RELIANCE INTELLIGENCE
THE TOKEN BUSINESS

Commercial & Go-to-Market Strategy

Turning India's Sovereign AI Factory into a Durable Revenue Engine

---

Prepared by Ilan Gleiser
Candidate, Chief Commercial Officer — Token Business
Prepared for the Reliance Intelligence leadership team · August 2026

Companion to the live Commercial Cockpit, P&L Primer and Compute Futures model at reliancecloud.live

Table of Contents

This document is a companion to the live Commercial Cockpit dashboard (reliancecloud.live) built to run the Token Business day to day. It assembles the commercial thesis, the pricing and SKU architecture, the worked financial model, the billing and revenue-assurance operating model, the compute-futures hedging strategy, the organization and 90-day plan, and the current live-dashboard performance snapshot into a single reference for tomorrow's discussion.

Executive Summary

Reliance Intelligence is building one of the largest AI compute platforms in the world: 120 MW of AI compute targeted at Jamnagar by end-2026 on NVIDIA GB300 systems, scaling from roughly 75,000 H100-equivalents on inference today past 200,000, backed by renewable power from Kutch and a ₹10 trillion, seven-year infrastructure commitment. Meta's 168 MW built-to-suit lease at Jamnagar proves the anchor-tenant model already works at this site.

The commercial mandate is simple to state and hard to execute: convert that capacity into contracted, metered, collected revenue faster than it depreciates. GPU infrastructure is a wasting asset with a four-to-six-year useful life — every idle GPU-hour is perishable inventory, exactly like an empty airline seat at departure. This document lays out the plan to do that, organized around five operating principles: sell certainty first (60–70% of capacity under multi-year take-or-pay before optimizing the residual); price in tokens, cost in GPU-hours; treat captive Reliance demand as the launch customer, not the business plan; treat billing as a product, not back-office plumbing; and lean on sovereignty — data residency, Indic-language models, government-grade compliance — as the moat hyperscalers cannot easily match on Indian soil.

Part I sets the strategic context: the opportunity, the five-principle mandate, the competitive landscape, and the two-sided business model (capacity sold by the GPU-hour, tokens sold by the million) that the rest of the plan is built on. Part II lays out the commercial architecture — seven SKUs sequenced across three waves, a four-model pricing system, go-to-market motions for five distinct buyer segments, and the marketplace commercial model. Part III is the financial engine: the cost stack, two fully worked P&Ls (a 1,024-GPU capacity cluster and a token-serving endpoint on the same fleet), a SKU-level ROI model, and the assembled income statement benchmarked against the closest public comparable, CoreWeave. Part IV covers the discipline this plan treats as home ground — metering, billing, settlement and revenue assurance — because 1–3% of revenue leaks in every usage-based business that does not control this chain tightly. Part V introduces compute futures as a financial overlay that hedges the uncommitted tail of capacity, lowers the cost of debt, and stabilizes the whole strategy. Part VI is the organization build, the first-90-days plan, and the KPI and governance rhythm that runs it. Part VII takes the reader inside the Commercial Cockpit itself — the live dashboard this

plan is operationalized in — walking through its seven views and the current scorecard against the mandate. Part VIII closes with why this plan and why this candidate, plus a sourced appendix.

Headline targets

- **60–70% of Wave-1 capacity** under multi-year take-or-pay contracts before the commissioning date, so the depreciation schedule is underwritten by contracted revenue, not spot hope.
- **$100M+ ARR run-rate** exiting year one, with 20+ enterprise logos live on token service, fine-tuning and private endpoints launched, and a 10+ model marketplace beta.
- **Revenue leakage under 0.5%** run through a three-way match control (allocation ledger vs. telemetry vs. billed invoice) borrowed directly from trading-desk break controls.
- **Net revenue retention of 120%+** and a documented, desk-level playbook for every commercial motion by the end of year one, so the organization scales on practice, not heroics.
- **A five-pillar commercial organization** led and staffed to at least 60% ready-now succession depth by month 12, so the plan does not depend on any single individual to keep running.

This document deliberately runs long because the mandate is broad: it is asked to cover the commercial thesis, the pricing and product architecture, the underwriting math, the operating discipline that protects the margin once it is earned, the financial hedge that protects it against a falling price environment, and the organization that runs all of it — in enough depth that each claim can be checked against its own worked assumptions rather than taken on faith. Read end to end, it is a business case; read section by section during tomorrow's discussion, each Part stands on its own with its own numbers and its own recommended actions.

Part I — Strategic Context

1. The Opportunity: Reliance's Sovereign AI Factory

Reliance Intelligence is building one of the largest AI compute platforms in the world. The commercial mandate is to convert that capacity into contracted, metered, collected revenue faster than it depreciates.

The asset base is extraordinary: 120 MW of AI compute at Jamnagar targeted by end-2026 on NVIDIA GB300 systems (75,000+ H100-equivalents on inference today, scaling past 200,000), renewable power from Kutch that structurally lowers cost per token, a nationwide Jio edge network, captive demand across the Reliance group, and a ₹10 trillion seven-year infrastructure commitment. Meta's 168 MW built-to-suit lease proves the anchor-tenant model works at Jamnagar.

GPU infrastructure is a depreciating asset with a four-to-six-year useful life; every idle GPU-hour is perishable inventory, like an empty airline seat. The commercial strategy therefore optimizes one master metric — revenue-weighted utilization — and is built on five principles.

The five operating principles

- **Sell certainty first.** Anchor the P&L in multi-year reserved and committed-capacity contracts (target 60–70% of capacity) before optimizing spot and on-demand token yield on the remainder.
- **Price in tokens, cost in GPU-hours.** Customers buy outcomes (tokens, fine-tunes, agent runs); the platform's cost basis is GPU-hours, power, and network. The pricing architecture must arbitrage that translation profitably, model by model.
- **Captive demand is the launch customer, not the business plan.** Jio, Retail, JioStar, and the five Jio AI services (JioBharatIQ, Vyapar, JioHealthIQ, JioLearnIQ, JioKrishiIQ) de-risk utilization in year one and serve as public reference architecture, but external enterprise, SaaS, and marketplace revenue is what makes the token business a standalone P&L.
- **Billing is a product.** A token business lives or dies on metering integrity, invoice trust, and revenue assurance. The OSS/BSS stack is treated as a first-class product with an SLA, not back-office plumbing.

- **Sovereignty is the moat.** Data residency under the DPDP Act, Indian-language models across 22 languages, and government-grade compliance are differentiators the hyperscalers cannot easily match at Indian price points on Indian soil.

The Jio distribution asset, in detail

Distribution is usually the hardest and slowest thing to build in enterprise cloud sales — most neoclouds spend years assembling an enterprise sales motion from zero. Reliance starts this business with the opposite problem: Jio already has the relationships, the billing infrastructure, and the channel; the task is to point that machine at a new product. Every Jio enterprise, government, and BFSI relationship is a warm door for the token business, and Jio's existing telecom billing rails and channel-partner network are a template — not a blueprint to be copied, but an operating muscle already built — for the OSS/BSS discipline this plan calls for in Part IV. The edge network adds a second, distinct asset: latency-tiered products (real-time voice, vision, agentic workflows) that no centralized cloud, however large, can match on round-trip time from an Indian access point. That is a genuinely defensible product category, not just a cost advantage.

Sovereignty and compliance, in detail

Data residency under India's Digital Personal Data Protection (DPDP) Act, first-class serving of 22 Indian languages, and government-grade compliance are not marketing claims in this plan — they are the two hardest procurement requirements for exactly the buyers with the highest willingness to pay: BFSI, healthcare, and government. A hyperscaler can open an Indian region; it cannot as easily replicate a national telecom's existing regulatory relationships, data-localization track record, or Indic-language model quality built for the country's actual linguistic diversity rather than translated as an afterthought. This is why Part II prices private enterprise endpoints at a 30–50% sovereignty premium and why Segment 3 (Enterprise and BFSI) is called the margin engine of the portfolio: sovereignty is the one advantage on this list that a well-capitalized competitor cannot simply buy its way past in twelve months.

2. Market Context and Right to Win

India's AI compute market is supply-constrained and price-sensitive at the same time. The IndiaAI Mission has anchored subsidized GPU access for startups and academia; hyperscalers (AWS, Azure, GCP) serve the premium enterprise tier from limited Indian regions; domestic providers (Yotta, E2E, Tata Communications, CtrlS, Sify) compete on

price and availability. Demand from Indian enterprises, GCCs, SaaS builders, and government programs is growing faster than any domestic supply response except Reliance's.

Where Reliance wins

- **Scale and cost:** gigawatt-trajectory capacity plus captive renewable power yields a structural cost-per-token advantage over both hyperscalers (imported margin stacks) and subscale domestic players.
- **Distribution:** Jio's enterprise relationships, telecom channel, and 500M+ consumer reach constitute a GTM asset no competitor holds. Every Jio enterprise account is a warm token-business lead.
- **Sovereignty and language:** sovereign hosting plus first-class Indic-language model serving addresses the two hardest requirements of Indian government, BFSI, and healthcare buyers.
- **Edge:** telecom edge inference across the Jio network enables latency-tiered products (real-time voice, vision, agentic workflows) that centralized clouds cannot price competitively in India.

Where discipline is required

- **Price-sensitive demand:** Indian on-demand GPU pricing runs well below US list; chasing spot volume at Indian street prices without committed anchors erodes margin. The answer is contract mix, not price leadership everywhere.
- **Hyperscaler counterattack:** expect aggressive India-region expansion and credits. Defense: sovereignty, edge latency, rupee-denominated committed pricing, and group distribution.
- **Capacity credibility:** selling ahead of commissioned megawatts destroys trust. Commercial commitments must be gated to the delivered-capacity curve agreed with Infrastructure leadership.

Competitive landscape at a glance

Four distinct competitor archetypes shape the India compute market, each with a different value proposition and a different weakness Reliance's asset base and distribution can exploit.

Archetype	Examples	Where they win	Where Reliance wins instead
Hyperscalers	AWS, Azure, GCP	Brand, global ecosystem, enterprise compliance pedigree	Rupee pricing, sovereignty, edge latency, group distribution
Domestic neoclouds	Yotta, E2E, CtrlS, Sify	Price, local presence, faster provisioning	Scale economics, renewable power cost advantage, captive demand
Telecom-adjacent	Tata Communications	Existing enterprise telecom relationships	Jio's larger distribution footprint plus a purpose-built AI stack
Subsidized / mission programs	IndiaAI Mission allocations	Price to startups and academia	Enterprise-grade SLAs and a path from subsidized pilot to commercial scale

Positioning is directional, drawn from the market-context assessment in this plan, not a formal competitive-intelligence study; validate pricing and share claims against Infrastructure and Corporate Strategy data in the first 30 days.

3. The Business Model in One Picture

A GPU cloud is, financially, a leasing business wearing a software company's clothes. The provider raises capital, converts it into a rapidly depreciating asset (GPU servers plus the power, cooling, and network needed to run them), and then sells time on that asset in two very different wrappers: capacity, priced per GPU-hour and sold largely through multi-year take-or-pay contracts, and tokens, priced per million tokens of model input and output and sold through a usage-metered API. The first wrapper looks like commercial real estate or aircraft leasing; the second looks like a telecom selling minutes. The whole P&L lives or dies on three numbers: the price achieved per GPU-hour (or its token-denominated equivalent), the utilization of the fleet, and the depreciation life assumed for the hardware.

The middle of the business — metering, rating, billing, settlement — is where reported revenue is actually manufactured, and it is where money silently leaks. Every GPU-second and every token must be captured as an event, aggregated correctly, rated against the right price book, invoiced, collected, and recognized under the right accounting treatment. Telecom operators learned decades ago that 1–3% of revenue routinely disappears between the network switch and the invoice; usage-metered compute has exactly the same anatomy and exactly the same failure modes. Revenue assurance — the discipline of proving that everything used was billed and everything billed was correct — is therefore not a back-office nicety but a direct, high-margin lever on the bottom line: a recovered dollar of leakage carries no incremental COGS.

Two revenue engines, one asset base

- **The capacity business (GPU-hours):** the provider sells dedicated GPU instances or whole clusters, metered in GPU-hours (usually rated per second or per minute). The dominant commercial form is the multi-year take-or-pay reservation: the customer commits to pay for a fixed quantity of capacity whether or not it is used, often 1–5 years, sometimes with prepayments. On-demand and spot capacity monetize whatever the committed book leaves stranded.
- **The token business (inference-as-a-service):** the provider runs models (its own, open-weight, or the customer's fine-tunes) behind an API and charges per million input tokens and per million output tokens, with distinct rates for cached input and batch processing. Here the provider absorbs the utilization risk that the capacity customer would otherwise carry — and prices for it.

The strategic difference between the two is who owns the utilization problem. In the capacity business, a take-or-pay contract transfers utilization risk to the customer: revenue is contractual, visibility is measured in years (CoreWeave disclosed a $99.4 billion revenue backlog at March 31, 2026), and the provider's residual risks are counterparty credit, delivery milestones, and residual value at contract end. In the token business the provider retains utilization risk — tokens only bill when traffic arrives — but earns a substantially higher achieved price per GPU-hour when traffic is dense, because one GPU-hour can be resold many times over as batched, multiplexed token throughput. A well-run operator uses the token layer as a yield-management overlay on capacity the reserved book does not fully absorb.

