



WHITE PAPER

Economics of the Rafay Token Factory

From GPU-hour monetization to token-metered AI services, measured on live production endpoints.

WRITTEN FOR: GPU cloud operators, neoclouds, telcos, and sovereign AI cloud providers evaluating how to monetize a GPU fleet beyond hourly rental.

August 2026

EXECUTIVE SUMMARY

The same fleet, monetized a second way

GPU-as-a-Service earns revenue by renting infrastructure time. Token Factory shifts the billing unit to what the infrastructure actually produces: tokens, requests, and governed workflows. That shift changes who the operator can sell to and how much a GPU can earn, without adding a single card to the fleet.

Rafay measured this on live production Token Factory endpoints. Across the tested open-model configurations, token-metered serving supported roughly **2x to 4.8x the annual revenue per GPU** of hourly rental at the most competitive open-model market prices, while opening the fleet to enterprises and developers that want governed model access delivered as a service rather than raw compute to manage themselves.

This paper makes the operator's case with measured data. It sets out what a Rafay-managed deployment serves and at what reliability, what that earns against GPU rental, what tenants consume, what end customers save on fit-for-purpose workloads, and how to validate every assumption against your own numbers. The full benchmark detail sits in the appendix so the commercial argument stays in view.

EXECUTIVE TAKEAWAY

Token Factory is an additive monetization layer on top of GPU-as-a-Service. Rafay's testing on a live production deployment sustained roughly **20M tokens per minute at zero errors** and sub-second time to first token, scaling linearly with replicas. At the most competitive open-model market prices this supports roughly **2x to 4.8x the annual revenue per GPU** of hourly rental, needing only about **14% token-active utilization** to match the rental baseline. These figures compare revenue rather than guarantee margin. Realized token price, workload mix, serving cost, and sustained tenant demand determine actual margin.

2–4.8x

Annual revenue per GPU versus hourly rental, at competitive open-model prices

~14%

Token-active utilization needed to match the on-demand rental baseline

20M/min

Sustained tokens at zero errors on the production endpoint

Three things to take from this paper

01

The same fleet earns more

Token metering raises annual revenue per GPU to roughly 2x to 4.8x that of hourly rental at competitive open-model prices, with no additional hardware. Catalog tiering, workload mix, and serving configuration move that figure upward from there.

02

The market gets bigger

GPU rental sells to buyers willing to run their own stack. Token-metered endpoints reach the enterprises, developers, and business units that want governed model access delivered as a service.

03

The downside is contained

Roughly 14% token-active utilization matches the rental baseline. Because Token Factory runs alongside rental on one control plane, committed rental revenue stays intact while the ceiling rises.

SECTION 01

Selling tokens changes the buyer, the billing unit, and the upside

A growing share of the customers Rafay serves do not want to manage clusters, runtimes, model weights, or billing logic. They want governed model access through APIs. A token-metered service reaches those buyers, and it lets better hardware translate directly into more billable output instead of a fixed hourly rate.

EXHIBIT 1

Dimension	GPU-as-a-Service	Token Factory / Token-as-a-Service
Billing unit	GPU-hour, node-hour, storage	Tokens, requests, workflows
Buyer experience	Customer manages infrastructure	Customer consumes model and API output
Main buyer	AI labs, startups, platform teams	Enterprises, developers, internal business units
Revenue driver	Hours rented × hourly rate	Tokens served × realized token price
Upside from better GPUs	Limited unless the hourly rate rises	Higher throughput becomes more billable output
Operator role	Infrastructure provider	AI service provider with billing, governance, SLAs

Exhibit 1: What changes commercially when the billing unit moves from GPU-hours to tokens.

SECTION 02

Revenue ceiling and break-even economics

Pricing the measured volume at published open-model market rates gives a defensible revenue floor. The rates used here, **\$0.14 per 1M input tokens** and **\$0.90 per 1M output tokens**, sit in the lower-middle of the open-model range. The GPU rental baseline uses RunPod's published H200 on-demand rate of **\$4.39 per hour**, with on-demand fleets realizing around 70% billed utilization and reserved fleets billing 100% of contracted hours.

