Sistema de inferencia: costo actual vs. sistema nuevo

De una infraestructura estática de GPU dedicada a una plataforma dinámica que escala a cero

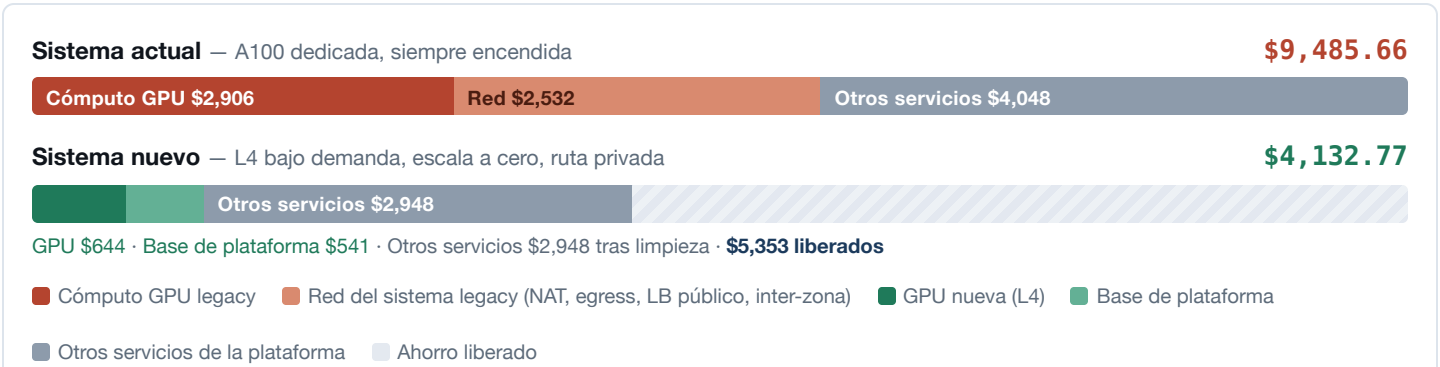
Vigencia: 1 de octubre de 2026
Elaborado 28-09-2026
Cifras mensuales en USD
Fuente: Cloud Billing (ago–sep 2026)

COSTO ACTUAL / MES
$9,486
Inferencia legacy $5,438 + otros servicios $4,048

COSTO NUEVO / MES
$4,133
Inferencia nueva $1,185 + otros servicios $2,948 tras limpieza

AHORRO MENSUAL
$5,353
Migración $4,253 + limpieza $1,100

Comparación del gasto mensual



Dónde se produce el ahorro

COMPONENTE	ACTUAL / MES	NUEVO / MES	AHORRO / MES	REDUCCIÓN
Cómputo GPU (A100 → L4 bajo demanda)	2,905.78	643.68	2,262.10	−77.8%
Base de plataforma de inferencia	—	541.02	−541.02	nuevo
Red: NAT, egress, balanceador público e inter-zona	2,531.80	0.00	2,531.80	−100%
Subtotal sistema de inferencia	**5,437.58**	**1,184.70**	**4,252.88**	**−78.2%**
Otros servicios de la plataforma — limpieza de recursos sin uso	4,048.07	2,948.07	1,100.00	−27.2%
Total	**9,485.66**	**4,132.77**	**5,352.88**	**−56.4%**

El costo de red desaparece porque el tráfico de inferencia deja de entrar y salir por el dominio público (balanceador, egress a internet y Cloud NAT) y pasa por la ruta privada interna entre `lucro-recognition-system` y `hub-gpu-prod`, en la misma región y zona: tráfico sin cargo.

$5,353 **de ahorro mensual desde el 1 de octubre de 2026,** en dos frentes independientes.

MIGRACIÓN DE INFERENCIA
$4,253
$5,438 → $1,185 · −78%

LIMPIEZA DE PLATAFORMA
$1,100
recursos sin uso · −27% de otros servicios

PROYECCIÓN DEL AHORRO ACUMULADO

OCT 2026	OCT – DIC 2026	OCT 26 – MAR 27	12 MESES
$5,353	$16,059	$32,117	$64,235
1 mes	3 meses	6 meses	año completo

Detalle por SKU — sistema actual

Facturación real de `lucro-recognition-system`, normalizada a un mes de 730 horas

Ventana facturada
630.7 h de cobertura continua
Factor de normalización 1.1574

Sistema de inferencia actual (legacy)

SKU	CANTIDAD/MES	UNIDAD	$/UNIDAD	USD/MES
CÓMPUTO GPU — A100 DEDICADA				
Nvidia Tesla A100 GPU running in Americas	730.84	hora	2.93391	2,144.22
A2 Instance Core running in Americas	8,770	hora	0.03161	277.23
A2 Instance Ram running in Americas	62,092	GiB·hora	0.00424	263.08
Autopilot A100 40GB Premium	730,651	unidad	0.0002295	167.65
Autopilot Accelerator CPU Premium	8,767,809	unidad	0.0000020	17.24
Nvidia Tesla A100 GPU en VM Spot/Preemptible	9.82	hora	1.69385	16.63
Autopilot Accelerator Memory Premium	60,988	unidad	0.0002294	13.99
Spot Preemptible A2 Instance Core running in Americas	117.86	hora	0.01825	2.15
Spot Preemptible A2 Instance Ram running in Americas	834.89	GiB·hora	0.00245	2.04
Autopilot A100 40GB Spot Premium	9,366	hora	0.0001400	1.31
Autopilot Accelerator Spot CPU Premium	112,392	hora	0.0000011	0.12
Autopilot Accelerator Spot Memory Premium	781.78	GiB·hora	0.0001300	0.10
Subtotal cómputo GPU				**2,905.78**
RED — SALIDA POR DOMINIO PÚBLICO				
Network Internet Data Transfer Out (Americas→Americas)	15,025	GiB	0.10444	1,569.18
Networking Cloud NAT Data Processing	15,171	GiB	0.04500	682.72
Network Inter Zone Data Transfer Out	13,461	GiB	0.01000	134.61
Global External ALB Inbound Data Processing (Iowa)	14,742	GiB	0.00800	117.93
Cloud LB Forwarding Rule Minimum Global	730.00	hora	0.02500	18.25
Networking Cloud NAT Gateway Uptime	3,898	hora	0.00140	5.46
Networking Cloud NAT IP Usage	730.00	hora	0.00500	3.65
Subtotal red				**2,531.80**
Subtotal sistema de inferencia legacy				**5,437.58**
Otros servicios de la plataforma				4,048.07
Total sistema actual				**9,485.66**