Financially, the two lines also differ in revenue quality. Capacity revenue under committed contracts is recognized ratably as the capacity is made available, is backed by remaining performance obligations, and supports debt financing against contracted cash flows. Token revenue is pure usage revenue: recognized as consumed, re-priced frequently (token prices have deflated sharply every year as hardware and software efficiency compound), and far more exposed to competition. Rating agencies, lenders, and equity analysts will price the two streams very differently, which is why disclosure of backlog, RPO, and contract mix matters so much in this sector.

Part II — Commercial Architecture

4. SKU Portfolio and Sequencing

Seven launch SKUs, sequenced in three waves against the Jamnagar commissioning curve. Wave 1 maximizes certainty and utilization; Wave 2 builds margin-rich differentiated products; Wave 3 builds the ecosystem flywheel.

SKU	What is sold	Buyer	Wave
Reserved GPU Capacity	Dedicated GB300 blocks, 1–3 year terms, take-or-pay	AI labs, GCCs, captive RIL units, Meta-style anchors	1
Token-as-a-Service	Metered inference per 1M tokens across a curated model catalog	Enterprises, SaaS, developers	1
Fine-Tuning-as-a-Service	Training runs priced per token processed plus hosted-model serving	Enterprises with proprietary data	2
Private Enterprise Endpoints	Provisioned-throughput, single-tenant serving with residency guarantees	BFSI, healthcare, government	2
Telecom Edge Inference	Latency-tiered inference at Jio edge sites, per-token premium	Real-time apps, voice, IoT, autonomous platforms	2
Agent Runtime	Metered agent-hours plus per-task pricing for autonomous workloads	SaaS builders, enterprise automation	3
Marketplace Revenue Share	Third-party models, agents, and datasets sold on-platform	Model builders, ISVs, telecom partners	3

Sequencing logic: reserved capacity and token serving can be sold today against the end-2026 commissioning date (contracts signed now, revenue recognized at delivery). Fine-tuning, private endpoints, and edge follow within two quarters of first capacity. Agent runtime and marketplace launch once the platform has reference customers and settlement rails proven at scale.

Illustrative SKU mix at maturity

Consistent with the SKU-level ROI model in Part III, the portfolio's trailing-twelve-month revenue and margin mix is expected to look approximately as follows once all three waves are live.

SKU	TTM revenue ($M)	Gross margin (mid)	Wave
Reserved GPU capacity	61.4	38%	Wave 1
Token-as-a-service	28.7	60%	Wave 1

SKU	TTM revenue ($M)	Gross margin (mid)	Wave
Private enterprise endpoints	12.3	62%	Wave 2
Fine-tuning-as-a-service	6.8	65%	Wave 2
Telecom edge inference	4.1	50%	Wave 2
Agent runtime	1.6	70%	Wave 3
Marketplace revenue share	0.9	25%	Wave 3

Illustrative figures consistent with the Strategy & GTM model; reserved capacity dominates revenue while token-as-a-service and private endpoints carry the portfolio's margin.

5. Pricing Architecture

Four commercial models, one discipline: every price traces to cost per token, which traces to GPU-hour cost, power, utilization, and model efficiency. The pricing council reviews monthly as GB300 performance data and market rates move.

Model	Structure	Design principles
Token-metered	Per 1M input/output tokens, tiered by model class (frontier, mid, Indic-optimized, embeddings)	Output tokens priced 3–4x input; volume tiers; rupee-denominated; batch and off-peak discounts to shape load into utilization valleys
GPU-hour on-demand	Per GPU-hour by SKU class	Priced at a premium to committed rates; capacity-managed so on-demand never crowds out committed obligations
Reserved / committed	1–3 year take-or-pay blocks; committed-spend drawdown pools	15–45% discount ladder by term and volume; anchor tenants underwrite the depreciation schedule; renewal motion begins at 60% of term
Outcome-based	Per resolved task, per document processed, per agent completion	Only where the platform controls the full workflow (agent runtime, Jio AI services); protects against token-price commoditization by pricing value, not volume

- **Packaging:** three enterprise tiers (Launch, Scale, Sovereign) bundling tokens, fine-tuning credits, endpoint throughput, and support SLAs; a self-serve developer tier with transparent card-billed pricing to seed the funnel.
- **Price governance:** a deal desk owns discount authority above thresholds; every non-standard deal carries a margin floor at cost per token plus target return on the underlying reserved capacity.

Capacity pricing: reserved, on-demand, spot

Capacity pricing is a curve, not a number. The same physical H100-hour trades at wildly different prices depending on commitment length, cluster quality (interconnect, storage, support), and channel. Illustrative market benchmarks as of mid-2026:

Product form	Typical H100 price / GPU-hr	Notes
Spot / interruptible	$0.55 – $1.70 (avg ≈ $1.69)	Preemptible; monetizes stranded capacity
On-demand, neocloud	$1.89 – $2.99	UpCloud $1.89, RunPod $1.99–2.59, Hyperstack $2.50
On-demand, hyperscaler	$5.99 – $6.98	AWS p5 $6.88, Azure NC-H100 $6.98
1-yr reserved contract	$1.70 – $2.35	Late-2025/early-2026 market range
Multi-yr take-or-pay cluster	negotiated (≈ $1.50 – $2.50 eff.)	Anchor deals; milestone delivery; prepayments
Market average (all forms)	≈ $3.65	Reserved avg ≈ $3.52 across listings

Sources: getdeploying.com H100 price index (48+ providers); zettabyte.space neocloud unit-economics analysis. Prices move quickly; treat as illustrative.

Three structural features of capacity pricing deserve emphasis. First, the commitment discount is really a utilization-risk transfer price: the gap between $6.98 on-demand at a hyperscaler and ~$2.00 on a one-year neocloud contract is what the market charges for holding idle-capacity risk, plus brand, ecosystem, and enterprise-compliance premia. Second, generation transitions reprice the whole curve: when Blackwell arrives at scale, H100 contract prices sag, which is why depreciation life and contract tenor must be jointly managed — a six-year depreciation schedule against a fleet whose market rate halves in year three is a structural mismatch. Third, large deals are increasingly structured, with delivery milestones (revenue starts when capacity is accepted), ramp schedules, prepayments that fund capex, and sometimes seller financing or vendor equity — all of which complicate both revenue recognition and cash settlement (Part IV).

Token pricing

Token pricing is a rate card per million tokens, always asymmetric between input and output. Output tokens are priced 3–5x input because generation is sequential (one forward pass per token) while input prefill parallelizes efficiently across the sequence. Illustrative rate-card points as of July–August 2026:

Model / tier	Input $/M tok	Output $/M tok	Notes
Frontier (GPT-5.6 Terra)	$2.50	$15.00	Premium closed model
Frontier (Claude Sonnet 5)	$2.00	$10.00	Cached input ≈ 10% of list
Fast tier (Gemini 2.5 Flash)	$0.30	$2.50	Flash-Lite input from $0.10
Value tier (DeepSeek-class)	$0.14 – $0.55	$2.19	Cached input as low as $0.0028
Open-weight small (Together/Groq)	$0.03 – $0.08	$0.08 – $0.12	Commodity floor

Sources: tldl.io verified LLM pricing tracker (July 2026); introl.com inference unit-economics guide. Cached-input discounts of ~90% and batch-API discounts of ~50% are now standard.

The rate card has grown real microstructure: cached-input pricing (typically ~10% of list) rewards prompt reuse and materially changes both revenue per request and metering requirements, since the biller must now distinguish cache hits from cache misses per request; batch pricing (typically 50% off) sells latency flexibility, which is to the token business what interruptible spot is to the capacity business; and long-context surcharges, image/audio token conversion rates, and per-request minimums round out the card. Every one of these distinctions is a metering obligation: a price that cannot be metered is a discount that was not intended.

The price book as a controlled object

Between list price and invoice sits the commercial machinery: enterprise discounts, committed-spend deals (customer commits to $X over a term, drawn down against usage at discounted rates), prepaid credit packs, free tiers, promotional credits, and partner/reseller margins. The operational requirement is that all of this lives in a versioned, dated, access-controlled price book that the rating engine consumes — not in spreadsheets, sales emails, or hand-edited invoice adjustments. Every revenue-assurance program that finds material leakage finds some of it here: an expired discount still applied, a negotiated rate never loaded, a currency mis-set, a free tier without an enforcement cap. Rate-table changes are treated with the same change-management rigor as production code deployments, because that is exactly what they are.

Enterprise packaging: three tiers

The three enterprise tiers referenced above translate the pricing architecture into something a buyer can actually evaluate in a 30–45 day decision cycle, each bundling

committed tokens, fine-tuning credits, endpoint throughput, and a support SLA at increasing levels of guarantee.

Tier	Target buyer	Bundle	Support SLA
Launch	Mid-market enterprise, first AI workload	Token allotment + shared endpoints + community/standard support	Business hours; 99.5% uptime
Scale	Growth-stage enterprise, multiple workloads	Larger token pool + fine-tuning credits + provisioned throughput option	24x5; 99.9% uptime; named CSM
Sovereign	BFSI, healthcare, government	Private single-tenant endpoints + data-residency guarantee + dedicated capacity option	24x7; 99.95% uptime; audit rights; dedicated TAM

The enterprise sales motion, step by step

Segment 3 (Enterprise and BFSI) is the margin engine of the plan, and its land-and-expand motion is deliberately time-boxed so pipeline does not stall in evaluation purgatory.

Stage	Typical duration	Owner	Exit criteria
Discovery & workload fit	1–2 weeks	Enterprise AE + solution architect	Named workload, success metric, and budget owner identified
Pilot design	1 week	Solution architect	Scoped private-endpoint or fine-tuning pilot with a fixed decision date
Pilot execution	2–4 weeks	Solution architect + Customer Success	Success metric hit; reference-able result documented
Commercial proposal	3–5 days	Enterprise AE + deal desk	Tiered package proposed against margin-floor discipline
Contracting	1–2 weeks	Deal desk + Legal	Committed-spend contract signed at or above target ARPA
Expansion	Ongoing, quarterly review	Customer Success	NRR tracked against the 120%+ target

Total land cycle target: 30–45 days from discovery to signature, consistent with Part II, Segment 3.

6. Go-to-Market by Segment

Segment 1: Captive Reliance businesses (utilization floor)

Jio network AI, JioBrain workloads, JioStar GenAI Media Studio, Retail demand forecasting, and the five Jio AI services migrate onto internal committed-capacity contracts at arm's-length transfer pricing. Target: 25–35% of Wave-1 capacity under internal MOUs

before external launch, priced to recover full cost plus a modest internal margin so external pricing integrity is never undermined.

Segment 2: Anchor external tenants (the Meta motion, repeated)

Five to eight named-account pursuits for multi-year reserved capacity: global hyperscalers needing India capacity, frontier labs seeking inference distribution in India, large GCCs (banks, retailers, pharma) with sovereign-AI mandates, and Indian AI startups with national-scale ambitions. This is a CEO-to-CEO sales motion supported by the CCO with bespoke structuring: take-or-pay with utilization credits, expansion options priced as capacity calls, and co-marketing rights.

Segment 3: Enterprise and BFSI (the margin engine)

Top 200 Indian enterprises and GCCs, led by regulated industries where sovereignty commands premium pricing. Land with a private endpoint or fine-tuning pilot scoped to a 30–45 day decision, expand into committed-spend contracts. Field organization of enterprise account executives paired with solution architects, comped on committed ARR and net revenue retention, not bookings.

Segment 4: SaaS, startups, and developers (the volume flywheel)

Self-serve token tier with transparent pricing, generous free tier for Indic-language models, IndiaAI Mission alignment for subsidized startup access, and a partner program for SaaS platforms embedding the API. This funnel feeds the marketplace and generates the usage data that sharpens cost-per-token management.

Segment 5: Telecom edge and government

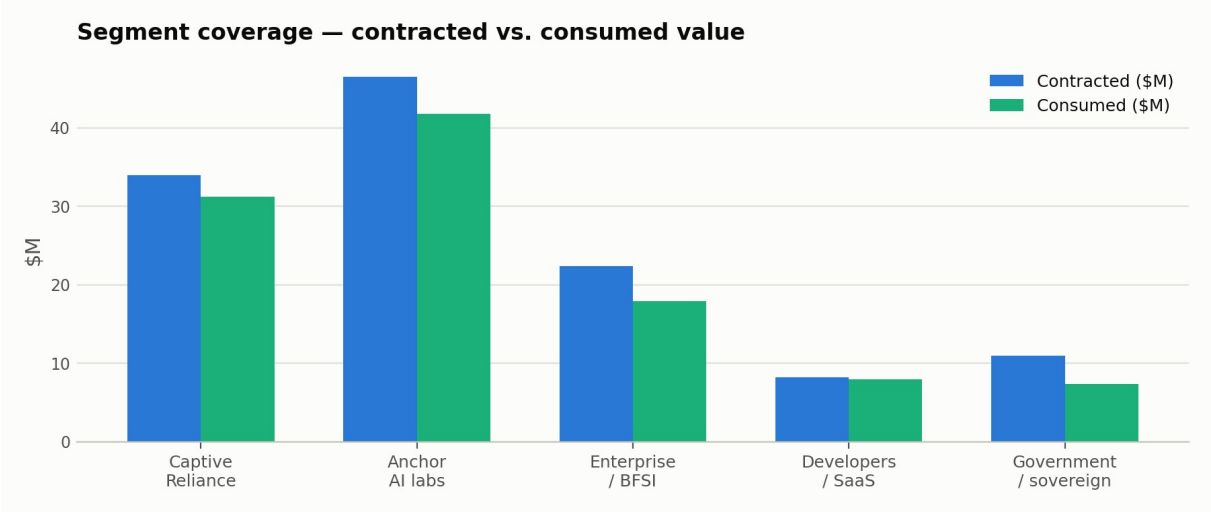
Edge inference sold through Jio's existing enterprise telecom channel as an attach to connectivity contracts; government and PSU demand pursued through sovereign-cloud framework agreements where residency, auditability, and Indic-language capability are the spec, and price competition is secondary.