EXHIBIT 2

Scenario	Revenue per GPU-hour	Annual revenue per GPU at 100% active	Annual revenue at realistic utilization
GPUaaS H200 on-demand rental	\$4.39	\$38,456	\$26,920 (70% billed)
GPUaaS H200 reserved	\$4.39	\$38,456	\$38,456 (100% billed)
Token serving, moderate concurrency (~38M tok/hr)	~\$10.10	~\$88,500	~\$53,100 (60% active)
Token serving, high concurrency (~84M tok/hr)	~\$22.40	~\$196,000	~\$117,500 (60% active)
Token serving, production endpoint (~123M tok/hr per replica)	~\$24.60	~\$215,000	~\$129,100 (60% active)

Exhibit 2: Measured revenue at open-model market prices versus GPU-hour rental baselines. The first four rows are per H200 GPU. The production endpoint row is measured per serving replica, and replica-to-GPU mapping is deployment-specific.

At realistic utilization, the measured production endpoint lands at roughly **4.8x the on-demand rental baseline** and 3.4x the reserved baseline. The lower-concurrency H200 configurations land between 2.0x and 4.4x. These figures behave as a floor because they price an efficient mixture-of-experts model at the most competitive public rates. Operators blending premium model SKUs, enterprise service tiers, or reserved endpoints move up from here. The full open-model price spread that supports a tiered catalog is in Appendix B.

Break-even: the demand needed to beat rental

Holding the measured 84M tokens per hour constant, the following shows how little token-active utilization it takes to match the on-demand rental baseline of roughly \$26,900 per GPU per year.

EXHIBIT 3

Blended realized price per 1M tokens	Revenue potential at 100% active	Token-active utilization needed to match rental
\$0.27 (measured market floor)	\$196,000 / GPU / year	13.7%
\$0.50	\$368,000 / GPU / year	7.3%
\$1.00	\$736,000 / GPU / year	3.7%

Exhibit 3: Break-even token-active utilization at three blended price points (H200, measured 84M tokens/hr). On the higher-throughput production endpoint, the market-floor threshold falls to roughly 12.5%.

WHAT THIS MEANS

At the most competitive open-model market pricing, roughly 14% token-active utilization matches the on-demand H200 rental baseline. At a \$0.50 blended price the threshold falls to about 7%. These are revenue break-even numbers rather than margin break-even. Margin still depends on serving efficiency, model mix, support, power, platform cost, and demand consistency. Section 08 provides the validation checklist.

SECTION 04

Designing the catalog for volume and value

Capacity per GPU becomes a pricing model only when it is paired with demand. Token demand does not distribute evenly. Across established token-metered providers, a small number of models drives the majority of volume, and high-frequency utility workloads such as embeddings and retrieval are often the single largest driver, while premium and specialist models carry lower volume at higher value.

That pattern points to a clear design principle. A tiered catalog captures both the volume and the value, and Rafay lets an operator publish, meter, and bill every tier from one control plane. The observed demand data and per-archetype consumption that inform this design are in Appendix C.

EXHIBIT 4

Model category	Role in token demand	Implication for your catalog
Embedding models	Can become the largest token-volume driver	Include embedding endpoints for RAG, search, retrieval, and enterprise knowledge workflows
Small instruct models	Steady, low-cost inference	Developer tiers, internal tools, high-frequency app workloads
Mid-sized models	Balanced production usage	Better quality than small models at lower serving cost than 70B-class
70B+ / high-parameter	Lower volume, higher value	Premium SKUs with SLA, reserved access, or custom pricing
Long-tail models	Experimentation and customer-specific use	Keeps customers on-platform through model optionality

Exhibit 4: Catalog design by model category, informed by observed provider demand patterns.

COMMERCIAL TAKEAWAY

The winning catalog is built around both volume capture and value capture: high-volume embedding models, efficient small models, mid-sized production models, premium larger models, and long-tail specialized endpoints. Track usage by model, tenant, endpoint, and time window so you know which models drive volume, which drive margin, and which SKUs deserve reserved capacity.

SECTION 05

Additive to GPU rental

Token Factory does not ask an operator to abandon GPU rental. It monetizes the same fleet a second way, and the two run together on one control plane. Rental delivers committed revenue on contracted hours. Token Factory raises the ceiling on hours that would otherwise bill at cost.

