



Fireworks AI

Serverless Standard lands builders on per-token open-model APIs (\$1 free credits; OpenAI/Anthropic compatible); cached input (~50%) and Batch (~50%) cut cost on repeated and async prompts; Priority and Fast tiers buy reliability and speed; On-Demand sells dedicated GPU-hours (H100/H200 \$8, B200 \$13, B300 \$15, GB300 \$20); Training (LoRA/SFT/DPO \$/1M and GPU-hour RFT) specializes models on the same stack; Reserved Capacity / Enterprise closes guaranteed throughput and BYOC.

BUSINESS MODEL CANVAS

Fireworks AI

Research date: Sep 21, 2026

Evidence-backed · sources in companion brief

KEY PARTNERS

- **NVIDIA** — investor + infrastructure partner narrative (FACT Series D / press).
- **Open-weight labs** — GLM, Kimi, DeepSeek, MiniMax, Qwen, Llama, Nemotron (FACT model library / homepage).
- **Series D leads** — Atreides, Index, TCV; Nvidia, Lightspeed, Evantic, Bessemer, Menlo, others (FACT blog).
- **Cloud surfaces** — Azure Foundry customer stories (FACT testimonials).
- **Named customers** — Cursor, Harvey, Notion, Vercel, Quora, Sourcegraph, Cresta, UiPath (FACT site/blog).
- OpenAI- and Anthropic-compatible API ecosystem for tool/SDK partners (FACT docs).

KEY ACTIVITIES

- Operate serverless multi-model inference (Standard / Priority / Fast).
- Provision On-Demand dedicated deployments (multi-region, custom models).
- Ship Managed Training, Serverless Training API, Dedicated Training API.
- Advance disaggregated serving stack (company inference narrative).
- Publish pricing + docs; Batch, cache, Reserved Throughput.
- Sell Reserved Capacity / enterprise quotas and BYOC.

KEY RESOURCES

- **GPU SKUs** — H100, H200, B200, B300, GB300 (FACT pricing tables).
- Disaggregated inference engine narrative (FACT inference page).
- 250+ models; day-zero launch partner posture (FACT inference FAQ).
- **Train → serve loop** — fine-tunes serve on same kernels (FACT inference).
- SOC 2 Type II / HIPAA-ready / GDPR claims (FACT FAQ).
- Named production customer relationships (FACT site/blog).

VALUE PROPOSITIONS

- **Serverless** — popular open models, no cold starts, pay per token (FACT docs).
- **OpenAI- and Anthropic-compatible APIs** — migrate by swapping a URL (FACT).
- Cached input ~50%; Batch ~50% on Standard rates (FACT docs).
- Priority + Fast serverless modes for reliability/speed (FACT docs).
- On-Demand dedicated GPUs for custom/LoRA and predictable throughput (FACT).
- Train and serve on one stack (SFT/DPO/RFT + Serverless Training API) (FACT).
- **Company claim** — specialized models dominate token mix; cost narrative vs closed APIs (blog).

CUSTOMER RELATIONSHIPS

- Self-serve signup + \$1 free credits + public pricing (FACT site).
- **Docs-led expansion** — serverless → on-demand for LoRA/custom (FACT docs).
- Batch/cache education to keep cost competitive (FACT docs).
- Priority/Fast as reliability and speed upsells (FACT docs).
- Reserved Capacity / Throughput + Contact sales (FACT inference).
- Reference customers named on site and Series D post (FACT claim).

CHANNELS

- **Primary** — fireworks.ai + /pricing public cards.
- docs.fireworks.ai quickstarts and serverless pricing tables.
- OpenAI- and Anthropic-compatible API / SDK discovery.
- **Company blog** — funding and specialized-intelligence narrative.
- NVIDIA + Azure Foundry partner/customer surfaces.
- Contact sales for Reserved Capacity, BYOC, enterprise quotas.

CUSTOMER SEGMENTS

- App/agent builders on serverless open models (FACT docs ICP).
- **AI-native production companies (named site logos** — FACT claim).
- Teams needing Priority/Fast or dedicated/custom endpoints (FACT).
- ML orgs training specialized models (SFT/DPO/RFT) (FACT pricing).
- Residency/enterprise buyers (US serverless, reserved, BYOC) (FACT).
- Coding/legal/agent products attaching train → serve (FACT testimonials).

COST STRUCTURE

- GPU / power / colo COGS behind serverless + on-demand (INFERENCE).
- Peak-capacity cost behind Priority/Fast and US premiums (INFERENCE).
- Serving-stack and kernel R&D (company inference narrative).
- Training compute COGS for SFT/DPO/RFT (INFERENCE from pricing).
- Series D-funded capacity + eng expansion (FACT blog claim).
- **GTM** — PLG content + enterprise sales (INFERENCE).
- No public Fireworks AI P&L line invented.

REVENUE STREAMS

- **Serverless Standard** — PAYG tokens; \$1 free credits (FACT pricing).
- **Examples** — GLM 5.2 \$1.40/\$0.14/\$4.40; Kimi K2.7 Code \$0.95/\$0.19/\$4.00; MiniMax M3 \$0.30/\$0.06/\$1.20 (FACT docs).
- Cached input discounts; Batch 50% of Standard (FACT docs).
- Priority / Fast premium token cards (FACT docs).
- **On-Demand** — H100/H200 \$8/hr; B200 \$13; B300 \$15; GB300 \$20 (FACT).
- **Training** — LoRA/full SFT & DPO \$/1M bands; RFT = GPU-hr; Serverless Training API (FACT).
- **Reserved / Enterprise** — Contact sales for guaranteed capacity (FACT).

Sources: fireworks.ai/blog/series-d-announcement · fireworks.ai/inference · fireworks.ai/pricing · docs.fireworks.ai/serverless/overview · docs.fireworks.ai/serverless/pricing
Not affiliated with Fireworks AI.