Segment coverage: contracted vs. consumed

The GTM cockpit tracks each segment's contracted value against what it has actually consumed to date — the single fastest read on whether a segment is under-penetrating its commitment or running ahead of it.

Segment	Contracted ($M)	Consumed ($M)	Consumption rate
Captive Reliance	34.0	31.2	92%
Anchor AI labs	46.5	41.8	90%
Enterprise / BFSI	22.4	17.9	80%
Developers / SaaS	8.2	8.0	98%
Government / sovereign	11.0	7.4	67%

Figures consistent with the Strategy & GTM model. Government / sovereign is the segment most worth watching: the lowest consumption rate of the five, and the segment where framework-agreement cycle time (not price) is the usual constraint.



Contracted vs. consumed value by segment ($M). Developers/SaaS is the tightest (98% consumption — a self-serve funnel converting almost exactly what it commits), while government/sovereign carries the widest gap between what is contracted and what has actually been drawn down.

7. Marketplace Commercial Model

- **Listing economics:** revenue share tiered by exclusivity and integration depth (approximately 80/20 for standard listings, improving toward 70/30 platform share for premium placement, fine-tuned variants, and bundled distribution through Jio channels).
- **Settlement mechanics:** monthly settlement to sellers with full metering transparency; the platform's billing system is the single source of truth for both sides, with seller-facing dashboards mirroring buyer-facing invoices.
- **Telecom partners:** edge-inference revenue shared with carrier partners where their spectrum and sites participate in delivery; modeled on interconnect settlement, a discipline the Jio organization already runs at scale.

- **Curation:** a quality-gated catalog, not an open bazaar: every listed model carries benchmark cards, safety attestations, and Indian-language capability scores, protecting enterprise trust in the platform.

Wave-3 targets: a marketplace beta with 10+ listed models in year one, scaling toward $3.4M of trailing-twelve-month GMV and a 65% coverage target on Wave-1 committed capacity as the anchor book fills in behind it — tracked live on the GTM & SKU Mix tab of the cockpit alongside a net-revenue-retention target of 122%.

Illustrative catalog composition at beta

A quality-gated catalog spanning frontier, Indic-optimized, and vertical fine-tuned listings gives the marketplace credibility with enterprise buyers from day one rather than reading as an open, unvetted bazaar.

Listing category	Example	Pricing type	Platform take
Frontier general-purpose	Third-party frontier LLM, hosted	Per-token, standard listing	20%
Indic-language specialist	22-language fine-tune of an open-weight model	Per-token, premium placement	30%
Vertical fine-tune — BFSI	Regulatory-document extraction model	Per-token + per-task hybrid	30%
Vertical fine-tune — Healthcare	Clinical-notes summarization model	Per-task, outcome-based	30%
Agent template	Customer-support resolution agent	Per-completion	25%
Embeddings / retrieval	Domain-tuned embedding model	Per-token, standard listing	20%
Dataset	Licensed Indic-language training corpus	One-time / subscription	20%
Telecom edge bundle	Real-time voice-agent inference, Jio edge	Per-token, edge premium + carrier share	30% platform / carrier split

Net revenue retention: reading the bridge

NRR is the single number that tells the board whether the existing customer base is growing or shrinking on its own, independent of new-logo sales — and at a 122% target, it means the base alone would grow revenue by 22% even with zero new customers. The bridge below shows how starting committed ARR moves to ending ARR across a representative quarter.

NRR bridge component	$M	Notes
Starting committed ARR	100.0	Beginning-of-quarter base
+ Expansion (upsell, usage growth)	+28.0	Capacity upgrades, token consumption growth, new SKU attach
− Contraction (downgrades)	-4.0	Reduced commitment at renewal, tier downgrades
− Churn (non-renewal)	-2.0	Full account loss — target: near zero for committed accounts
Ending committed ARR (existing base only)	122.0	NRR = 122.0 / 100.0 = 122%

Illustrative bridge consistent with the 122% NRR target in the GTM data model. Expansion is the dominant lever by design — Segment 3's land-and-expand motion (Part II, above) and Segment 2's capacity-call expansion options (Part II, Segment 2) are both structured to drive this line.

Part III — The Financial Engine

This part is the underwriting case: what a GPU-hour actually costs, what it earns under two different commercial wrappers on the identical fleet, and how the pieces assemble into an income statement benchmarked against the sector's clearest public comparable. Every number here is an illustrative modeled estimate built from public market rates and standard cloud unit economics as of mid-2026 — not Reliance internals — stated with its assumptions so it can be pressure-tested and replaced with actuals in the first 30 days.

8. The Cost Stack

The cost side of the P&L is dominated by capital recovery. A useful mental model: roughly half to two-thirds of the fully-loaded cost of a GPU-hour is depreciation and financing on the hardware itself; power, facilities, network, and people share the remainder.

- **Capex.** an 8×H100 HGX server runs ≈ $300k; network fabric (InfiniBand/RoCE), storage, and integration add 15–25%, landing all-in cluster capex around $40–48k per GPU. Blackwell-class systems are substantially more per GPU but deliver more throughput per dollar.
- **Depreciation life.** the single most consequential accounting judgment in the sector. The same server generates ≈ $50k/yr of depreciation on a 6-year life and ≈ $75k/yr on a 4-year life. Long lives flatter EBIT today at the cost of overstating the asset's late-life earning power; the market debate over 5–6 year GPU lives versus 3–4 year competitive relevance is, in essence, a debate about whether reported margins in this sector are real.
- **Power and facilities.** an H100 draws ~700W (B200 up to ~1,000W); a dense rack draws 40–80kW, 3–8x a conventional enterprise rack. At a 1.3 PUE and $0.09/kWh, power is roughly $0.15 per GPU-hour; colocation space and cooling, priced per kW-month, add roughly $0.20–0.25.
- **Financing.** fleets are heavily debt-financed against contracted cash flows. CoreWeave carried $24.9B of debt at Q1 2026 with net interest expense of $536M in the quarter — over 25% of revenue — with recent facilities priced around SOFR+2.25% floating / ~5.9% fixed, against earlier GPU-backed facilities in the 8–11% range. Interest sits below EBITDA but is utterly real cash: a debt-financed fleet breaks even in the high-$1s per GPU-hour before a dollar of profit.

- **People and platform.** site reliability, datacenter ops, support, and the software layer (scheduler, storage, billing itself). Small per GPU-hour (≈ $0.10 in the worked example) but scales with product complexity — the token business carries meaningfully more engineering cost than raw capacity.

The Kutch renewable-power advantage

Power is typically 15–25% of an all-in GPU-hour cost, and it is the one line in the cost stack where Reliance's asset base creates a structural, durable gap versus every competitor in this document's landscape table. A captive solar transfer rate meaningfully below grid-blended industrial tariffs compounds directly into cost per token: at the $0.09/kWh figure used throughout this document's worked examples, power contributes roughly $0.15 per GPU-hour to the cost stack; a competitor buying grid power at a materially higher blended industrial rate carries a proportionally higher cost floor on every GPU-hour it sells, a gap that persists for the life of the power purchase agreement regardless of how efficiently either side runs its servers. This is the reason the market-context assessment in Part I treats scale-and-cost as a right-to-win rather than a temporary edge: it is financed and contracted years in advance, not a pricing decision that can be matched next quarter.

The practical implication for the commercial organization is to treat the power cost advantage as a margin source to be banked, not necessarily a price cut to be passed through. Because on-demand and spot pricing in the India market is already compressed (Part II, Section 5), competing away the power advantage through list-price cuts would forfeit exactly the differentiated economics that fund the depreciation schedule; the plan instead directs the advantage toward the sovereignty-premium SKUs and the token-margin curve (Part III, Section 10), where it shows up as superior gross margin rather than as a race to the bottom on headline price.

9. Worked Example A: 1,024-GPU Capacity Cluster P&L

Assumptions: 128 servers x 8 H100 = 1,024 GPUs; $300k/server plus 18% for fabric, storage, and integration → $45.3M capex ($44.3k/GPU); 5-year straight-line depreciation; 1.275 kW IT load per GPU, PUE 1.3, $0.09/kWh; colocation at $135/kW-month on IT load; ops and support at $0.10/GPU-hr; 70% debt at 8.0% interest-only for simplicity. The entire cluster is sold under a take-or-pay reservation at an effective $2.10/GPU-hr, so billable hours = 8,760 x 1,024 = 8.97M GPU-hours regardless of customer utilization.

Line item	$/GPU-hr	$M / year	% of revenue
Revenue (take-or-pay @ $2.10)	2.100	18.84	100.0%
Power (1.275 kW x 1.3 PUE x $0.09)	(0.149)	(1.34)	7.1%
Colocation ($135/kW-mo)	(0.236)	(2.12)	11.2%
Ops, support, platform	(0.100)	(0.90)	4.8%
EBITDA	1.615	14.49	76.9%
Depreciation (5-yr SL on $45.3M)	(1.010)	(9.06)	48.1%
EBIT	0.604	5.42	28.8%
Interest (70% LTV @ 8.0%)	(0.283)	(2.54)	13.5%
Pre-tax income	0.322	2.89	15.3%

Cluster-level illustration. EBITDA margin of ~77% at full contract coverage compares to CoreWeave's reported 56% adjusted EBITDA margin in Q1 2026 (company level, diluted by ramping capacity, on-demand mix, and opex).

Read the sensitivity, because it is the whole business. At $2.10 with full take-or-pay coverage this cluster earns a 15% pre-tax margin and pays back its full capex in about 3.8 years of pre-principal cash flow — uncomfortably close to the useful life. Move the achieved price to $1.70 (the bottom of the 2026 contract range) and pre-tax income falls to roughly minus $0.7M: below breakeven after financing. Shorten depreciation to 4 years at $2.10 and EBIT drops from $5.4M to $3.2M. Leave 20% of the cluster unsold (no take-or-pay, merchant exposure) and EBITDA falls by ~$3.8M against a fixed cost base. The equity case is leveraged to price, tenor, and utilization simultaneously — which is precisely why the committed backlog, not the GPU count, is the asset that lenders underwrite.

10. Worked Example B: Token Unit Economics on the Same Fleet

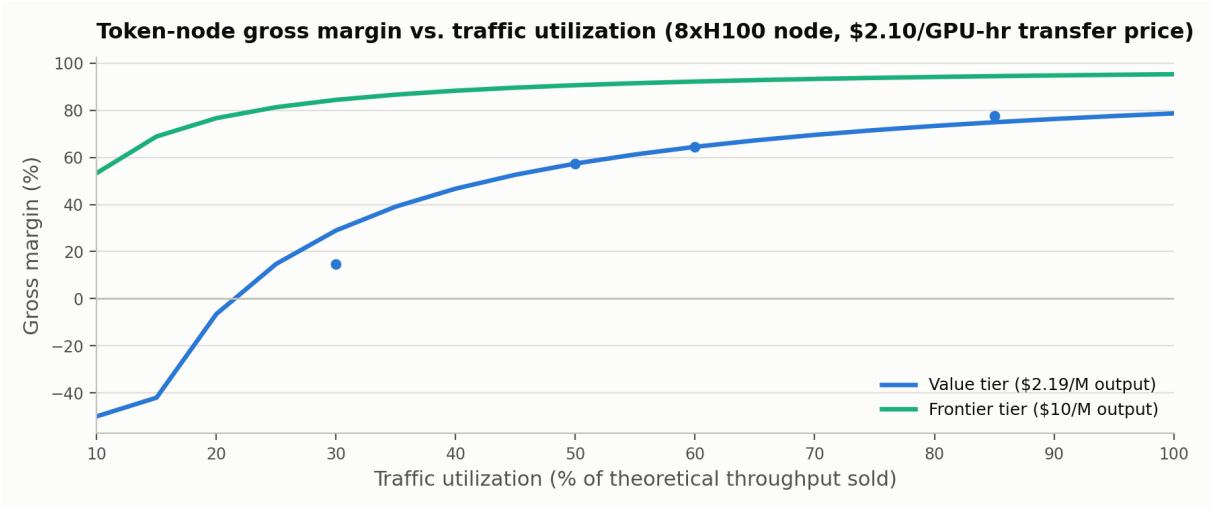
Now run an inference endpoint on one 8xH100 node from the same fleet, charged internally at the $2.10/GPU-hr transfer price → $16.80/node-hr. Serving a 70B-class open-weight model with tensor parallelism across the node, a well-tuned stack (continuous batching, paged KV-cache, speculative decoding) sustains ≈ 10,000 output tokens/second aggregate at healthy batch depth, and prefill throughput roughly 25x that for input tokens. Utilization here means traffic density: the fraction of theoretical token throughput actually sold across the day's peaks and troughs.

Traffic utilization	Output tok sold / node-hr	Cost / 1M output tok	Margin @ $2.19/M	Margin @ $10/M
30%	10.8M	$1.56	29%	84%

Traffic utilization	Output tok sold / node-hr	Cost / 1M output tok	Margin @ $2.19/M	Margin @ $10/M
50%	18.0M	$0.93	57%	91%
60%	21.6M	$0.78	64%	92%
85%	30.6M	$0.55	75%	95%

Input tokens cost ≈ $0.03/M to serve at 60% utilization against list prices of $0.14–$2.50/M — richly profitable, which is why cached-input discounts of 90% are commercially survivable.

