RUNMYML · PARTNER BRIEF

GPU capacity partnership and pilot

Renting idle GPUs for training and managed inference

> **RunMyML plans to rent idle GPU capacity and use it to run training and inference for customers. We would operate inference runtimes on partner hardware and support training jobs from existing MLOps tools. We are looking for an infrastructure partner and a real customer workload for a first pilot.**

Suppliers and customers

GPU providers, cluster owners and companies with reserved capacity may have paid capacity available between jobs or outside peak periods. They could offer approved windows to external teams and recover value from hours that would otherwise remain unused.

Customers could bring a training or fine-tuning job, or a model they want RunMyML to run for inference. Training teams could reserve suitable GPU windows; inference customers would agree the model, batch or serving workload, duration and performance requirements.

Proposed service

Suppliers would set the resources, available windows and commercial terms. We would rent agreed GPU capacity for customer workloads, from individual GPUs to multiple nodes. Inference would include operating the model runtime, preparing weights and data, and measuring usage and performance.

Market references include SF Compute for GPU capacity resale and Touchmark for reserved inference. Our proposed service adds the execution and onboarding workflow on partner infrastructure.

- Confirmed bookings would cover an agreed window; interruptible offers would be identified separately.
- Quotes would include GPU time, setup, storage, data transfer and platform charges.
- The supplier would retain control over external access, resource allocation and security requirements. Resale of reserved capacity would require the relevant permissions.

Flexible onboarding and execution

Customers would keep their existing development and experiment workflow. We would start with the connection method that fits the first workload:

- Connect selected experiments through an integration with tools such as W&B, MLflow or ClearML, using available configuration and artifacts and returning logs and metrics.
- Bring a repository or container with a launch command, resource requirements and authorized data access.
- Use a guided job launch, model endpoint or direct VM access, according to the provider's capabilities.

Our execution layer would check software and driver compatibility, prepare data near the selected GPUs and manage the launch. Training support would include checkpoints and results; inference support would include batch processing, model serving and planned traffic spikes.

Pilot proposal

The first pilot would test a complete transaction and workload on partner infrastructure: agree the capacity and price, prepare the job, run it and reconcile usage and payment. The partner and customer would set the hardware, duration and security requirements, including access for development.

- Agree one initial customer workload and the inputs required to run it.
- Connect an agreed pool of resources and test onboarding, isolation and metering.
- Complete a paid workload, review the economics and attempt a repeat booking.

Contributions and commercial terms

- Infrastructure partner: provide agreed capacity, access boundaries and operational support; confirm that external use and any resale are permitted.
- Our team: develop the booking and execution integration, help source customer demand, prepare workloads and handle the agreed customer workflow.
- Customer: provide the workload, environment, authorized data access, budget and acceptance criteria.

Training bookings would carry an agreed fee. For inference operated by RunMyML, customers would pay for the agreed inference service and we would pay the provider for GPU time. Prices would need to cover setup, storage, transfer, operations and unused time.

Success measures

We would measure capacity sold, supplier revenue, the customer's full cost, onboarding time, execution success and willingness to book again. Benchmarking would support resource selection and verify that time, performance and model quality meet the workload's needs.

The pilot could focus on training or inference. For inference, success would include delivering model outputs within the agreed volume, latency and service window, alongside measuring the full cost of rented GPU capacity.

Founder experience

Grigorii Penkin has practical experience with ML experimentation and GPU workload execution. In a previous role, his team used SkyPilot before adopting an in-house execution approach. He has also evaluated dstack. This experience informs the focus on preparing existing code, environments and data for execution on external GPUs.

Introductions requested

The project is at the concept and prototyping stage. We would welcome introductions to GPU providers or authorized capacity holders with available resources, ML teams seeking temporary compute, and people responsible for GPU infrastructure or ML platforms.