Clastiq AI Platform

The AI workstation that arrives as a platform

NVIDIA GPU hardware and Clastiq Enterprise, delivered, installed and supported as one unit. Open models behind an OpenAI-compatible gateway, GPUs shared fairly down to a fraction of a card, and full visibility of who is using what. Plug it in, sign in, serve your first model the same day. Nothing leaves the box.

SOLUTION BRIEF

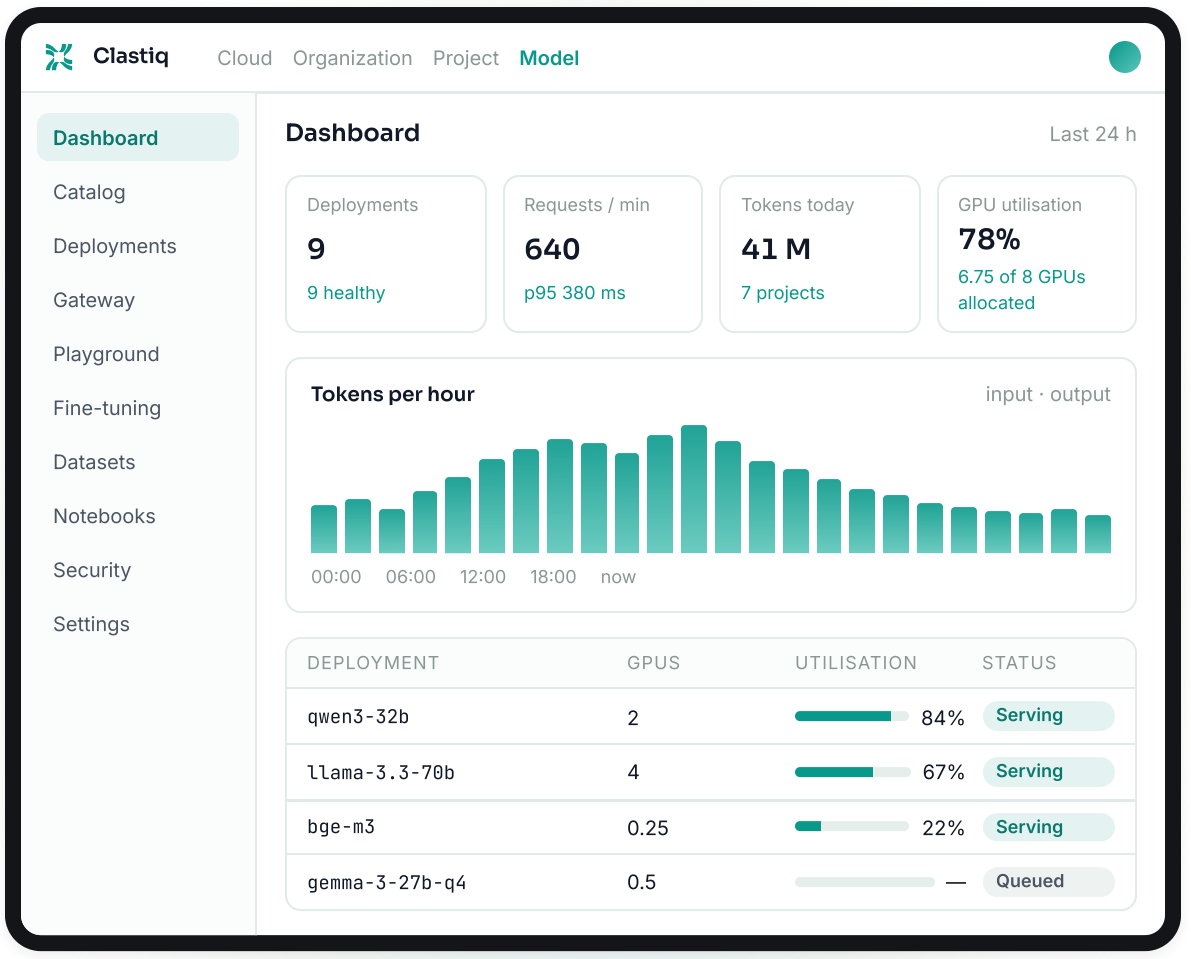
NVIDIA hardware, Clastiq Enterprise and support in one solution

THE SOLUTION

One box, every team

Clastiq AI Platform is an NVIDIA GPU system with Clastiq Enterprise pre-installed. It gives each team a project of its own: members, roles, quotas, storage and keys. Teams deploy models, start databases and tools from the marketplace, run their own containers and notebooks, and fine-tune on the shared GPUs without ever touching the machine.