Two lessons fall out. First, utilization is the margin: the identical node produces a 29% or a 75% gross margin at the same rate card depending purely on traffic density, which is why serverless token providers obsess over routing, multiplexing, multi-tenancy, and autoscaling-to-zero. Second, the token wrapper can dramatically out-earn the capacity wrapper on the same silicon: at 60% utilization and even the value-tier $2.19/M output price, the node generates ≈ $47/hr of token revenue against a $16.80 internal capacity cost — an achieved price of ~$5.90 per GPU-hour, nearly 3x the take-or-pay rate. That spread is the reward for absorbing utilization risk and operating a much more complex metering, serving, and billing stack. It is also why token prices deflate relentlessly: every efficiency gain (quantization ~50%, speculative decoding 2–3x latency, better batching — compounding to ~16x versus naive deployment) becomes price competition within quarters.



Gross margin vs. traffic utilization for the same 8xH100 node at two rate-card tiers. The value tier only clears a healthy margin once traffic density is well above 40%; the frontier tier is profitable almost everywhere — the commercial argument for prioritizing frontier and mid-tier model traffic onto capacity that is still filling up.

11. SKU-Level ROI Model

Cost basis assumptions

- **All-in GPU-hour cost:** $1.40–$1.80 per H100-equivalent GPU-hour, comprising GB300 capex amortized over 5 years, power at Kutch solar transfer rates, facility (PUE assumption 1.2–1.3), network, and operations. Renewable power is the structural advantage: power is typically 15–25% of all-in cost, and captive solar compresses it.
- **Utilization:** margins quoted at 70% revenue-weighted utilization for always-on SKUs; reserved capacity is utilization-independent by construction (take-or-pay).
- **Pricing references:** public India and global rates for on-demand GPU-hours, per-token serving, and sovereign-cloud premiums observed across hyperscaler India regions and domestic providers.

ROI by SKU

SKU	Gross margin	Capital intensity	Payback	Primary risk
Reserved GPU capacity	30–45%	Very high	3.5–4.5 yrs	Lowest margin, but zero utilization risk; underwrites the depreciation schedule
Token-as-a-service	50–70% at high util	High	2.5–3.5 yrs	Token prices fell ~10x in two years; margin decays unless serving efficiency keeps pace
Fine-tuning-as-a-service	60–70%	Medium (bursty)	2–3 yrs	Smaller TAM; lumpy demand
Private enterprise endpoints	55–70%	High	2.5–3 yrs	Best risk-adjusted ROI: 30–50% sovereignty premium, sticky BFSI contracts, low churn
Telecom edge inference	40–60%	Medium; less efficient/watt	3–4 yrs	Premium pricing real, but per-unit cost higher; volume unproven
Agent runtime	70%+ potential	Low incremental	<2 yrs if demand holds	Outcome pricing decouples from token commoditization, but demand is speculative today
Marketplace revenue share	20–30% take	Minimal	Fastest ROI/$ once GMV exists	Cold-start: GMV small for 12+ months

Payback is against allocated capex. Illustrative modeled estimates, not Reliance internals.

Portfolio read

Reserved capacity carries the worst margin but buys the certainty that pays for the asset. Private endpoints are the best risk-adjusted return and the SKU where India-specific advantage (sovereignty, DPDP compliance, Indic-language serving) is hardest to compete away. Token-as-a-service shows good headline margin but is most exposed to commoditization. Agent runtime and marketplace are inexpensive options on the future, not year-one P&L.

Claims to validate in the first 30 days

- **Cost floor:** the $1.40–$1.80 all-in GPU-hour figure depends on actual PUE, solar transfer pricing, and GB300 delivered performance; validate against Infrastructure's cost model.
- **Sovereignty premium:** the 30–50% private-endpoint premium should be benchmarked against what Yotta, hyperscaler India regions, and domestic providers actually charge BFSI and government buyers.
- **Serving yield:** tokens per GPU-hour is a function of batching, quantization, and model mix; it can swing token-SKU margin 20 points in either direction and is the single highest-leverage number in the model.

12. The Assembled P&L and Public Comparables

Putting the pieces together, the income statement of a combined capacity-and-token provider reads as follows, with the operational system that feeds each line noted alongside it. Revenue splits into committed capacity (ratable recognition off the contract and delivery ledger), on-demand/spot (metered GPU-seconds, rated), and token/inference (metered tokens, rated), net of estimated SLA credits and credit memos (variable consideration). Cost of revenue carries power, colocation and datacenter opex, cluster support, and — under the dominant convention — depreciation of revenue-generating hardware, making reported gross margin a direct function of the depreciation-life judgment. Operating expense holds platform engineering, sales, G&A, and the billing/assurance function itself. Below EBIT, interest expense on the fleet financing is the line that separates the sector's cheerful adjusted-EBITDA story from its GAAP reality.

P&L line	Fed by	Key risk / control
Committed capacity revenue	Contract ledger + delivery	Acceptance timing; counterparty credit

P&L line	Fed by	Key risk / control
	milestones	
On-demand & spot revenue	GPU-second metering → rating	Completeness; orphaned allocations
Token revenue	Token metering → rating	Tokenizer drift; cache-hit classification
(SLA credits & memos)	Incident system; credit governance	Over-crediting; estimation of variable consideration
Power & facilities	Telemetry + supplier invoices	Supplier take-or-pay mismatch
Depreciation	Fixed-asset register	Life assumption vs. competitive life
Interest expense	Debt schedule	Rate, covenants, contracted-cashflow coverage

Illustrative three-year trajectory

Extending the model-year build (Appendix A) forward at a moderating growth rate consistent with the Wave 2/3 SKU launches in Part II gives a directional multi-year shape for board planning purposes. These are illustrative projections, not commitments — the first genuine three-year plan should be rebuilt once Part III's 30-day validation claims are closed out against actuals.

Metric	Year 1	Year 2	Year 3
Fleet size (GPUs, exit)	8,192	~14,000	~20,000+
Revenue run-rate (exit, $M)	~$140	~$260	~$420
EBITDA margin	65–70%	60–65%	58–63%
Committed-capacity coverage	72–77%	70–75%	65–70%
Contribution from Wave 2/3 SKUs	<10%	25–30%	35–40%

Illustrative, not a formal forecast. EBITDA margin is expected to moderate as Wave 2/3 SKUs (fine-tuning, edge, agent runtime, marketplace) — lower-margin at launch and margin-accretive at scale — take a growing share of the mix, and as committed-capacity coverage normalizes from an early-period high toward the steady-state 60–70% target as the fleet itself grows faster than any single wave of anchor contracts.

The public reference point makes the shape concrete. In Q1 2026 CoreWeave reported revenue of $2.08B (+112% YoY), adjusted EBITDA of $1.16B (56% margin, down from 62% a year earlier), an operating loss of $144M (-7% margin), and a net loss of $740M (-36% margin), with D&A of $1.15B and net interest expense of $536M in the quarter against $24.9B of debt and $7.7B of quarterly capex. That is the sector's margin bridge in one line: a 56-point EBITDA margin becomes a negative operating margin once the quarter's depreciation is charged, and a deeply negative net margin once the financing of the fleet is

paid — while a $99.4B backlog argues that today's losses are the funded construction phase of tomorrow's contracted revenue. Whether that argument closes depends on exactly the variables this section has priced: achieved $/GPU-hour versus the cost stack, real hardware life versus depreciation schedules, utilization of the uncommitted fleet, counterparty credit in a concentrated backlog — and a metering-to-settlement chain tight enough that the revenue actually contracted is the revenue actually collected.

A closing heuristic for the operator: the capacity business is won at contract signature (price, tenor, credit), the token business is won in the serving stack (utilization, throughput per dollar), and both are kept honest in the metering-to-settlement chain covered next, in Part IV. Pricing sets the ceiling on the P&L; the cost stack sets the floor; metering, billing, settlement, and revenue assurance determine how much of the space between them is actually kept.

13. Scenario Planning: Stress-Testing the Model

The cockpit's model is built to be stressed, not just admired. Four named scenarios — one base case and three deliberately adverse or favorable cases — run the same monthly build with different multipliers on price, coverage, opex, and fleet size, so the board can see the plan's resilience before the market tests it.

Scenario	Price	Coverage	Opex	Fleet	Annual EBITDA	Δ vs. base
Base case	1.00x	1.00x	1.00x	1.00x	$87.3M	—
Price crash (hedged)	0.85x	0.95x	1.00x	1.00x	$69.0M	-$18.3M
Demand shock (upside)	1.05x	1.15x	1.00x	1.10x	$110.4M	+$23.0M
Cost inflation	1.00x	0.95x	1.15x	1.00x	$80.1M	-$7.2M

Modeled on the cockpit's base-case monthly build (Aug 2025–Jul 2026, illustrative figures). The price-crash case assumes the compute-futures hedge described in Part V is in place on the merchant tail; without it, the EBITDA impact of a comparable price move is materially larger.

Three readings matter more than the headline numbers. First, the price-crash case is survivable — EBITDA falls by roughly a fifth, not by half — specifically because the hedge overlay in Part V is doing its job on the uncommitted tail; this is the strongest quantitative argument in the entire plan for standing up the futures desk early rather than treating it as a year-two nicety. Second, the demand-shock case shows the upside is asymmetric: a 10%

larger fleet plus 15% better coverage plus a modest price uplift compounds to a 26% EBITDA gain, which is the case for erring toward capacity aggressiveness once the commissioning curve is credible. Third, cost inflation — a 15% rise in cash opex, most plausibly from power or colocation cost pressure — is the most likely of the three adverse scenarios and the one most directly within the CCO organization's control to offset, through the load-shaping and cost-per-token discipline in Part IV, Section 18.

Part IV — Billing, Metering, and Revenue Assurance: The System of Record

This is the discipline most token businesses get wrong, and the one this plan treats as home ground. Revenue leakage in usage-based businesses typically runs 1–3% of revenue; at $100M+ ARR that is a leadership-team-sized sum lost annually. The plan below is how this gets run week to week, and it is operationalized live in the Commercial Cockpit's Revenue Assurance and Meter-to-Cash views (Part VII).

- **Metering integrity:** token and GPU-hour metering instrumented at the serving layer with immutable event logs; daily automated reconciliation between metering, rating, and invoicing, with variance thresholds that page the FinOps team — the same control philosophy as trade-settlement breaks on a derivatives desk.
- **Order-to-cash ownership:** one pipeline from onboarding, KYC, and credit assessment through rating, invoicing, collections, and dunning; enterprise net-terms risk managed with credit limits and prepay conversion triggers.
- **Revenue assurance function:** a dedicated team inside the CCO org running leakage analytics: unbilled-usage detection, rate-plan drift, discount-policy compliance, marketplace settlement audit, and fraud/abuse monitoring (prompt-loop abuse, token-farming, resale violations).
- **FinOps for customers:** customer-facing cost dashboards, budgets, and anomaly alerts. Counterintuitive but essential: helping customers avoid bill shock is the single strongest driver of renewal and expansion trust in consumption businesses.
- **Build/buy posture:** rate and mediate on a proven usage-billing engine rather than building from scratch; differentiate in the metering layer and the reconciliation controls, and keep the system auditable to Big-4 standard from day one, because anchor-tenant contracts will demand it.

14. Metering: Turning Physics into Billable Events

Metering is the sensory system of the P&L. Everything downstream — invoice, revenue, cash, audit — is a transformation of metering events, so an unmetered event is unbillable revenue lost forever, and a double-metered event is a dispute, a credit memo, and reputational damage. The two product lines meter very different things.

Product	Billable meters	Primary source of truth
Capacity	GPU-seconds by SKU; instance-hours; storage GB-mo; egress GB; IP/API extras	Scheduler/control-plane allocation ledger, cross-checked vs. node telemetry
Token API	Input tokens; output tokens; cached-input tokens; batch vs. interactive; per-model rates; images/audio	API gateway request logs with tokenizer counts, cross-checked vs. serving-engine counters
Both	SLA downtime (credit meters); committed-spend drawdown; free-tier consumption	Incident system; billing ledger

The reference pipeline has five stages, and modern usage-billing platforms (Lago, Metronome, Orb, and in-house equivalents at scale players) all converge on the same shape: (1) event capture at the source of truth, emitting one immutable usage event per billable action with customer ID, resource, quantity, and timestamp; (2) ingestion through a durable, idempotent stream — every event carries a unique transaction ID so retries, replays, and race conditions cannot double-bill (production systems ingest up to ~1M events/second); (3) aggregation into billable metrics per customer per period (SUM of GPU-seconds, SUM of output tokens, MAX of concurrent instances, COUNT of requests); (4) rating against the versioned price book; and (5) storage of both raw events and rated results, because disputes are settled by replaying raw events, not by arguing about aggregates.

Where metering breaks, in practice: late events (a node agent buffers during a network partition and flushes after the billing period closed — a documented late-event policy and a re-rating window are required); duplicate events (idempotency keys must be enforced at ingestion, not trusted from emitters); clock skew between nodes assigning events to the wrong period; lost events when agents crash — which is why capacity metering should be derived from the allocation ledger (intent) and reconciled against node heartbeats (reality), never from a single fragile source; orphaned resources that keep metering after the control plane thinks they are gone (or worse, stop metering while still consuming power); and tokenizer drift in the token business — the tokenizer version is billing-critical configuration, because a silent tokenizer change re-prices every request by a few percent in one direction or the other. Every one of these failure modes has a matching revenue-assurance control below.

