

Public models bill pay-per-prediction on shared GPUs (active time only — setup/idle free) via hardware-second or per-image/token/video meters; fine-tunes train on multi-GPU clocks (e.g. FLUX ~\$1.46 typical); private models and Deployments bill setup+idle+active on dedicated instances; multi-GPU/H200 need committed spend; Enterprise adds volume discounts, SLAs, and reserved compute — all without you managing CUDA or queues.

KEY PARTNERS

- **Cloudflare** — joining / parent narrative Nov–Dec 2025 (FACT blogs).
- **Catalog publishers** — Black Forest Labs, Ideogram, Recraft, DeepSeek, Wan (FACT pricing).
- Community model creators contributing thousands of models (FACT homepage/pricing).
- Cog open-source packaging for custom models (FACT homepage).
- NVIDIA GPU SKUs in public hardware table (FACT pricing).
- Enterprise “trusted by” customer logos (FACT marketing claim).

KEY ACTIVITIES

- Serve public predictions on shared hardware pools.
- Host official + community model catalog with cost estimates.
- Run fine-tuning / training jobs (e.g. FLUX).
- Deploy custom Cog models; autoscale private traffic.
- Operate Deployments (min/max instances, stable endpoints).
- Publish pricing/docs; sell Enterprise volume & SLAs.

KEY RESOURCES

- **GPU SKUs** — T4, L40S, A100, H100, H200 (+ multi-GPU) (FACT pricing).
- 50,000+ model catalog claim (FACT enterprise).
- Cog open-source packaging + generated API servers (FACT homepage).
- Public vs private vs Deployment billing modes (FACT billing).
- Client libraries / HTTP API / docs (FACT docs).
- Cloudflare Developer Platform adjacency (FACT blogs).

VALUE PROPOSITIONS

- One-line API / web Run for production-ready models (FACT homepage).
- **Public** — pay active time only; setup/idle free (FACT billing).
- Transparent hardware \$/sec and per-output estimates (FACT pricing).
- Fine-tune without owning multi-GPU clusters (FACT docs).
- Cog custom models with autoscale (FACT homepage/pricing).
- Deployments for stable endpoints and cold-boot control (FACT billing).
- **Enterprise** — one contract, 50k+ models, SLAs (FACT enterprise).

CUSTOMER RELATIONSHIPS

- Self-serve signup + public pricing cards (FACT site).
- Model pages show run-time and cost estimates (FACT docs).
- **Docs-led expansion** — public → Deployments when steady (FACT billing).
- Prepaid credit or monthly arrears billing (FACT billing).
- **Enterprise** — dedicated AM, priority support, SLAs (FACT enterprise).
- Community model publishing as relationship flywheel (FACT homepage).

CHANNELS

- **Primary** — replicate.com model explore + Run UI.
- Public /pricing hardware and output examples.
- docs.replicate.com quickstarts and billing topics.
- Client libraries / HTTP API / OpenAPI.
- Cog open-source + GitHub Actions publish guides.
- Cloudflare blog/dev platform adjacency; Enterprise form.

CUSTOMER SEGMENTS

- App/creative builders on public model APIs (FACT homepage ICP).
- Products metering images/video/tokens via official models (FACT pricing).
- Teams fine-tuning bases like FLUX (FACT docs).
- Engineers deploying custom Cog models (FACT homepage).
- Prod apps needing Deployments / no shared queue (FACT billing).
- **Enterprises** — one contract, SLAs, reserved compute (FACT enterprise).

COST STRUCTURE

- GPU / power / colo COGS behind predictions (INFERENCE).
- Absorbed setup/idle on public shared pools (INFERENCE from billing).
- Dedicated capacity for private models & Deployments (INFERENCE).
- Platform/Cog/serving R&D (INFERENCE from product shape).
- Catalog and community operations (INFERENCE).
- **GTM** — PLG content + Enterprise sales (INFERENCE).
- No public Replicate P&L line invented.

REVENUE STREAMS

- **Public predictions** — active GPU-sec or per-output; setup/idle free (FACT).
- **Hardware examples** — T4 \$0.81/hr · A100 \$5.04 · H100 \$5.49 (FACT).
- **Official meters** — FLUX Pro \$0.04/img; Wan 480p \$0.09/s video (FACT).
- **Fine-tune training** — e.g. FLUX 8×H100 \$0.0122/sec (FACT docs).
- **Private models** — setup + idle + active (fast-boot exception) (FACT).
- **Deployments** — bill online instances; autoscale (FACT billing).
- **Enterprise** — volume discounts, SLAs, reserved compute (FACT).

Sources: replicate.com/blog/replicate-cloudflare · replicate.com/pricing · replicate.com/enterprise · replicate.com/home · replicate.com/docs/topics · replicate.com/docs/get-started
Not affiliated with Replicate.