

Free GroqCloud Console lands builders on LPU inference; Developer pay-per-token APIs convert usage into recurring revenue; Enterprise adds custom models, regional endpoints, LoRA, and dedicated capacity; GroqMetal/GroqRack and infra partners extend private and sovereign deployments.

KEY PARTNERS

- **Open-weight model labs hosted on GroqCloud (Llama, GPT-OSS, Qwen, Whisper — docs).**
- NVIDIA non-exclusive licensing / LPX platform path (company news Jun 2026).
- Data-center operators across North America, Europe, MEA, APAC (13 DCs).
- Bell Canada exclusive inference for Bell AI Fabric (May 2025 newsroom).
- **Infra investors and growth capital (Disruptive, Infinitum — \$650M claim).**
- Developer ecosystem around OpenAI-compatible API and playground.

KEY ACTIVITIES

- Operate LPU / LPX inference clusters in 13 data centers.
- Run GroqCloud Console, OpenAI-compatible API, playground.
- **Host production models —** GPT-OSS, Whisper, Qwen, Llama enterprise, Compound.
- **Ship Developer features —** Flex, Batch, spend limits, prompt caching.
- **Deliver Enterprise —** custom models, regional endpoints, LoRA inference.
- **Expand capacity (company —** toward 200 MW by end of 2027).

KEY RESOURCES

- LPU / LPX inference accelerators and compiler stack.
- **GroqCloud —** public, private, co-cloud instances.
- 13 data centers · claimed >5M developers (Jun 2026 newsroom).
- Production model catalog + Compound tool-using systems.
- **Enterprise controls —** LoRA, SSO/SCIM, audit logs, regional endpoints.
- \$650M growth capital (company news Jun 2026).

VALUE PROPOSITIONS

- LPU inference speed + predictable cost-per-token.
- Free Console land → Developer pay-per-token PLG.
- OpenAI-compatible API for easy migration.
- **Hosted open weights —** GPT-OSS, Whisper, Qwen, Llama enterprise.
- **Enterprise —** custom models, LoRA, regional endpoints, SSO.
- GroqRack / private / co-cloud for sovereignty and air-gap.

CUSTOMER RELATIONSHIPS

- Self-serve Free signup at console.groq.com.
- **Developer upgrade —** payment method + chat support.
- Usage dashboard and spend limits for cost control.
- Docs, models list, playground for builders.
- **Enterprise —** dedicated support, LoRA access, SSO/SCIM.
- **Expansion path —** Free → Developer → Enterprise / Rack.

CHANNELS

- **Primary —** groq.com + console.groq.com Free signup.
- Developer docs and OpenAI-compatible API endpoint.
- Playground / demos for speed proof.
- Direct Enterprise sales for dedicated / LoRA / Rack.
- Partner programs (e.g. Bell Canada AI Fabric).
- Company newsroom and developer community amplification.

CUSTOMER SEGMENTS

- Indie builders and students on Free Console.
- Startups / AI-native teams on Developer PAYG.
- Product orgs needing OpenAI-compatible fast inference.
- **Enterprises —** custom models, LoRA, SSO/SCIM.
- Regulated / air-gapped buyers on GroqRack.
- Sovereign / telco partners (e.g. Bell Canada).

COST STRUCTURE

- LPU / LPX hardware + power / colocation COGS.
- Free-tier inference as demand-gen cost.
- **R&D —** compiler, GroqCloud, Compound, Enterprise features.
- **Sales & marketing —** PLG content + enterprise field.
- Support / success for Developer chat and Enterprise CSM.
- **Compliance —** SOC 2, GDPR, HIPAA, private tenancy overhead.

REVENUE STREAMS

- **Free plan —** \$0 demand gen (FACT — plans).
- **Developer PAYG tokens —** card/bank required (FACT — billing).
- **GPT-OSS 20B \$0.075/\$0.30 · 120B \$0.15/\$0.60 per 1M (FACT — docs).**
- Whisper Turbo \$0.04/hr · Whisper V3 \$0.111/hr (FACT).
- **Enterprise contracts —** custom models, LoRA, SSO (sales).
- GroqMetal / GroqRack dedicated and on-prem capacity.
- Partner / sovereign deals (e.g. Bell Canada).