15. Billing: Rating, Invoicing, and the Contract Layer

Rating converts aggregated usage into money: quantity x applicable rate, walked through the discount waterfall (list → volume/tier → negotiated rate → committed-spend discount → promotional credits) in a deterministic, versioned, replayable order. The non-negotiable engineering property is reproducibility: given the same events and the same price-book version, the rating engine must produce the same invoice, byte for byte, months later. Without that, disputes and audits become archaeology.

Invoicing then assembles rated usage into the customer-facing document: usage lines by SKU and model, committed-fee lines, prepaid-credit drawdowns, SLA credits, proration for mid-period starts and stops, and taxes (which, for compute sold across borders, are a genuine sub-specialty — place-of-supply rules, VAT reverse charges, and cross-border variation all attach at invoicing). Dunning — the retry-and-escalate machinery for failed payments — matters enormously in the self-serve token business where cards fail constantly, and barely at all in the enterprise capacity business where the risk is concentrated slow-pay instead.

The contract layer adds structures that pure usage billing does not have. Take-or-pay invoices bill the committed quantity monthly or quarterly regardless of consumption, with usage above commitment billed as overage and usage below simply forfeited (or, in some contracts, banked into a limited rollover — a term with direct revenue-recognition consequences). Committed-spend deals invert this: the customer prepays or commits to $X, and every month's rated usage draws the balance down, with true-up invoicing if the term ends short. Milestone billing governs new-build clusters: invoicing begins per contract when capacity passes acceptance tests, so delivery-date slippage is directly a revenue event. SLA credits flow the other way — downtime beyond the committed availability generates credits against the next invoice, making incident metering a billing input. Each of these structures is a place where the invoice can drift from the contract, and each therefore appears again in the revenue-assurance control set below.

16. Settlement and Revenue Recognition

Cash settlement

Settlement is the unglamorous conversion of invoices into cash, and its texture differs completely across the two lines. The token/self-serve business settles like consumer SaaS:

card-on-file or prepaid credits, processor fees of 2–3%, chargebacks, and fraud — of which the sector's signature is stolen-card GPU consumption (often for crypto mining or resale), where the provider eats both the chargeback and the burned capacity. Prepaid credits deserve special respect: they are customer liabilities (deferred revenue) until consumed, they age, and their breakage (expiry unconsumed) is recognizable revenue only under defined conditions. The enterprise capacity business settles like industrial leasing: invoices on net-30/60 terms, wires, occasional letters of credit or parent guarantees on large take-or-pay books, and DSO management. Concentration is the defining risk: when a handful of AI labs and hyperscalers are most of the backlog — CoreWeave's disclosed counterparties include Meta (a $21B commitment), Anthropic, Cohere, Jane Street, and Mistral — receivables risk is really single-name credit risk, and the finance function should treat it with credit-committee discipline, not accounts-receivable routine.

Two further settlement webs surround the core. Channel settlement: capacity and tokens sold through hyperscaler marketplaces, brokers, aggregators, and resellers arrive net of channel fees (marketplace listing fees, reseller margins), requiring reconciliation of the platform's settlement report against internal metering — an inter-company version of the meter-to-bill problem. Supplier settlement: on the cost side sit take-or-pay obligations of the platform's own — colocation and power contracts with minimum commitments, and revenue-share arrangements with datacenter partners — so the provider is simultaneously a seller and a buyer of committed capacity, and a mismatch in tenor between the two books (long supplier commitments against short customer contracts, or vice versa) is a structural P&L risk that belongs on the CFO's dashboard.

Revenue recognition (ASC 606 view)

Usage revenue — on-demand GPU-hours and tokens — is the easy case: recognized as consumed, in the period metered. Take-or-pay capacity is recognized ratably as the stand-ready obligation is satisfied, i.e., as capacity is made available, whether or not the customer uses it; unbilled committed amounts sit in remaining performance obligations (RPO), and amounts billed or prepaid ahead of availability sit in deferred revenue. This is why backlog is the sector's favorite disclosure — CoreWeave's $99.4B revenue backlog is contracted future revenue awaiting delivery — and also why it deserves scrutiny: backlog converts to revenue only as fast as data centers are energized and clusters pass acceptance, and it can be repriced, renegotiated, or (in credit stress) defaulted. Variable consideration cuts revenue at recognition, not just at invoicing: expected SLA credits and expected credit-

memo rates should be estimated and netted, so a provider with chronic incidents recognizes less revenue than its meters would suggest. Prepayment breakage, milestone acceptance timing, and rollover rights in take-or-pay deals are the recurring judgment areas an auditor will probe.

17. Revenue Assurance: Where the Money Leaks and the Control Framework

Revenue assurance answers two questions continuously: is everything delivered being billed (completeness), and is everything billed correct (accuracy)? Telecom, which invented the discipline, consistently found 1–3% of revenue leaking before controls; usage-metered compute has the same anatomy — high-volume events, complex rating, negotiated exceptions — and no reason to expect a better baseline. At a 56% EBITDA margin, recovering one point of revenue leakage is worth roughly two points of EBITDA growth for zero incremental COGS.

Where the money leaks

Leakage mode	Mechanism	Typical detection
Unmetered usage	Agent crash, orphaned allocation, event loss in pipeline	Capacity ledger vs. billed-hours reconciliation
Meter-to-bill gap	Events captured but dropped/late in aggregation or rating	Event-count completeness checks per stage
Rating errors	Wrong/stale price book, expired discount still live, FX error	Invoice re-rating; rate-table change audit
Entitlement drift	Free tier uncapped, internal/test usage on prod SKUs	Entitlement audit; margin-by-customer outliers
SLA over-crediting	Credits granted beyond contractual formula	Credit memo vs. incident-record match
Fraud	Stolen cards, abuse of free tiers, resale of discounted capacity	Velocity checks, KYC, anomaly detection
Contract leakage	Overage never billed; rollover mis-tracked; milestone billing missed	Contract-to-invoice audit per account
Tokenizer/meter drift	Counting change silently re-prices all requests	Golden-request regression tests on billing

The control framework

The backbone control is the three-way match, run daily and automated: (1) what the control plane says was allocated (the intent ledger), against (2) what telemetry says

actually ran (node heartbeats, serving-engine counters), against (3) what the billing system rated and invoiced. Discrepancies land in a workflow queue with materiality thresholds, root-cause categories, and closure SLAs — exactly like a trading firm's breaks process. Around that backbone: completeness checks that event counts reconcile across every pipeline stage; re-rating audits that replay a sample of invoices from raw events against the price book; change control on rate tables and tokenizer versions with four-eyes approval and effective-dating; entitlement reviews that catch uncapped free tiers and forgotten discounts; margin-by-customer monitoring, because a customer whose gross margin quietly went negative is either mis-rated or mis-contracted; and credit-memo governance, since discretionary credits are the softest channel through which revenue leaks. The token business adds golden-request tests: a fixed corpus of requests run through the full meter-rate-bill path on every deploy, with billed amounts asserted to the cent.

KPIs for the revenue-assurance function

KPI	Definition	Healthy zone
Leakage rate	Identified leakage / total revenue, annualized	< 0.5% mature; 1–3% typical at start
Meter-to-cash yield	Cash collected / value of metered usage at contract rates	> 98%
Billing accuracy	1 − (credit memos from billing error / billed revenue)	> 99.5%
Recovery rate	Leakage recovered / leakage identified	> 70%
Dispute rate & aging	Disputed invoice value / billed; days to resolve	< 1%; < 30 days
DSO by segment	Receivables days, split enterprise vs. self-serve	Enterprise < 60; self-serve < 5
Unbilled usage aging	Metered-but-uninvoiced value by age bucket	No bucket > 1 cycle
SLA credit ratio	SLA credits / revenue	< 0.5% and matched to incidents

18. Cost per Token, Utilization, and Capacity Alignment

- **Master metric:** revenue-weighted utilization of commissioned capacity, reviewed weekly with Platform and Infrastructure leadership; target trajectory from 50% at commissioning to 80%+ within four quarters of each tranche.
- **Cost per token:** managed as a living model per model-class: GPU-hour cost (depreciation, power at Kutch solar rates, network, facilities) divided by tokens per GPU-hour (batching, quantization, GB300 inference efficiency). Every serving-stack

optimization the platform team ships is commercial margin; the CCO org funds and prioritizes that roadmap jointly.

- **Load shaping:** off-peak and batch pricing moves elastic workloads (training, batch inference, embeddings) into utilization valleys; edge pricing captures premium for latency instead of leaving it on the table.
- **Capacity gating:** a single capacity ledger, jointly owned with Infrastructure, gates every committed-capacity contract against the commissioning curve plus a reserve buffer, so the sales organization can sell aggressively without ever selling capacity that does not exist.

Part V — Risk Management: Compute Futures and Hedging

Reliance Intelligence is building one of the largest AI compute platforms in the world: a targeted 120 MW of compute at Jamnagar on NVIDIA GB300 systems, backed by a ₹10 trillion, seven-year infrastructure commitment. Managing the rapid depreciation of that hardware is the critical financial hurdle sitting alongside the commercial plan.

GPU infrastructure has a useful life of only four to six years. An idle or underpriced GPU-hour is not deferred revenue — it is destroyed revenue, exactly like an empty airline seat at departure. The asset keeps depreciating whether or not it is sold, so the entire commercial question is how much certainty can be attached to future GPU-hours before they perish.

The emerging compute futures market — contracts listed by CME Group and ICE referencing GPU rental indices (H100, B200 and successors) — gives an owner of physical capacity four distinct mechanisms: hedge revenue risk, de-risk capital financing, stabilize the capacity strategy, and govern pricing against a public benchmark.

What the contracts look like

These are early-stage, index-referenced instruments, not a mature futures curve like oil or interest rates — treat the specifics below as illustrative of the contract shape, not as confirmed terms, and validate against the exchanges' current listings before any position is sized.

Feature	Illustrative shape
Underlying	A published GPU rental-rate index (e.g., H100-equivalent $/GPU-hr), aggregated across a panel of reporting providers
Settlement	Cash-settled against the index at expiry — no physical GPU delivery
Tenor	Near-dated contracts most liquid (1–3 months); longer tenors thin and widening in spread
Contract size	Notional GPU-hours per contract, analogous to how power futures quote in MWh
Margining	Standard exchange variation margin, marked to market daily

19. Hedging Revenue Risk Against Price Crashes

As a massive builder and owner of GPU infrastructure, Reliance is naturally long physical compute. That exposure is highly vulnerable to falling rental prices and dropping utilization.

- **The risk.** When next-generation chips scale, or when more efficient models, quantization and serving software reduce the compute required per unit of output, GPU rental prices sag dramatically — often faster than the depreciation schedule assumes.
- **The futures solution.** Sell compute futures (a short position) against uncommitted on-demand and spot capacity. If physical rental rates drop, gains on the short futures offset the lower rental revenue collected on the physical fleet, stabilizing the P&L.
- **What stays unhedged.** Take-or-pay reserved capacity is already contractually fixed; hedging it again converts a stable book into a speculative one. Hedge the merchant tail, not the contracted core.

The offset, in one line

Hedge P&L = hedged GPU-hours x (strike − settled index). A short at $2.10 against a market that settles at $1.70 pays $0.40 per hedged GPU-hour — precisely the revenue the physical fleet failed to earn.

20. De-Risking Capital Financing and Lowering Capital Costs

Financing a gigawatt-scale data center buildout requires massive debt underwritten by lenders who look closely at debt service coverage.

- **The risk.** Lenders are hesitant to finance volatile, rapidly depreciating technology assets carrying unhedged merchant exposure. Unhedged spot revenue is discounted heavily or excluded from the borrowing base entirely.
- **The futures solution.** A liquid forward curve gives the market a public, transparent, tradable reference price for future compute. Lenders can value future revenue streams against that curve instead of a management forecast.
- **The outcome.** Executing (or covenanting) hedges converts merchant revenue into predictable cash flow, improves modelled DSCR, lowers borrowing cost, and makes capital more comfortable underwriting the expansion.

Financing input	Unhedged merchant capacity	Hedged merchant capacity
Revenue basis for lenders	Management forecast, heavily discounted	Exchange forward curve, mark-to-market
Advance rate on merchant EBITDA	Low or zero	Materially higher
DSCR volatility	Tracks spot price swings	Compressed to the hedged band
Cost of debt	Wider spread for price risk	Tighter spread; risk transferred to the market

21. Stabilizing the 60/70 Capacity Strategy

The core go-to-market strategy is to sell certainty first: anchor 60–70% of Jamnagar capacity in multi-year, take-or-pay reserved contracts and monetize the remainder with on-demand token serving.

Compute futures act as a financial yield-management overlay on that remainder. Instead of aggressively discounting physical contracts or chasing cheap spot volume in price-sensitive local markets, the desk can financially lock target yields on uncommitted capacity — preserving the physical price list while still de-risking the revenue.

Capacity layer	Share	Revenue certainty	Futures role
Reserved / take-or-pay	60–70%	Contractual	None — already fixed
On-demand & token serving	20–30%	Merchant, volatile	Short futures on the expected sold-hour base
Spot / preemptible	5–15%	Fully merchant	Opportunistic hedging; keep upside optionality

Hedge only the volume the fleet can confidently deliver: hedging above realistic sold hours converts a hedge into a naked short.

22. Price Governance and Deal Structuring

In a fragmented, opaque physical cloud market, buyers and providers have historically negotiated blind. Exchange-traded benchmarks change that.

