

GPU operations for asset owners

Operate GPU capacity, improve service margins and plan its release.

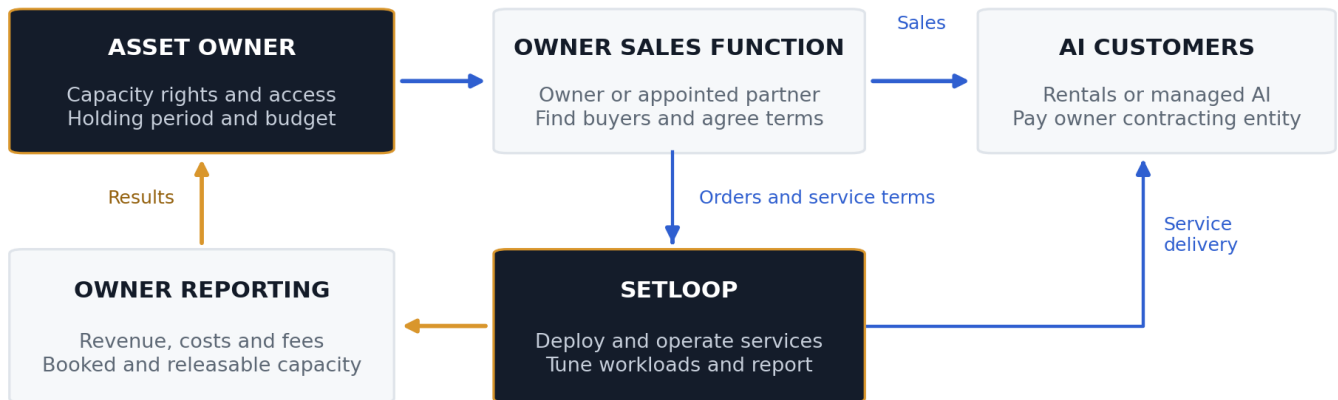
Setloop designs and operates GPU cloud services for organisations that control hardware or contracted capacity. We combine cloud engineering with inference optimisation, LLM security and automated operations. The proposed mandate covers service delivery, operating costs and customer commitments across the owner's holding period.

Compute futures and physical capacity

ICE and Ornn announced cash-settled GPU rental-price futures in May 2026. Architect and Compute Desk announced a network connecting futures positions with capacity delivery in July. Both announcements carry regulatory conditions. Financial exposure alone creates no fleet to operate. [1-2]

Our prospective clients hold usable capacity and the right to sell services from it: asset owners, funds and physical trading desks. Hardware owners may need the full operating service; capacity buyers need a scope that fits their provider's access, resale and support terms.

Customer acquisition and operating responsibilities



Customers pay the owner entity; it pays hosting costs and Setloop fees.

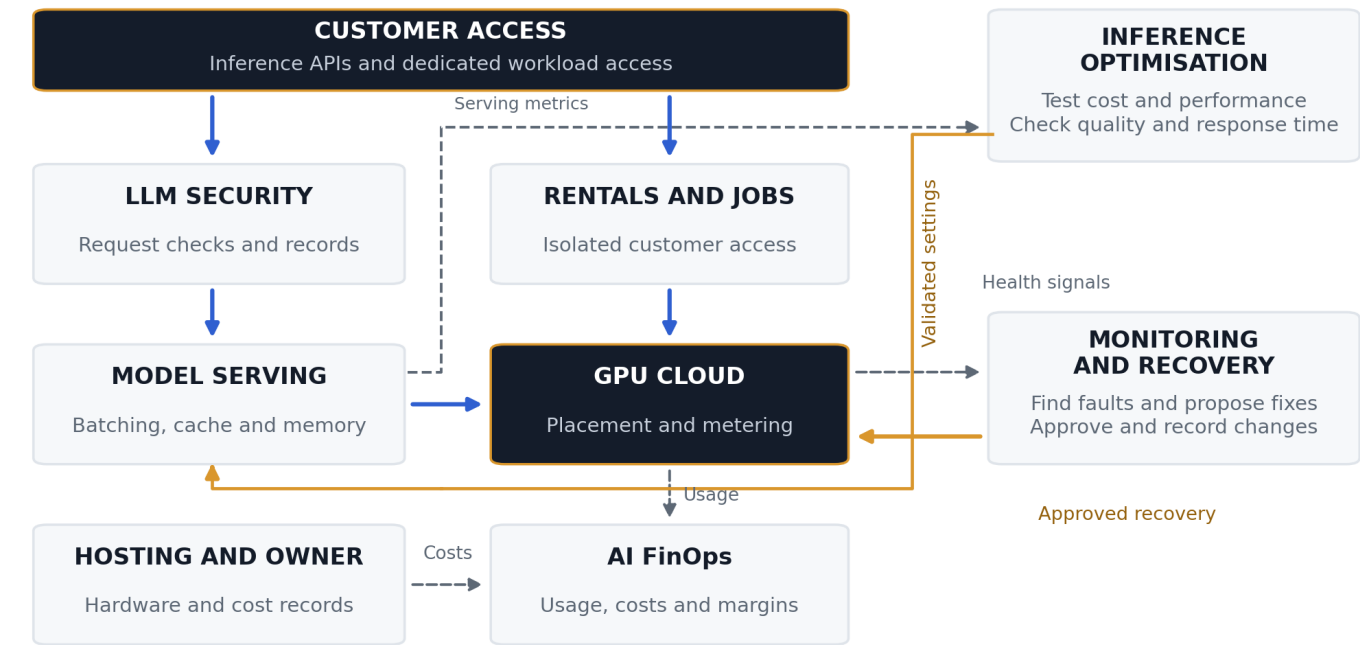
Proposed starting model: the owner or its appointed sales partner supplies qualified demand, signs customer contracts, collects payment and carries credit risk. Setloop supports technical sales and runs the contracted service. Customer sourcing by Setloop would require a separate sales scope.