One vendor, one invoice, one support contract for hardware and software. The administrator decides how the GPUs are split. Every request, every GPU minute and every token is visible on one dashboard.



The model console: deployments, gateway traffic and GPU utilisation for the whole workstation on one screen, from one hour to seven days.

¼ GPU
the smallest slice a notebook or embedding model can take, so eight cards serve dozens of jobs

4
inference engines: vLLM, SGLang, llama.cpp and TensorRT-LLM, chosen per deployment

1
OpenAI-compatible endpoint for every model, with keys and token quotas per project

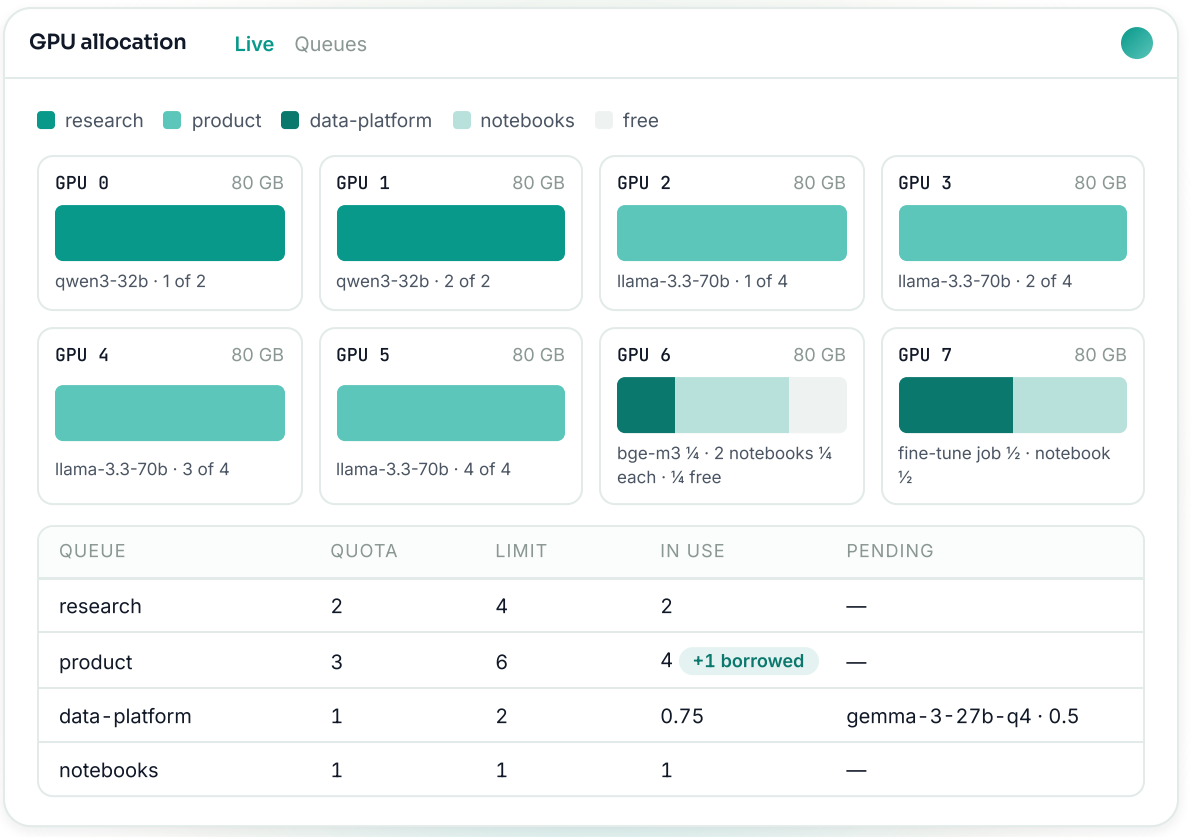
0
bytes of weights, prompts or usage data leaving the box

GPU SHARING

Every GPU fully used, every team fairly served

The GPUs are the most valuable part of the box, and Clastiq treats them that way. The administrator gives each project a queue with a guaranteed GPU quota and a hard limit. When other teams are idle, a queue may borrow beyond its quota according to its weight. When demand returns, borrowed capacity is handed back.