POR QUÉ COSTABA LO QUE COSTABA

- **GPU siempre encendida.** Un nodo A100 dedicado facturado prácticamente 24/7, más un nodo de pruebas, se use o no.
- **Se factura el nodo completo.** Autopilot con clase Accelerator cobra el nodo entero, no lo que pide el pod.
- **Salida por internet.** El tráfico de inferencia entraba y salía por el dominio público: 14,742 GiB por el balanceador, 15,025 GiB de egress y 15,171 GiB procesados por Cloud NAT. En total $2,532/mes.

QUÉ CAMBIA CON EL SISTEMA NUEVO

- **Escala a cero nodos.** Los pools de GPU arrancan en 0 nodos y se apagan al quedar ociosos; solo producción mantiene 1 GPU encendida 24/7.
- **GPU proporcional a la carga.** L4 a $0.854 por hora-GPU frente a $3.91 efectivos de la A100, con placement dinámico por modelo y hasta 10 nodos bajo demanda en lugar de un único nodo compartido.
- **Ruta privada interna.** El tráfico viaja por VPC peering entre proyectos, en la misma región y zona: sin Cloud NAT y sin egress a internet.

Detalle por SKU — sistema nuevo

Plataforma de inferencia dinámica sobre `hub-gpu-prod`

Escenario modelado
1 GPU en prod 24/7 (730 h)
24 h de pruebas · batch no incluido

Sistema de inferencia nuevo

SKU	CANTIDAD/MES	UNIDAD	$/UNIDAD	USD/MES
COSTO DE GPU — L4 BAJO DEMANDA				
Nvidia L4 GPU running in Americas — prod	730	hora	0.56004	408.83
G2 Instance Core running in Americas — prod	5,841	hora	0.02499	145.96
G2 Instance Ram running in Americas — prod	23,365	GiB·hora	0.00293	68.40
Nvidia L4 GPU running in Americas — test	24	hora	0.56004	13.44
G2 Instance Core running in Americas — test	192	hora	0.02499	4.80
G2 Instance Ram running in Americas — test	768	GiB·hora	0.00293	2.25
Subtotal GPU	**754**	**horas GPU**	**0.85370**	**643.68**
BASE DE PLATAFORMA				
Hyperdisk ML Throughput in Iowa	1,600.00	MiBps·mes	0.120000	192.00
Balanced PD Capacity	833.81	GiB·mes	0.100000	83.38
Regional Kubernetes Clusters	730	hora	0.100000	73.00
E2 Instance Core	3,715.22	hora	0.014370	53.39
Prometheus Samples Ingested	564.01	millón de muestras	0.060000	33.84
E2 Instance Ram	14,851.67	GiB·hora	0.001926	28.61
Cloud LB Forwarding Rule Minimum for Americas	730	hora	0.025000	18.24
Cloud LB Forwarding Rule Minimum Global	730	hora	0.025000	18.24
Hyperdisk ML Capacity in Iowa	150	GiB·mes	0.080000	12.00
Storage, snapshots, Cloud Build y ~53 SKUs menores	—	unidades mixtas	—	28.32
Subtotal base de plataforma				**541.02**
Total sistema de inferencia nuevo				**1,184.70**

Otros servicios de la plataforma — $4,048/mes · ajenos a inferencia, idénticos en ambos escenarios

SERVICIO	USD/MES	% DEL BLOQUE
AlloyDB	1,135.89	28.1%
BigQuery	922.33	22.8%
Kubernetes Engine	892.43	22.0%
Cloud Storage	345.98	8.5%
Cloud SQL	235.07	5.8%
Compute Engine, Cloud Logging, Artifact Registry, Dataplex y otros	516.38	12.8%
Total otros servicios hoy	**4,048.07**	**100%**
Limpieza identificada — recursos sin uso a retirar	−1,100.00	−27.2%
Otros servicios después de la limpieza	**2,948.07**	**72.8%**

Método y supuestos. Cifras derivadas de la facturación real de Google Cloud para `lucro-recognition-system` (ventana de 630.7 h de cobertura continua, ago–sep 2026) y `hub-gpu-prod`, normalizadas a un mes de 730 horas (factor 1.1574). El escenario del sistema nuevo modela 1 GPU en producción 24/7 y 24 h de pruebas, sin carga batch. El bloque de red del sistema legacy incluye Cloud NAT, egress a internet, el balanceador público y el tráfico inter-zona: todo tráfico de inferencia que deja de existir al pasar a la ruta privada. La limpieza de plataforma es una oportunidad identificada sobre otros servicios, independiente de la migración. Workers y gateway no migran este mes.