 UTC by John
Updated 2026-06-10 02:29:40 UTC by John