EXHIBIT 5

Dimension	GPU rental as a service	Token Factory on the same fleet
Revenue per H200 / year	\$26,900 on-demand at 70% billed; \$38,500 reserved at 100% billed	\$53,000 to \$129,000 at measured throughput and 60% token-active utilization
Revenue character	Committed and predictable	Consumption-based and expanding
Upside from better hardware	Limited unless hourly rates rise	Direct. Higher throughput becomes more billable output
Buyers reached	AI labs, startups, platform teams that manage their own stack	Enterprises, developers, and business units that want governed endpoints
What Rafay provides	Bare metal and VM provisioning, cluster lifecycle, tenancy	Model catalog, token metering, quotas, chargeback, SLAs, autoscaling

Exhibit 5: What each monetization model contributes on the same Rafay-managed fleet.

One pool, one control plane

Most operators lead with enterprise GPU rental, because it is the faster path to committed revenue and the more familiar sale. Token Factory then layers on top: first for enterprise tenants that want governed model endpoints without running inference themselves, then for broader developer consumption as the tenant base grows. Reserved rental revenue funds that expansion, and because both run on one Rafay control plane, capacity moves between them without a second platform investment.

Rafay operates the fleet as a single pool allocated by use case rather than two separate estates. Model deployments reserve a baseline, inference services autoscale with request volume, and the cluster scales as aggregate demand moves. Capacity freed by one workload becomes available to another, which keeps blended utilization high enough for both models to pay. Metering and governance at the control plane also unlock pricing patterns that raw hourly rental cannot express, each monetizing a different part of the demand curve.

EXHIBIT 6

Consumption pattern	What the operator sells	Where it fits
Scheduled capacity	Compute available in defined windows, released afterwards	Teaching environments, notebooks, and periodic internal workloads that should not pay around the clock
Batch and deferred inference	Lower-priced throughput executed when utilization is low	Latency-tolerant jobs that fill overnight and off-peak capacity

Consumption pattern	What the operator sells	Where it fits
Burst above baseline	Overflow capacity above a reserved floor	Enterprises with predictable baselines and unpredictable peaks
Reserved token endpoints	Guaranteed endpoint capacity with an SLA	Enterprise tenants requiring committed performance rather than best-effort

Exhibit 6: Consumption patterns Rafay can meter and bill beyond standard hourly rental.

THE PRACTICAL ANSWER

Run both. Reserved GPU rental delivers near-term committed revenue, Token Factory raises the revenue ceiling on the same hardware, and scheduled, batch, and burst patterns monetize capacity neither model would otherwise fill. The requirement in every case is one control plane that provisions the infrastructure and meters, governs, and bills what runs on it. That is what Rafay provides.

SECTION 06

The end-user case: open-model endpoints versus frontier APIs

The operator economics are one side of the coin. The other is the argument the operator's commercial team makes to its own customers: what does a workload cost on a governed Token Factory endpoint compared with consuming frontier APIs directly. For workloads where an open model meets the task, the gap is large.

Priced at published August 2026 list rates, a representative enterprise workload of 1B tokens per month, split 5:1 between input and output, costs at the extremes:

EXHIBIT 7

Consumption path	Representative list price (input / output per 1M)	Monthly cost, 1B tokens
Frontier reasoning and flagship APIs	~\$3.00–\$5.00 / ~\$15–\$30	~\$5,000–\$9,200
Frontier lightweight APIs	~\$1.00 / ~\$5.00	~\$1,700
Token Factory open-model endpoint (efficient mid-sized model, market rate)	\$0.14 / \$0.90	~\$270

Exhibit 7: Monthly cost of a 1B token workload (833M input / 167M output) across provider categories at published list prices, as of August 2026. Frontier ranges are representative of published flagship and lightweight tiers rather than a single vendor SKU, and move frequently; confirm current rates before quoting externally.

On the serving side, that same 1B token monthly workload consumes roughly **12 H200 GPU-hours** at measured high-concurrency throughput. That small footprint is why the operator can price well below frontier list rates and still hold margin, and why an end customer running high-volume utility workloads such as RAG, retrieval, and internal assistants has a strong cost case for consuming through the operator.