- **A standardized reference.** Public price discovery from H100 / B200 rental index futures gives the commercial deal desk an external benchmark to price against, instead of anecdote and last-quarter's win rate.

- **Structured physical contracts.** Write index-referenced deals: floors, collars, and expansion options structured as capacity calls, with the premium priced off the forward curve.
- **Margin floor discipline.** Every discount can be tested against a risk-adjusted market index, so concessions are approved against a hard floor rather than negotiated ad hoc.

Operationally this belongs in the same control fabric as metering and billing — see the revenue-assurance section of Part IV and the leakage controls in the cockpit (Part VII).

23. Hedge Mechanics — Worked Example

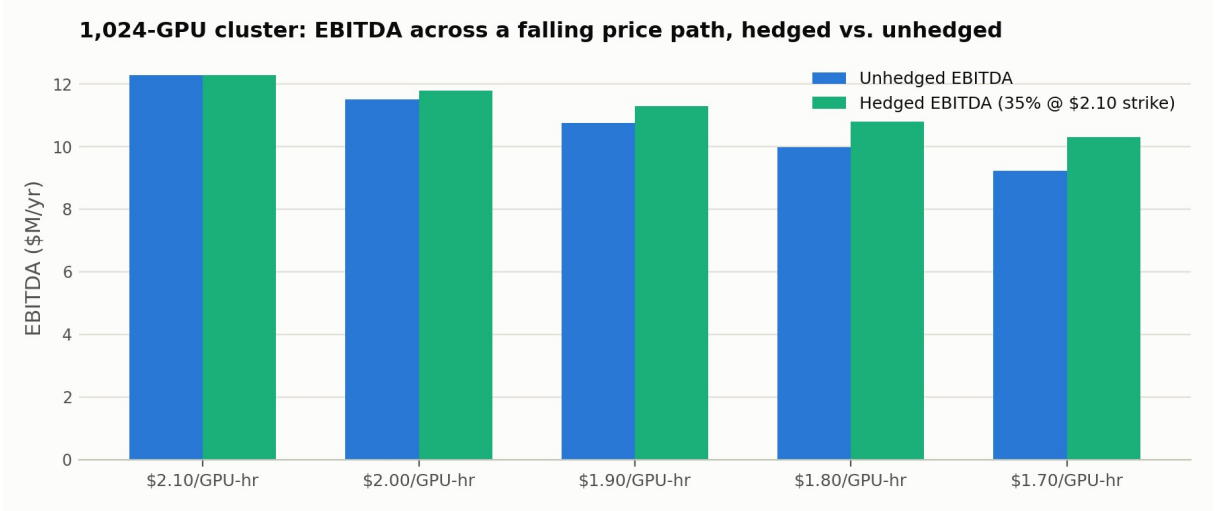
Take one 1,024-GPU cluster at 85% utilization: 1,024 x 8,760 x 0.85 ≈ 7.6M sold GPU-hours a year. Hedging 35% of that base is roughly 2.7M GPU-hours short at a $2.10 strike.

Step	Calculation	Value
Capacity hours	1,024 GPUs x 8,760 hrs	8.97M GPU-hrs
Sold hours @ 85% util	8.97M x 0.85	7.62M GPU-hrs
Hedged notional @ 35%	7.62M x 0.35	2.67M GPU-hrs
Spot falls $2.10 → $1.70	7.62M x $0.40 lost	-$3.05M revenue
Futures settlement	2.67M x $0.40 gained	+$1.07M
Net revenue impact	-3.05 + 1.07	-$1.98M (35% offset)

A 100% hedge ratio would fully neutralize the price move — and equally forfeit the upside if spot rallies. Hedge ratio is the policy dial.

24. The 1,024-GPU Sensitivity Model

Unit economics and P&L sensitivity of a single 1,024-GPU cluster, showing exactly how a price drop from $2.10 to $1.70 per hour hits cash flow and debt payback — and how much of that a futures overlay claws back, at 85% utilization with 35% of sold hours hedged at a $2.10 strike.



Unhedged vs. hedged EBITDA for a 1,024-GPU cluster across the $2.10 → $1.70 price path (85% utilization, 35% hedge ratio, $2.10 strike). The hedge overlay recovers roughly a third of the EBITDA erosion at the bottom of the path without touching the contracted core.

Spot $/GPU-hr	Revenue	EBITDA	Hedge P&L	Hedged EBITDA
$2.10	$18.4M	$12.3M	$0.0M	$12.3M
$2.00	$17.5M	$11.5M	$0.3M	$11.8M
$1.90	$16.6M	$10.8M	$0.5M	$11.3M
$1.80	$15.7M	$10.0M	$0.8M	$10.8M
$1.70	$14.8M	$9.2M	$1.1M	$10.3M

Illustrative synthetic economics for one 1,024-GPU cluster; not any specific company's numbers.

25. Risks and Limits

- **Basis risk.** The index settles on a reference GPU class and geography; realized Jamnagar rates differ by SKU mix, contract type and power cost. The hedge offsets the index move, not the exact realized delta.
- **Liquidity.** These contracts are new. Depth at long tenors is thin, so size positions to what can be exited, and stagger tenors rather than concentrating in one expiry.
- **Margin and collateral.** Short futures require variation margin. A spot rally is a P&L win on the physical fleet but a cash drain on the hedge — treasury must fund that timing mismatch.
- **Accounting.** Hedge accounting designation determines whether settlement flows through OCI or hits earnings each period; document the hedge relationship before the first trade.

- **Governance.** Hedging is a risk-transfer policy, not a trading desk. Board-approved limits on ratio, tenor and counterparty keep it that way.

Part VI — Organization and Execution

26. Organization Build

Five pillars under the CCO, each with a named leader by day 90 and succession depth by month 12.

Pillar	Mandate	First hire profile
OSS/BSS & Billing Platform	Metering, rating, invoicing, settlement; the system of record	Telecom BSS or cloud-billing platform leader
Enterprise Commercial & Sales	Anchor tenants, enterprise ARR, deal desk, pricing execution	India enterprise-cloud sales leader with BFSI depth
Marketplace & Partner Ecosystem	Model/agent marketplace, ISVs, telecom partners, rev-share ops	Platform-ecosystem builder from cloud marketplace background
FinOps & Revenue Assurance	Cost/token model, leakage controls, utilization analytics, audit	Revenue-assurance or trading-controls leader
Customer Success & Capacity Mgmt	Onboarding, adoption, renewals, capacity ledger, NRR	Consumption-cloud CS leader comfortable with capacity planning

Illustrative year-one headcount plan

Headcount builds behind revenue, not ahead of it, with the exception of the OSS/BSS and Revenue Assurance pillars, which are deliberately front-loaded because the billing platform must be live before external SKUs generalize beyond the first anchor contracts.

Pillar	Q1 (leader + core)	Q2	Q3	Q4 (year-end)
OSS/BSS & Billing Platform	1 + 3	+4	+3	~14
Enterprise Commercial & Sales	1 + 2	+5	+6	~19
Marketplace & Partner Ecosystem	1 + 1	+2	+3	~9
FinOps & Revenue Assurance	1 + 2	+3	+2	~11
Customer Success & Capacity Mgmt	1 + 1	+2	+4	~11
Total CCO organization	13	+16	+18	~64

Illustrative build; the Q1 core team is intentionally lean — five named leaders plus nine early hires — so the first 90 days (Section 27) can move on decisions rather than wait on headcount.

Culture: a build-from-scratch commercial organization with desk-level P&L discipline — weekly revenue-weighted-utilization reviews, monthly pricing council, quarterly renewal and margin reviews with Finance, and playbooks written down from the first deal onward so the organization scales on documented practice rather than heroics.

27. First 90 Days and Year-One Roadmap

Horizon	Commercial priorities
Days 0–30	Capacity ledger agreed with Infrastructure; internal captive-demand MOUs scoped; pricing architecture v1 and margin floors approved; billing-platform build/buy decision framed; top-8 anchor pursuit list agreed with CEO
Days 31–60	First two anchor-tenant term sheets in negotiation; deal desk and discount governance live; OSS/BSS vendor selected and metering spec frozen; leadership hiring for all five pillars launched
Days 61–90	Captive MOUs signed at transfer prices; enterprise tiering and packaging published; revenue-assurance control framework documented; developer self-serve tier in design; board readout: contracted-capacity coverage vs. commissioning curve
Quarters 2–4	60%+ of Wave-1 capacity under contract before commissioning; token service GA with 20+ enterprise logos; fine-tuning and private endpoints launched; edge pilot with two real-time customers; marketplace beta with 10+ listed models; exit the year at a $100M+ ARR run-rate with leakage under 0.5% and documented playbooks for every motion

28. KPIs and Commercial Governance

Dimension	Primary metrics
Revenue	Contracted ARR; billed revenue; committed vs. on-demand mix; ARPA by segment; NRR (target 120%+)
Utilization	Revenue-weighted utilization; contracted-capacity coverage of commissioning curve; peak/off-peak load ratio
Margin	Cost per token by model class; gross margin per SKU; discount depth vs. policy; margin floor exceptions
Assurance	Revenue leakage rate (target <0.5%); metering-to-invoice reconciliation breaks; DSO; settlement accuracy
Ecosystem	Marketplace GMV and platform take; active developers; partner-sourced revenue share

Governance cadence: weekly utilization and pipeline review; monthly pricing council and revenue-assurance report; quarterly board pack covering contracted coverage, margin trajectory, and renewal health — the same rhythm a desk head and CRO runs, applied to a token P&L.

Sample monthly business review agenda

Segment	Content	Owner
Scorecard walk	Mandate scorecard status changes since last month (Part VII, Section 32)	CCO
Pipeline & coverage	Contracted-capacity coverage vs. commissioning curve; anchor pursuit status	Enterprise Commercial lead
Revenue assurance	Leakage identified/recovered; open findings aging; control exceptions	FinOps & Revenue Assurance lead
Pricing council read-out	Discount depth vs. policy; margin floor exceptions; rate-card changes	Deal desk
Customer health	NRR bridge; renewal calendar; top accounts at risk	Customer Success lead
Ecosystem	Marketplace GMV, active listings, partner-sourced revenue	Marketplace & Partner lead

29. Risk Register: What Could Break This Plan

A consolidated view of the risks flagged throughout this document, so the leadership team can weigh them together rather than encountering each one in isolation deep inside a section. Likelihood and impact are qualitative judgments to be pressure-tested with Infrastructure, Finance, and Legal in the first 30 days.

Risk	Where it's discussed	Likelihood	Impact	Primary mitigation
Selling ahead of commissioned capacity	Part I, §2	Medium	High	Single capacity ledger jointly owned with Infrastructure; contracts gated to the commissioning curve
Token price commoditization erodes margin	Part I, §3; Part III, §11	High	Medium	Serving-efficiency roadmap funded jointly with Platform; portfolio weighted toward private endpoints and outcome pricing
Hyperscaler India-region price war	Part I, §2	Medium	Medium	Compete on sovereignty, edge latency, and group distribution, not list price
Revenue leakage above 0.5% target	Part IV, §17; Part VII, §32	High (currently occurring)	Medium	Three-way match control, revenue-assurance hire in first 30 days, golden-request billing tests
Committed revenue share stalls below 80%	Part VII, §32	Medium (currently occurring)	High	Renewal motion starting at 60% of term; four near-term renewals prioritized now
GPU price crash faster than depreciation schedule	Part V, §19	Medium	High	Compute-futures hedge on the merchant tail; margin floors in the deal desk
Lenders discount unhedged merchant	Part V, §20	Medium	High	Board-approved hedge ratio policy; forward-curve reference pricing for debt

Risk	Where it's discussed	Likelihood	Impact	Primary mitigation
revenue				underwriting
Commercial-org succession gap (33% ready-now)	Part VI, §26; Part VII, §32	High (currently occurring)	Medium	Named second-in-command for each of the five pillars by month 12
Counterparty credit concentration	Part IV, §17	Low–Medium	High	Credit-committee discipline on anchor accounts; parent guarantees / letters of credit on large take-or-pay books
Billing platform build/buy decision slips	Part VI, §27	Medium	High	Vendor decision framed in days 0–30, selected by day 60, metering spec frozen alongside it

30. First-Week Priorities Checklist

A day-one, week-one action list distilled from the 90-day roadmap in Section 27, for the first CCO staff meeting.

- Confirm the delivered-capacity commissioning curve with Infrastructure and agree the capacity-ledger data feed
- Request the current draft of any internal captive-demand MOUs already in progress with Jio, Retail, and JioStar
- Assemble the top-8 anchor-tenant pursuit list with the CEO's office and prioritize the first two outreach calls
- Pull the current billing/metering vendor shortlist (or confirm none exists) to frame the build/buy decision
- Meet Finance to align on the depreciation-life and financing assumptions used in this document against actuals
- Identify acting or interim leads for each of the five organizational pillars pending permanent hires
- Schedule the first weekly revenue-weighted-utilization review and the first monthly pricing council
- Request read access to whatever revenue, contract, and usage data already exists, however partial

Part VII — Inside the Commercial Cockpit

Everything in Parts I through VI is operationalized in a live Commercial Cockpit — the dashboard built alongside this document at reliancecloud.live — so the mandate is not a slide, it is a system that recomputes against live scenario assumptions (price, coverage, opex, fleet size, hedge ratio and strike) every time an input changes. This part walks the cockpit's seven views and reports the current snapshot against the mandate.