Fleet and customer requirements

Before launch, confirm GPU inventory, location, permitted use, technical access, hosting costs, buyer demand and the available term. Begin with rentals or a named inference workload. Fine-tuning, rollout generation and RL services follow a committed buyer and a separate delivery assessment.

Control serving costs and service quality

Setloop’s engineering work spans production GPU infrastructure, distributed model serving with vLLM and SGLang, and multi-tenant platforms. This lets us address workload efficiency, isolation and recovery within the same engagement. [8]



Proposed deployment architecture. Blue: service execution. Dashed: measurements and cost records. Amber: tested or authorised changes. Provider responsibilities and integration readiness are confirmed before deployment.

Capability	Owner benefit and delivery scope
GPU cloud services	Provisioning, isolation and billing integration for rentals. Managed inference is a scoped deployment; the published platform lists scalable inference as a roadmap service. [9]
Inference optimisation	Test serving settings to lower cost per token while meeting quality and response-time targets. Optimisation is tailored to the selected model, hardware and traffic. [5]
LLM security	Check model requests for security risks, limit usage and keep audit records. Available through a published product; validate detection rules and performance overhead per deployment. [3]
Automated monitoring and recovery	Connect incident signals, identify likely faults and propose fixes. Begin with monitoring only; agree permissions, human approval and support ownership before enabling changes. [4]
AI FinOps	Reconcile usage and cost. Fleet margins require integration of hosting bills, owner costs, customer receipts and Setloop fees into the reporting scope. [6]

Specialist services: automated model experiments support repeatable evaluation and recovery after interrupted runs. AI security testing examines defined threats and compares defences; its managed platform is in development. Both are separately scoped. [5,7]

Compare earnings with direct rental

Managed services should justify their extra operating work. Compare them with an achievable rental offer for the same hardware, location and available period. Include customer acquisition, setup and release costs before choosing the service mix.

Decision factor	Direct rental	Managed inference or AI
Revenue basis	Contracted GPU hours and collection terms.	Committed usage or service fees; account for traffic variability.
Delivery costs	Hosting, provisioning, support and sales costs.	The same costs plus serving, security, tuning and model support.
Owner flexibility	Rental term and return conditions.	Service term, replacement capacity and customer migration obligations.
Decision measure	Net operating contribution over the available period.	Additional contribution after setup, migration and service costs.