CAPACITY OPPORTUNITY COST

Those 12 GPU-hours are the bridge between per-workload and per-GPU economics. At the \$4.39 on-demand rate, they represent about \$53 of rentable capacity. Served as tokens at the market-floor open-model price, the same hours produce roughly \$270 in revenue, about 5x more from identical capacity and broadly in line with the per-GPU revenue multiple in Section 03. This measures revenue against the rental price of the capacity it displaces, and assumes those hours are fully token-active. It is not a margin figure: it excludes serving cost, power, platform, and support, which Section 08 puts to the operator as questions to answer with its own numbers.

READ THIS COMPARISON CAREFULLY

Frontier models and open models are not interchangeable on every task, and frontier providers offer caching and batch discounts that lower effective rates. The comparison holds for workloads where an open model meets the quality bar, which observed market demand shows a large and growing share of enterprise token volume. Position the Token Factory as the cost-efficient path for fit-for-purpose workloads, not as a like-for-like replacement for frontier reasoning.

SECTION 07

Proven in production: a Tier 1 telco AI cloud

The following describes an operational Tier 1 Asia-Pacific telecom deployment running a multi-tier GPU portfolio. It supports 30 or more active tenants across enterprise AI categories, and it is the source of the production benchmarks used throughout this paper.

EXHIBIT 8

Production signal	Why it matters for an operator
30+ active tenants on one sovereign deployment	The platform supports a multi-tenant AI services business
Reserved and on-demand GPU consumption	Committed workloads and variable API consumption run on one platform
Tenants across IT services, research, government, healthcare, financial services, and AI-native startups	Maps to multiple buying motions: per-developer, per-API, sovereign, regulated, and startup model-serving
Models exposed as token-metered endpoints	Model usage becomes billable units: tokens, requests, or API calls
Open-source and partner models on one billing and control plane	A model catalog without forcing each customer to run model infrastructure

Exhibit 8: Production signals from Rafay's published Tier 1 APAC telco deployment.

What the control plane makes billable

Economics only materialize if usage can be packaged, metered, and invoiced. Rafay Token Factory gives the operator four commercial levers over the same infrastructure, each enforced at the platform layer rather than built per customer.

EXHIBIT 9

Pricing lever	Options	Purpose
Unit price	Per 1M input tokens, per 1M output tokens, per request	Convert model usage into billable revenue
Service tier	Developer, Professional, Enterprise, Custom	Segment by quota, SLA, support, and model access
Consumption mode	On-demand, reserved, committed spend	Balance usage flexibility with predictable revenue
Governance layer	Tenant quotas, rate limits, endpoint metering	Enable billing, chargeback, cost control, and auditability

Exhibit 9: The four commercial levers Rafay Token Factory enforces at the platform layer.

A representative production catalog on this deployment spans lightweight embedded endpoints (Llama 3.1 8B-Instruct), general-purpose enterprise APIs (Llama 70B-Instruct), multilingual and regional-language models

(Qwen 3 32B), cost-optimized reasoning (DeepSeek-distill-Llama 70B), and voice transcription for call-center and accessibility workloads (Whisper). Every model is published, metered, and billed from the same control plane.

SECTION 08

Validating the Token Factory model for your environment

The figures in this paper are a floor built on measured production data. The way to turn a revenue model into a margin model is to answer five questions with your own numbers.

EXHIBIT 10

Question to validate	Why it matters
What is your actual realized GPUaaS rate per GPU-hour?	Determines the true baseline to beat
What blended token price can you sustain by model family?	Determines top-line revenue per unit of inference
What is your actual serving cost per 1M tokens?	Turns the revenue model into a gross margin model
Which models produce the best tokens per GPU-hour?	Determines the catalog SKUs to prioritize
What utilization can you sustain across tenants?	Determines whether the revenue ceiling is achievable

Exhibit 10: Five questions that convert the revenue model into a margin model.