31. Seven Views of the Business

- **P&L cockpit.** KPI tiles, revenue by product line, margin-bridge waterfall, and fleet & token utilization — the same monthly model worked through in Part III, running live from Aug 2025 to Jul 2026.
- **Meter to cash.** The funnel from metered usage through rated, invoiced, and collected revenue; yield trend; and an unbilled-usage aging table — the operational heartbeat of Part IV's settlement discipline.
- **Revenue assurance.** Leakage by mode, identified vs. recovered dollars, and an open-findings table with severity and status — the live instrument panel for the control framework in Part IV, Section 17.
- **Customers & contracts.** Revenue by customer, take-or-pay drawdown, contracted backlog, and a customer-economics table spanning committed, on-demand and token-API accounts.
- **GTM & SKU mix.** The seven-SKU portfolio, segment coverage (contracted vs. consumed), and the commercial targets from Part II tracked against actuals.
- **Token economics.** The node-level margin model from Part III, Section 10, with live sliders on transfer rate, traffic utilization and output price.
- **Compute futures.** The hedging model from Part V, Section 24, with live sliders on spot price, utilization, hedge ratio and strike.

Each cockpit card links back to the relevant section of the P&L Primer, this Strategy document, or the Compute Futures page, so a reviewer can move from a headline number to the underlying assumption in one click — the same discipline this document tries to bring to a 50-page format: every number traceable to its source.

32. Current Performance Snapshot Against the Mandate

The cockpit translates the commercial mandate into a live scorecard of measurable KPIs, each with an owning metric and an evidencing tab. As of the most recent monthly close, the scorecard reads two metrics on track, five on watch, and three off track — a candid, unfiltered view of where the plan is ahead of itself and where it needs executive attention in the first 90 days.

Metric	Current	Target	Status
EBITDA margin	69.4% (+0.6 pts MoM)	≥ 45%	On track
SKU portfolio deployment	7 of 7 launch SKUs live	7 of 7 live	On track
Meter-to-cash throughput	97.0%	≥ 98%	Watch
Price realization vs. list	79.0% ($3.2M non-standard discounts)	≥ 80%	Watch
Cost per million output tokens	$0.63	≤ $0.60	Watch
Token gross margin	71%	≥ 70%	Watch
Order-to-cash SLO compliance	3 of 6 SLOs breaching	6 of 6 met	Watch
Revenue leakage	0.88% ($781K, 31 open disputes)	≤ 0.50%	Off track
Committed revenue share	68% (4 contracts need renewal)	≥ 80%	Off track
Commercial-org succession (ready-now)	33% coverage	≥ 60%	Off track

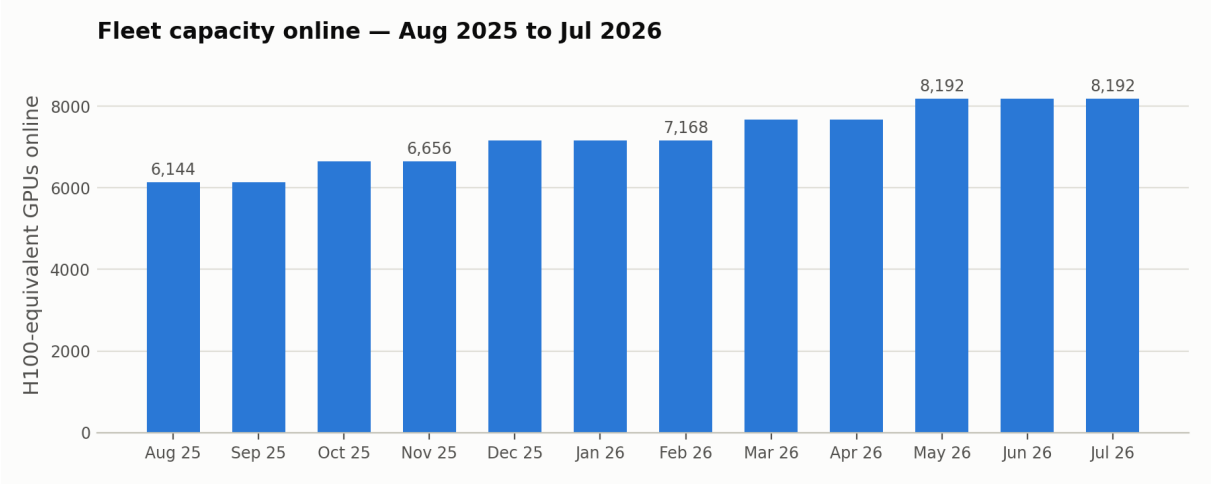
Reading the three off-track lines together tells one connected story rather than three separate problems. Committed revenue share sits 12 points below target because four contracts are approaching renewal without a locked next term — exactly the renewal-motion discipline Part II calls for beginning at 60% of term. Revenue leakage at 0.88% against a 0.50% target is 31 open disputes worth $781K, concentrated in the modes the control framework in Part IV, Section 17 is built to catch: unmetered usage after teardown, expired discounts still live, and batch requests rated at interactive prices. And succession coverage at 33% is a direct read on the organization build in Part VI, Section 26 — three of five pillars still lack a credible second-in-command, which is a single-point-of-failure risk the board pack should carry explicitly until it clears 60%.

None of the watch-list items are alarming in isolation — meter-to-cash at 97.0% against a 98% target, price realization at 79.0% against 80%, and cost per million tokens at $0.63 against a $0.60 ceiling are all within a point or two of target — but together they describe a business that is earning very well (69.4% EBITDA margin, 71% token gross margin) while

still tightening the operational plumbing that Part IV argues is where a mature token business actually protects its margin over multiple years, not just this quarter.

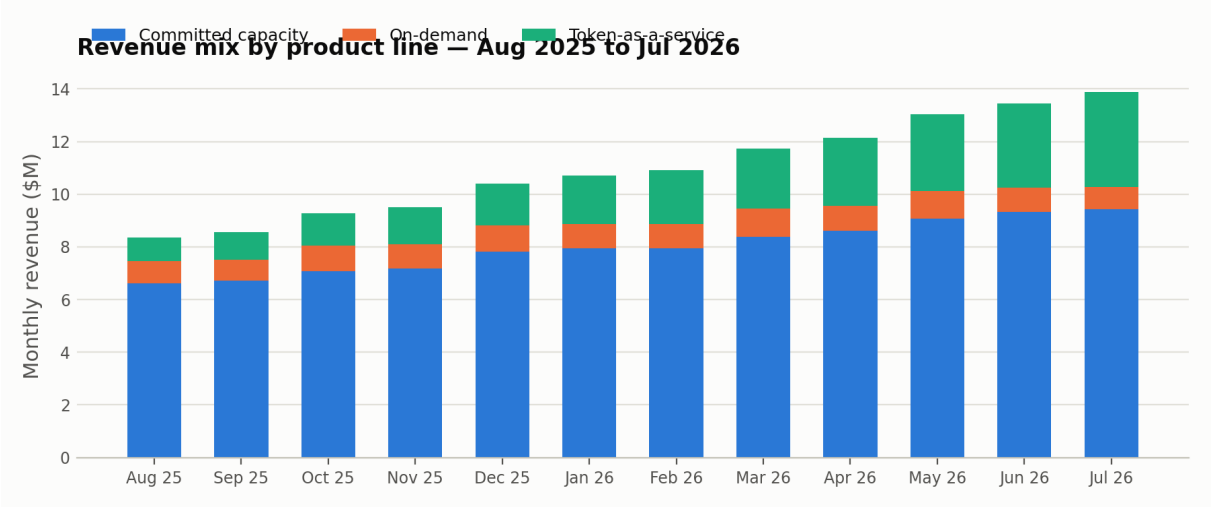
33. The Twelve-Month Build: Fleet, Coverage, and Revenue Mix

The cockpit's base-case model runs the fleet from 6,144 GPUs online in August 2025 to 8,192 by April 2026 and held through July 2026, with committed-capacity coverage climbing from 72% to 77% over the same period — both consistent with the Wave-1 target of 60–70% minimum coverage, tracking comfortably ahead of the floor.



Fleet capacity online, August 2025 through July 2026 — three commissioning steps take the platform from 6,144 to 8,192 GPUs across the model year.

Revenue mix over the same period shows committed capacity carrying the large majority of the base, on-demand providing a thin, roughly stable overlay, and token-as-a-service growing from under $1M to $3.6M a month as the token business ramps behind the capacity anchor — precisely the sequencing Part II's SKU roadmap calls for: sell certainty first, let the token layer's yield-management role scale in behind it.



Monthly revenue by product line. Committed capacity (blue) anchors the base; on-demand (orange) is a stable overlay; token-as-a-service (aqua) more than triples across the model year as Wave-1 and Wave-2 SKUs ramp.

34. Customers, Contracts, and Concentration

The customer book today is concentrated in a small number of committed accounts, exactly as the anchor-tenant motion in Part II, Segment 2 intends for year one, with a growing token self-serve tail that will diversify the base over time.

Customer	Segment	Revenue ($M)	Gross margin	Committed drawdown
Atlas Labs	Committed	63.4	72%	47%
Helios AI	Committed	34.1	70%	52%
Token self-serve (2,410 accounts)	Token API	24.7	61%	3%
Vector Capital	Committed	12.8	68%	64%
Nimbus Research	On-demand	7.9	55%	31%
Corte Systems	Committed	6.2	66%	44%
Other	Mixed	13.9	52%	28%

Take-or-pay drawdown for the four largest committed accounts, and their renewal calendar, is the single most important input to the committed-revenue-share KPI above.

Customer	Committed ($M)	Drawn ($M)	Renewal
Atlas Labs	52.0	47.9	Q2 FY27
Helios AI	28.6	27.1	Q4 FY26
Vector Capital	11.0	8.4	Q1 FY27

Customer	Committed ($M)	Drawn ($M)	Renewal
Corte Systems	5.5	5.3	Q3 FY26

Helios AI and Corte Systems renew within the next two quarters — first-order priorities for the renewal motion described in Part II, Section 5.

35. Revenue Leakage — Open Findings

The leakage identified by mode in the trailing period, aging of unbilled usage, and the current open-findings queue give the CCO organization a concrete first-30-days worklist consistent with the control framework in Part IV.

Leakage mode	Identified ($K, TTM)
Unmetered usage	1,850
Rating errors	1,420
Entitlement drift	960
Contract leakage	780
Fraud & chargebacks	610
Meter-to-bill gap	520
SLA over-crediting	340
Tokenizer drift	180

Finding	Mode	Severity	Value	Status
Orphaned A100 pool metering after teardown	Unmetered usage	Critical	$212K	Open
Expired FY25 discount still applied — Vector Capital	Rating errors	Serious	$148K	In review
Batch-API requests rated at interactive price	Rating errors	Serious	$121K	Fix deployed
Free-tier org spawning linked accounts	Fraud & chargebacks	Warning	$64K	Open
SLA credit granted beyond formula — Mar incident	SLA over-crediting	Warning	$38K	Recovered
Late-event window misconfigured on node agents v2.4	Meter-to-bill gap	Good	$17K	Closed

The two critical/serious findings — orphaned A100 metering ($212K) and the Vector Capital expired discount ($148K) — account for over 45% of currently open leakage value and are the natural first targets for the revenue-assurance hire in the first 30 days.

36. What Tomorrow's Discussion Should Resolve

Three decisions would move this from a strategy document into an executing organization: confirming the capacity ledger and commissioning curve with Infrastructure so the 60–70% coverage target can be gated correctly; agreeing the top-eight anchor pursuit list so the CEO-to-CEO motion in Part II, Segment 2 can start immediately; and approving the OSS/BSS build/buy decision so the billing platform — the product this plan treats as equal in priority to pricing — has an owner and a timeline inside the first 60 days.

Part VIII — Closing

37. Why This Plan, Why Me

Every element of this plan is drawn from businesses already run, not businesses studied. Consumption pricing and committed-capacity contracts come from a $250M+ ARR GPU and foundation-model business built at AWS. Client-franchise building and multi-million dollar desk P&L ownership come from eleven years at Morgan Stanley. Settlement, controls, and revenue assurance come from two CRO seats and a derivatives career, where a break in the metering-to-settlement chain is treated exactly as it should be treated here: a P&L event, not an IT ticket. Build-from-scratch commercial construction comes from founding ventures and scaling functions in an emerging market, where the playbook has to be written before it can be followed.

The token business at Reliance Intelligence is the convergence of those disciplines at national scale. It is a leasing business, a telecom, and a trading desk all at once — and the candidate best suited to run it is someone who has already run each of those pieces separately and knows exactly where they interact: where pricing meets the cost stack, where billing meets revenue recognition, where hedging meets financing, and where a five-pillar organization meets a 90-day clock. I would be honored to build it.

38. Anticipated Questions from the Leadership Team

A candid answer to the questions this plan is most likely to draw tomorrow, so the discussion can move quickly to decisions rather than clarification.

"Why not simply discount to match hyperscaler or domestic-neocloud pricing?"

Because price is not the axis Reliance should compete on everywhere. Part I, Section 2 argues the platform's durable advantages are scale-driven cost, sovereignty, and distribution — not being the cheapest GPU-hour in India, which is a race to the bottom against subsidized and subscale competitors alike. The pricing architecture in Part II instead uses committed-capacity discounts to reward certainty, reserves list-price discipline for genuinely differentiated SKUs (private endpoints, sovereignty premium), and treats aggressive spot pricing as a tool for monetizing otherwise-stranded capacity, not a headline strategy.

"What happens if GB300 delivery or the Jamnagar commissioning curve slips?"