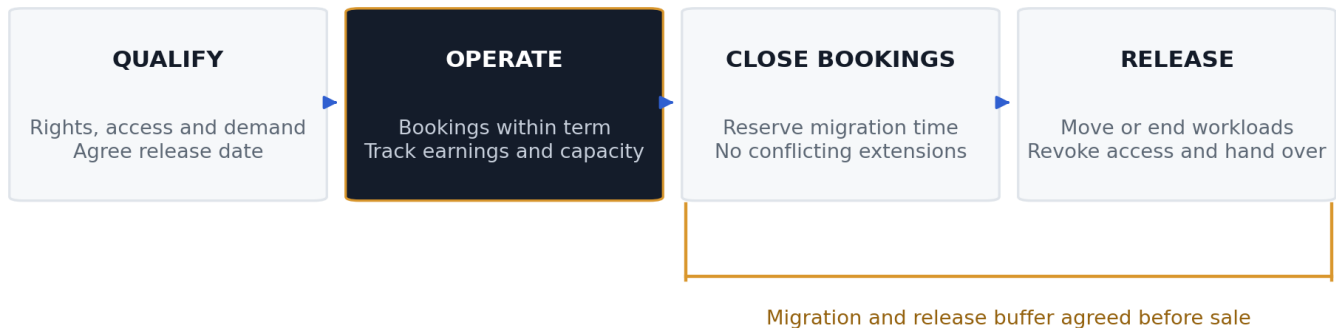
The additional revenue must cover the additional costs

Illustrative arithmetic only: if managed inference adds £5,000 per month in operating costs, £6,000 in setup and £3,000 in release costs over three months, it needs more than £8,000 extra monthly revenue to beat rental. This assumes the same fleet and period, with all other costs unchanged; these figures are not Setloop prices or a forecast.

Owner reporting shows collected revenue less operating costs, setup, agreed fees and release costs. Compare contribution per available GPU-hour, including idle and down hours, within the same GPU class. Report financing, depreciation and resale separately. For inference, also measure cost per million served tokens by model and input/output mix at agreed quality and latency.

Bookings must fit the owner's release date

PROPOSED CAPACITY POLICY



At onboarding, agree the last service date, notice period and release buffer. Reject bookings that exceed that window unless replacement capacity and migration are agreed and funded. Report booked, idle and releasable capacity separately; selling the hardware does not cancel customer obligations.

Before release, test migration or service closure, reconcile billing, remove customer data under the agreed retention policy and revoke access. The contract assigns replacement-capacity costs, migration costs and any early-exit charges. This policy is part of the proposed mandate, not a claim of unrestricted asset liquidity.

Validate one service before expanding

Start with a paid fleet assessment. It determines the supported service, integration work, rental alternative and minimum viable operating term. Proceed to a proposed 60–90 day pilot only when access, hosting, support and customer demand are ready.

Agree the evidence required for acceptance

01 Assess feasibility. Confirm operating and resale rights, fleet condition, committed demand, full costs and release dates. Identify what can be deployed, what needs integration and who delivers each part.

02 Launch one service. Test tenant isolation, metering, billing, incident handling and recovery. For inference, establish quality, latency and cost baselines on representative traffic before tuning.

03 Decide whether to expand. Compare collected revenue and net operating contribution with the rental alternative. Demonstrate the release procedure. Expand only if the agreed service and commercial thresholds are met.

Stop or rescope if access rights fail, isolation or recovery tests fail, costs exceed the agreed limit, or the release date cannot be honoured. Without paying demand, acceptance covers technical readiness and measured delivery cost only; it cannot establish revenue or commercial viability.

Contracted scope and proposed fees

Item	Agreement before deployment
Setup fee	Assessment, integration and acceptance tests; fixed scope and budget.
Operating fee	Named operator, support hours, incident response targets, escalation cover and reporting. The owner funds hardware and hosting.
Facility support	The hosting provider handles power, cooling and onsite repair under an agreed escalation path. Setloop's software responsibilities are listed separately.
Performance fee	Optional. Based on a defined contribution measure and rental baseline, with cost allocation, fee calculation and attribution agreed in advance.

Technical evidence and deployment readiness

Our published products and whitepapers describe the engineering approach. Fleet-specific acceptance requires a working deployment, benchmark records, billing reconciliation and a tested recovery and release procedure. The proposal identifies the delivery team, relevant experience, supported scale and support commitments.

Request a fleet assessment. Send the GPU inventory, locations, access and resale terms, available period, hosting costs and existing customer demand. The assessment defines the service scope, delivery plan and pilot budget.

[Discuss your fleet](#) • sales@setloop.io

Sources

[1] ICE and Ornn • [2] Architect and Compute Desk • [3] LLM security • [4] Monitoring and recovery
 [5] Inference optimisation and model experiments • [6] AI FinOps • [7] AI security testing
 [8] Engineering experience • [9] GPU cloud services • [Setloop whitepapers](#)

Business proposal • 21 September 2026. Commercial terms and the capacity-release policy are proposed. No exchange affiliation, benchmark improvement or investment return is assumed.