THE BOTTOM LINE

Token Factory raises the revenue ceiling on infrastructure an operator already owns. The throughput, reliability, and scaling behavior are measured on a live production deployment, the break-even threshold is low, and the model runs alongside GPU rental rather than replacing it. What makes it commercial is a control plane that provisions the infrastructure and meters, governs, and bills what runs on it. **Rafay works with operators to validate these numbers against their own prices, workload mix, and demand, and to put the platform into production.**

Benchmark methodology and full results

How Rafay measured

Every figure comes from Rafay's performance testing on Qwen3-class models, measured with NVIDIA AIPerf against live Token Factory endpoints. Tests ran as concurrency sweeps in 120-second measurement windows with a 30-second warm-up excluded, using server-reported token counting and a fixed enterprise-representative prompt mix. Requests carried five input tokens for every output token, so billable volume is larger than output volume. A GPU generating 14M output tokens per hour is billing roughly 84M total tokens per hour.

EXHIBIT A1 · MEASURED THROUGHPUT PER GPU (H200)

Configuration	Concurrency	Output tok/s per GPU	Tokens per GPU-hour (in + out)
Qwen3.6-35B-A3B, 2 GPUs	50	~1,754	~38M
Qwen3.6-35B-A3B FP8, 6 GPUs	400	~3,882	~84M
Qwen3-30B-A3B, 4 GPUs	400	~3,517	~76M
Qwen3-30B-A3B, 8 GPUs	800	~3,513	~76M
Qwen3-30B-A3B, 16 GPUs (10K in / 1K out)	1,200	~2,075	~82M

Exhibit A1: Measured throughput on H200 GPUs (5,000 input / 1,000 output token requests except where noted), NVIDIA AIPerf. Throughput moves from roughly 1,750 to 3,900 output tokens per second per GPU as concurrency rises, with near-linear scaling from 4 to 8 GPUs.

EXHIBIT A2 · LINEAR REPLICA SCALING

Deployment configuration	Replicas	Peak sustained throughput	Throughput per replica
Qwen3 27B-class, GPU-resident KV cache	4	~3.0M tok/min	~0.76M tok/min
Qwen3 27B-class, GPU-resident KV cache	8	~6.2M tok/min	~0.77M tok/min
Qwen3 27B-class, GPU-resident KV cache	12	~9.3M tok/min	~0.78M tok/min
Qwen3 30B-A3B, CPU KV cache offload	5	~10.4M tok/min	~2.09M tok/min
Qwen3 30B-A3B, CPU KV cache offload	10	~20.6M tok/min	~2.06M tok/min

Exhibit A2: Peak sustained throughput by replica count, measured on a live Tier 1 telco Token Factory endpoint (NVIDIA AIPerf). Scaling is linear across every configuration. Offloading the KV cache to host memory raised throughput per replica by roughly 2.6x, because cached prompt prefixes survive under load instead of being evicted and recomputed.

EXHIBIT A3 · RELIABILITY AND LATENCY UNDER SUSTAINED LOAD

Concurrent requests	Total throughput	Error rate	TTFT p50	Inter-token latency p50
128	~7.9M tok/min	0.0%	0.21s	13ms
512	~15.1M tok/min	0.0%	0.31s	27ms
1,000	~19.5M tok/min	0.0%	0.41s	37ms
1,500	~20.6M tok/min	0.0%	0.59s	39ms
2,000	~18.6M tok/min	0.3%	0.85s	62ms

Exhibit A3: Throughput, error rate, and latency across a concurrency sweep on the ten-replica production deployment. Throughput plateaus around 20M tokens per minute and rolls off gently rather than collapsing.

Serving capacity at a committed SLA

Peak throughput is not the right capacity planning anchor, because raw throughput keeps rising after per-request latency has already broken the service level a paying tenant expects. Rafay's goodput testing sweeps concurrency across workload profiles on H100 GPUs and identifies the knee where time-to-first-token or inter-token latency SLOs first fail. The result answers the question every operator asks: how many concurrent requests can a given GPU configuration serve at a committed SLA.

EXHIBIT A4 • SUSTAINABLE CONCURRENCY AT COMMITTED SLOS

Workload profile	2× H100 sustainable concurrency	4× H100 sustainable concurrency
Short interactive requests (~1K input)	64	128
Iterative debugging (~2K input)	64	64 to 128
Feature building (~4K input)	16	16 to 32
Document and repository heavy (~8K input)	16	16 to 32
Mixed agentic sessions	16	16 to 32

Exhibit A4: Sustainable concurrent request levels at committed SLOs (time-to-first-token under 2 to 5 seconds, inter-token latency p95 under 50ms), from Rafay goodput testing. Input-heavy workloads hit the prefill bottleneck first, so an 8K-input workload sustains roughly a quarter of the concurrency of short interactive traffic. Catalog pricing, tenant quotas, and SLA tiers should reflect this difference.