The entire commercial plan is deliberately gated to the delivered-capacity curve, not the announced one — Part I, Section 2 states this explicitly as a discipline requirement, and the capacity-ledger control in Part IV, Section 18 exists precisely to prevent the sales organization from selling capacity that does not yet exist. A slippage does not break the plan's logic; it delays the revenue timeline proportionally, and the 90-day roadmap's first milestone (Section 27) is agreeing that ledger with Infrastructure before any external commercial commitment is made.

"How is this different from just running a cloud business?"

Two ways. First, the two-sided commercial architecture (capacity plus tokens, Part I Section 3) means the business absorbs utilization risk deliberately on one line to earn a materially higher achieved price on the other — a yield-management discipline more common in airlines and telecom than in traditional enterprise IT sales. Second, the billing-as-a-product philosophy (Part IV) and the compute-futures hedging overlay (Part V) both borrow directly from trading-desk and derivatives discipline; this is a commercial P&L that behaves, underneath the product language, like a leasing book with a trading desk bolted on for the merchant tail.

"What is the single biggest execution risk?"

Per the risk register in Part VI, Section 28, the two highest-likelihood risks currently visible are revenue leakage above target and committed-revenue-share renewal risk — both already showing amber in the cockpit's live scorecard (Part VII, Section 32). Both are addressable with the same lever: getting the FinOps & Revenue Assurance and Enterprise Commercial pillars staffed and operating within the first 90 days, which is why organization build is sequenced ahead of, not after, the external commercial launch.

"Why does the plan spend so much time on billing and hedging instead of just selling?"

Because in a usage-metered business, unsold capacity and unbilled usage are both permanently lost revenue, and the two disciplines this plan gives equal weight to sales — metering/billing integrity (Part IV) and futures hedging (Part V) — are the mechanisms that make sure revenue that was earned is actually collected, and that a falling price environment does not undo a good sales quarter. A token business with excellent sales and weak billing discipline is a business that reports strong bookings and mysteriously

disappointing cash collection eighteen months later; this plan is built to avoid that outcome from day one.

"How much of the captive Reliance demand should count toward the revenue targets?"

All of it, priced honestly. Part I's third principle is explicit that captive demand is the launch customer, not the business plan, so internal MOUs are transfer-priced to recover full cost plus a modest margin (Part II, Segment 1) rather than subsidized, and the $100M+ ARR exit target in the Executive Summary is not meaningful unless it holds up under that same pricing discipline applied to Jio, Retail, JioStar, and the five Jio AI services. The risk to manage is the opposite one: under-pricing internal demand to hit a utilization number would flatter year-one optics while quietly undermining the external price list this entire plan is built to defend.

"What is the fastest path to a reference customer we can talk about publicly?"

Segment 2's anchor-tenant motion (Part II) is designed to produce exactly that within the first two quarters — a five-to-eight-name pursuit list, a CEO-to-CEO sales motion, and bespoke structuring modeled explicitly on the precedent Meta's 168 MW Jamnagar lease already set. A signed anchor deal is simultaneously a revenue outcome, a financing input (Part V, Section 20 — lenders value contracted revenue at a materially higher advance rate than a forecast), and the public reference architecture that makes every subsequent enterprise sales conversation in Segment 3 shorter.

Appendix A — Full Twelve-Month Financial Model

The complete monthly build behind every chart and summary table in Part III and Part VII, base case (scenario multipliers all at 1.00x). Utilization is total sold hours (committed plus on-demand plus token-equivalent) as a share of theoretical fleet capacity hours; pre-tax income crosses into positive territory in December 2025 as fleet size and coverage both scale, and the July 2026 EBITDA margin of 69.4% is the figure quoted as the current snapshot in Part VII, Section 32.

Month	GPUs	GPU-hrs (M)	Revenue ($M)	Opex ($M)	EBITDA ($M)	EBITDA %	Dep. ($M)	Interest ($M)	Pre-tax ($M)	Util. %
Aug 25	6,144	4.49	8.35	3.18	5.18	62.0%	4.53	1.27	-0.62	83.3%
Sep 25	6,144	4.49	8.57	3.20	5.37	62.7%	4.53	1.27	-0.43	84.

Month	GPUs	GPU-hrs (M)	Revenue ($M)	Opex ($M)	EBITDA ($M)	EBITDA %	Dep. ($M)	Interest ($M)	Pre-tax ($M)	Util. %
										7%
Oct 25	6,656	4.86	9.29	3.41	5.88	63.3%	4.91	1.37	-0.40	83.8%
Nov 25	6,656	4.86	9.50	3.44	6.07	63.8%	4.91	1.37	-0.22	85.0%
Dec 25	7,168	5.23	10.41	3.64	6.77	65.0%	5.29	1.48	0.00	86.1%
Jan 26	7,168	5.23	10.71	3.66	7.05	65.8%	5.29	1.48	0.28	87.4%
Feb 26	7,168	5.23	10.91	3.69	7.22	66.2%	5.29	1.48	0.45	88.0%
Mar 26	7,680	5.61	11.74	3.90	7.84	66.8%	5.66	1.59	0.59	87.7%
Apr 26	7,680	5.61	12.15	3.94	8.21	67.6%	5.66	1.59	0.96	89.6%
May 26	8,192	5.98	13.03	4.16	8.87	68.1%	6.04	1.69	1.14	89.3%
Jun 26	8,192	5.98	13.46	4.20	9.26	68.8%	6.04	1.69	1.53	91.3%
Jul 26	8,192	5.98	13.88	4.25	9.62	69.4%	6.04	1.69	1.89	92.6%

Illustrative base-case monthly model consistent with the cockpit's default scenario. Depreciation and interest scale directly with fleet size (5-year straight-line depreciation; 70% LTV at 8.0% interest), which is why pre-tax income lags EBITDA growth in the early months of each commissioning step.

Appendix B — Glossary of Terms

Term	Definition
Agent runtime	A SKU billing metered agent-hours and per-task completions for autonomous AI workloads.
ARR	Annual recurring revenue — committed or subscription-like revenue annualized.
ARPA	Average revenue per account.
ASC 606	US GAAP revenue-recognition standard governing when and how much revenue is recognized.
Backlog / RPO	Remaining performance obligations — contracted future revenue not yet recognized.
Basis risk	The risk that a hedge's reference index moves differently from the hedger's actual realized price.
Batch pricing	A discounted rate (typically ~50% off) for token requests that can tolerate latency, sold in exchange for scheduling flexibility.
Cached-input pricing	A steep discount (typically ~90% off list) for input tokens served from a prompt cache rather than freshly processed.

Term	Definition
Committed-spend deal	A contract where the customer commits to $X over a term, drawn down against usage at a discounted rate.
Cost per token	All-in GPU-hour cost divided by tokens served per GPU-hour; the core unit-economics metric for the token business.
Deal desk	The function holding discount and pricing-exception authority above defined thresholds.
Deferred revenue	Cash received or billed for goods/services not yet delivered; a liability until recognized.
DSCR	Debt service coverage ratio — cash flow available to cover debt payments.
DSO	Days sales outstanding — average days to collect receivables.
Dunning	Automated retry-and-escalation process for failed or overdue payments.
Golden-request test	A fixed corpus of billing requests replayed through the full meter-rate-bill path on every deploy to catch pricing regressions.
GMV	Gross merchandise value — total value transacted through the marketplace.
GPU-hour	One GPU running for one hour; the base unit of capacity pricing and cost.
Hedge ratio	The share of exposure covered by a futures or derivative position.
Idempotency key	A unique identifier on an event that prevents it from being billed twice if retried or replayed.
Meter-to-cash yield	Cash collected as a share of the value of metered usage at contract rates.
NRR	Net revenue retention — revenue from the existing customer base this period vs. the same base last period.
On-demand	Capacity sold without a multi-year commitment, priced at a premium to reserved rates.
Price book	The versioned, access-controlled system of record for every rate, discount, and promotional credit the rating engine applies.
Price realization	Actual achieved price as a share of list price, after discounts.
PUE	Power usage effectiveness — total facility power divided by IT-equipment power; lower is more efficient.
RAG status	Red/Amber/Green — a simple status classification; this document uses On track / Watch / Off track.
Rating (billing)	Converting metered usage into a priced line item using the price book.
Revenue assurance	The discipline of ensuring all usage is billed and all billing is correct.
Revenue-weighted utilization	Fleet utilization weighted by the revenue each hour of use generates, not just raw occupancy.
RPO	See Backlog.
SKU	Stock-keeping unit — here, a distinct commercial product (e.g., Reserved GPU Capacity, Token-as-a-Service).
SLA / SLO	Service-level agreement / objective — a committed or targeted operational standard, e.g., uptime.
Stand-ready obligation	A commitment to make capacity available regardless of whether the customer uses it — the basis for ratable revenue recognition on take-or-pay contracts.

Term	Definition
Take-or-pay	A contract obligating the customer to pay for committed capacity whether or not it is used.
Three-way match	Reconciliation of allocation ledger, telemetry, and billed invoice — the core revenue-assurance control.
Token	The base unit of language-model input/output; priced per million tokens.
Traffic utilization	The share of a token-serving node's theoretical throughput actually sold, as distinct from raw GPU occupancy.
Transfer price	The internal price charged between business units for the same underlying resource.
Variable consideration	Revenue-recognition adjustment for expected discounts, credits, or refunds.
Yield management	Dynamically pricing a perishable, capacity-constrained asset to maximize revenue — the discipline this plan borrows from airlines for GPU-hours.

Appendix C — Data Tables and Methodology Note

All illustrative dollar figures, unit economics, and worked P&Ls in this document are modeled estimates built from public market benchmarks as of mid-2026 (H100/H200/B200 rental pricing indices, LLM API rate cards, and CoreWeave's public disclosures), calibrated to the assumptions stated at each worked example — not Reliance's internal cost or contract data. They are opening hypotheses for diligence, explicitly built to be pressure-tested and replaced with actuals in the candidate's first 30 days, as flagged in Part III, Section 11. The live Commercial Cockpit dashboard (reliancecloud.live) runs the same assumption set as a working model rather than a static document, so any input — price, coverage, opex, fleet size, hedge ratio, or strike — can be changed and the full P&L, meter-to-cash, revenue-assurance, customer, and GTM views recompute immediately.

Core rate assumptions used throughout

Assumption	Value
Committed capacity rate	$2.05 – $2.10 / GPU-hr
On-demand rate	$2.60 / GPU-hr
Cash operating cost	$0.485 – $0.49 / GPU-hr
Depreciation life	5 years, straight-line
Financing	70% loan-to-value at 8.0% interest
All-in cluster capex	$40,000 – $45,300 per GPU
Base fleet size (model year)	6,144 → 8,192 GPUs

Appendix D — Sources

- CoreWeave Q1 2026 earnings press release (SEC EDGAR): revenue, adjusted EBITDA, backlog, debt, capex, interest — sec.gov/Archives/edgar/data/1769628/000176962826000220/coreweave1q26earningspress.htm
- GetDeploying H100 cloud price index across 48+ providers — getdeploying.com/gpus/nvidia-h100
- Zettabyte, "The unit economics of debt-financed GPU clouds" — zettabyte.space/blog/neocloud-unit-economics-gpu-cloud
- TLDL verified LLM API pricing tracker (July 2026) — tldl.io/resources/llm-api-pricing
- Introl, "Inference Unit Economics: The True Cost Per Million Tokens" — introl.com/blog/inference-unit-economics-true-cost-per-million-tokens-guide
- Lago, "How to architect billing systems to power usage-based pricing" — getlago.com/blog/architect-billing-systems

Worked examples are illustrative constructions from the benchmarks above, not any specific company's economics. This document is general information, not investment, accounting, or legal advice.

Appendix E — Document Map

This document, the P&L Primer, the Compute Futures page, and the live Commercial Cockpit are one connected reference set. Part I draws on the cockpit's executive framing; Part III is the P&L Primer's Sections 1–5 and 10 restated as strategy; Part IV is the Primer's Sections 6–9; Part V is the Compute Futures page in full; Part VII is a direct tour of the cockpit's seven tabs and current data. A reader who needs the deeper arithmetic behind any table in this document — the full monthly model, the metering pipeline detail, or the live hedge sensitivity sliders — should open the corresponding page on the cockpit rather than re-deriving it here.

Appendix F — Suggested Agenda for Tomorrow's Discussion

A proposed structure for a 60-minute session, built so the conversation reaches the three decisions in Part VII, Section 36 with time to spare.

Time	Topic	Reference
0–10 min	The opportunity and the five-principle mandate	Part I, Section 1
10–20 min	Commercial architecture: SKUs, pricing, segments	Part II
20–30 min	The financial engine and current live performance snapshot	Part III; Part VII, Section 32
30–40 min	Billing/revenue-assurance discipline and the compute-futures hedge	Part IV; Part V
40–50 min	Organization, 90-day plan, and risk register	Part VI
50–60 min	Three decisions: capacity ledger, anchor pursuit list, billing build/buy	Part VII, Section 36

Appendix G — Index of Tables and Figures

For quick navigation during tomorrow's discussion, every table and chart in this document, in order of appearance.

- Table — Suggested agenda for tomorrow's discussion (Appendix F)

A Closing Note

This document and the live Commercial Cockpit it accompanies were built to be read together and updated together: the numbers here are a snapshot, and the cockpit is the place they keep moving. The intent tomorrow is not to defend every figure as final, but to agree the plan's shape, the three near-term decisions in Part VII, and the mandate the first 90 days should be measured against — and then to start replacing every illustrative assumption in this document with Reliance's own actuals, one validated claim at a time.