APPENDIX B • OPEN-MODEL PRICE SPREAD

Where the market prices open models

Open-model rates vary by more than an order of magnitude. That spread is the point: it gives the operator a catalog strategy rather than one number, and it sets the realistic range for a blended realized price. The rates used in the body of this paper sit deliberately in the lower-middle of this range so the revenue figures behave as a floor. All prices below were verified against provider pricing pages as of August 2026; token prices change frequently, so an operator should confirm current rates before using any figure in a customer-facing quote.

EXHIBIT B1 • PUBLISHED OPEN-MODEL API PRICES BY TIER

Model tier	Example public listing	Price per 1M (in / out)	Blended at 12:1 in:out
Entry / lightweight	GPT-OSS 20B	\$0.05 / \$0.20	~\$0.06
Efficient small	GPT-OSS 120B	\$0.15 / \$0.60	~\$0.19
Mid-sized production	Llama 4 Maverick	\$0.27 / \$0.85	~\$0.32
Mid-sized production	MiniMax M2.7	\$0.30 / \$1.20	~\$0.37
Flagship open	DeepSeek V3.1	\$0.60 / \$1.70	~\$0.68
Premium open	Kimi K2.6	\$1.20 / \$4.50	~\$1.45
Reasoning-class open	DeepSeek R1	\$3.00 / \$7.00	~\$3.31

Exhibit B1: Published open-model API prices across catalog tiers, verified against provider pricing pages as of August 2026. Token prices move frequently; confirm current rates before quoting externally. An operator that stocks a tiered catalog, from lightweight endpoints through premium and reasoning SKUs, raises its blended realized price without changing a single GPU.

APPENDIX C • DEMAND AND CONSUMPTION DATA

What tenants consume and what the market runs

Token consumption by user archetype

Not all users burn tokens at the same rate. A person checking a fact consumes a fraction of what a coding assistant or an automated pipeline consumes. Rafay's benchmark mix was built from four request profiles spanning short

lookups to very large context calls. The token volumes per request are Rafay's tested and server-measured figures. The archetype labels interpret which customer type each profile represents.

EXHIBIT C1 · TOKEN CONSUMPTION BY USER ARCHETYPE

User archetype	Tested request profile (in / out)	Requests per working day	Tokens per user per month
Casual information lookup	350 / 120	15	~0.16M
Productivity and drafting	3,200 / 300	40	~3.1M
Developer and code assistant	8,300 / 420	120	~23.0M
Automated agentic pipeline	23,000 / 580	300	~155.6M

Exhibit C1: Token consumption by user archetype. Request profiles are from Rafay benchmark testing. Archetype mapping, 15 to 300 requests per working day, and 22 working days are planning assumptions.

What the market consumes today

Two established token-metered providers publish per-model usage through OpenRouter, which gives a directional read on where enterprise token volume actually lands. Nebius runs roughly 430B to 500B tokens per 30 days across a 12-model catalog. Together runs roughly 480B to 550B across 24 models. The figures below are estimated from published usage charts and are directional rather than exact.

EXHIBIT C2 · HIGHEST-CONSUMPTION MODELS AT TWO TOKEN-METERED PROVIDERS

Provider	Model	Est. tokens / 30 days	What the volume indicates
Nebius	Qwen3 Embedding 8B	~350B	Embeddings dominate: RAG, semantic search, retrieval, document indexing
Nebius	Llama 3.1 8B Instruct	~22B	Low-cost developer and in-application endpoints
Nebius	Qwen3 32B	~18B	Mid-sized models carrying production workloads
Nebius	Llama 3.3 70B Instruct	~16B	Premium general-purpose enterprise API
Together	MiniMax M2.7	~210B	High-volume general-purpose generation
Together	Qwen3.5-9B	~150B	Efficient endpoints for cost-sensitive production traffic
Together	MiniMax M2.5	~55B	Secondary generative capacity and model choice
Together	Qwen3.5 397B A17B	~14B	Specialist high-capability workloads at lower volume

Exhibit C2: Highest-consumption models at two token-metered providers, estimated from published OpenRouter usage data.

Two patterns are worth designing around. Demand concentrates: at both providers a handful of models drives the large majority of volume. And the highest-volume workloads are enterprise utility rather than premium chat. Nebius shows a single embedding model accounting for most observed usage, which points to retrieval and knowledge-base pipelines as the anchor tenant workload. Together shows the alternative shape, where a broad generative catalog spreads volume across several mid-sized models. Both are viable, and Rafay lets an operator publish, meter, and bill either shape from one catalog.

APPENDIX D · SOURCES AND CALCULATION NOTES

Sources

#	Source	Used for
1	NVIDIA Technical Blog, “Building Token-Metered AI Services on Telco AI Factories,” May 21, 2026	Industry framing for token-metered AI services on telco AI factories
2	Rafay, “How Rafay Turns NeoClouds and Telco AI Clouds into Token-Metered Revenue Engines,” May 21, 2026	Production deployment pattern, 30+ tenant signal, tenant categories, model catalog
3	RunPod GPU cloud pricing page (runpod.io/pricing)	Public H100, H200 (\$4.39/hr baseline), and B200 GPU-hour price checks
4	Lambda AI Cloud pricing page (lambda.ai/pricing)	Public H100 and B200 GPU-hour price sanity checks
5	Together AI pricing page (together.ai/pricing)	Public token API price sanity check for open-model inference
6	OpenRouter State of AI report	100T-token usage study and model/task usage framing
7	Rafay benchmark report: Qwen3.6-35B-A3B-FP8 on vLLM, H100 goodput study (NVIDIA AIPerf, May 2026)	Sustainable concurrency at committed SLOs by workload profile
8	Rafay benchmark report: Qwen3 / Qwen3.6 on H200 GPUs, Tier 1 telco Token Factory endpoint (NVIDIA AIPerf)	Measured output tokens per second per GPU across 2 to 16 GPU deployments
9	OpenAI API pricing documentation (developers.openai.com)	Frontier flagship and lightweight list prices, verified August 2026
10	Anthropic pricing documentation (platform.claude.com)	Frontier flagship and lightweight list prices, verified August 2026
11	Rafay benchmark series: Qwen3 27B-class concurrency ramps at 4, 8, and 12 replicas, live Tier 1 telco endpoint (NVIDIA AIPerf)	Linear replica scaling; effect of GPU-resident KV cache on throughput ceiling
12	Rafay benchmark series: Qwen3 30B-A3B concurrency ramps at 5 and 10 replicas with CPU KV cache offload, same endpoint	Peak sustained throughput, error rates, latency percentiles, prefix cache behavior
13	Together AI and OpenRouter published model pricing pages	Open-model price spread across catalog tiers, verified August 2026
14	OpenRouter provider pages: Nebius Token Factory and Together	Estimated per-model 30-day token volumes and demand concentration signals

Calculation formulas

Metric	Formula
GPUaaS annual revenue	Hourly GPU price × GPU utilization × 8,760 hours/year
Revenue multiple	Token Factory annual revenue ÷ matching GPUaaS baseline
H200 GPUaaS on-demand baseline	$\$4.39/\text{hr} \times 0.70 \times 8,760 = \$26,920$
Tokens processed per GPU-hour	Output tok/s per GPU × 3,600 × 6 (5:1 in:out); e.g. $3,882 \times 3,600 \times 6 \approx 84\text{M}$
Revenue per GPU-hour	(Input tokens × \$0.14/M) + (output tokens × \$0.90/M); e.g. $69.9\text{M} \times \$0.14 + 14.0\text{M} \times \$0.90 \approx \$22.40$
End-user monthly cost (1B tokens)	$833\text{M} \times \text{input price per 1M} + 167\text{M} \times \text{output price per 1M}$

Metric	Formula
Throughput per replica	Peak total tokens/min ÷ replica count; e.g. 20.55M ÷ 10 ≈ 2.06M tok/min ≈ 123M tok/hour
Tokens per user per month	(Input + output tokens per request) × requests per working day × 22 working days
Blended realized price	Total revenue ÷ total tokens served; varies with output share of the workload mix