- **Fractional GPUs.** A notebook, an embedding model or a small chat model can request a quarter or half of a card. Several workloads share one GPU safely, each with its own memory budget.
- **Multi-GPU models.** A 70B-class model spans four cards with tensor parallelism and still appears as one deployment.
- **Queued, never refused.** Work that exceeds the quota waits in the project's queue and starts the moment capacity frees up. Users see where their job stands.
- **Autoscaling within limits.** Deployments add replicas on load and give them back when it drops, never beyond the project's ceiling.



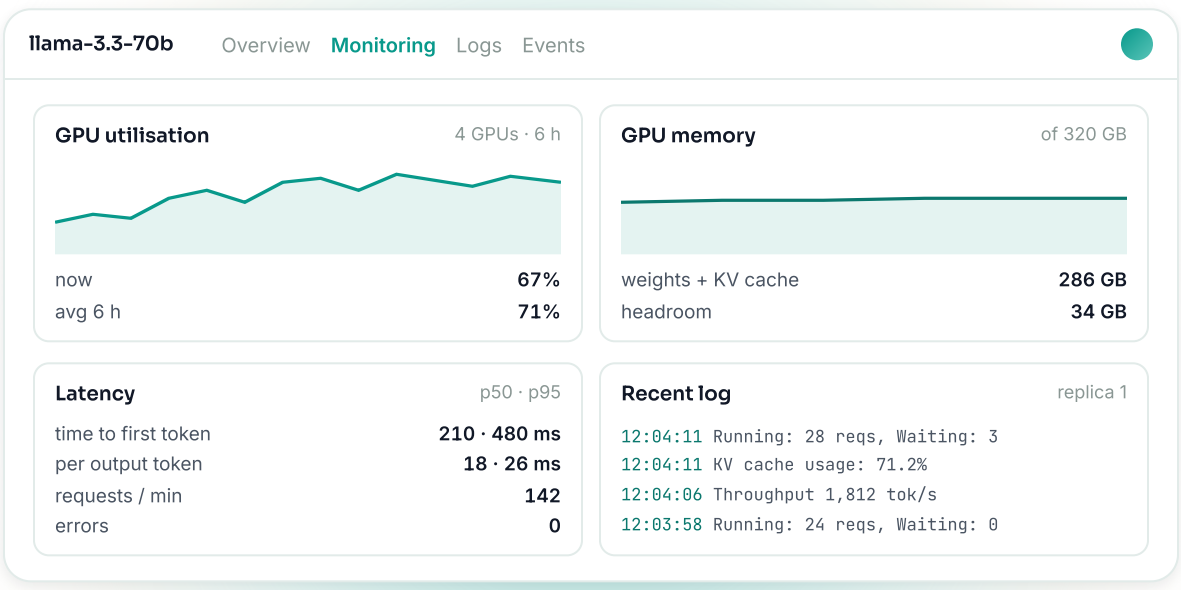
Whole cards for large models, fractions for small ones. Each team's queue has a guaranteed share, a hard ceiling and a weight for borrowing idle capacity.

OBSERVABILITY

See every GPU minute and every token

You paid for the GPUs; you should see them work. The model dashboard shows utilisation and memory per GPU and per deployment, request rates, token volume and latency, from the last hour to the last seven days. Project members see the same for their own workloads.

- **Per deployment:** GPU utilisation, GPU memory, replicas, requests per minute, tokens in and out, time to first token, errors.
- **Per project:** CPU, memory, GPU and storage against quota, with usage history and the workloads behind it.
- **Per workload:** live logs, events and resource graphs, and alerts when a model stops answering.
- **For the administrator:** the whole system at a glance, GPU temperatures and power included, plus ready-made monitoring dashboards for deeper analysis.



One deployment, one screen: GPU, memory, latency and logs side by side.

MODEL CONSOLE

Deploy any open model in minutes

From a Hugging Face checkpoint or a GGUF file to a production endpoint with autoscaling, guardrails and per-project keys. No infrastructure knowledge required from the team that needs the model.

Model catalog

A curated library of chat, completion, embedding and vision-language models, plus the entries your teams add. Weights are downloaded once into the local model store and reused by every deployment.

`Hugging Face` `GGUF` `private models`

Multi-engine deployments

Pick the engine per deployment. Fractional GPUs for small models, tensor parallelism for large ones, fp8 KV cache, speculative drafts and an engine cache so replicas start in parallel.

`vLLM` `SGLang` `llama.cpp` `TensorRT-LLM`

API gateway

One OpenAI-compatible endpoint in front of every model. Project-scoped keys, tokens-per-minute quotas, routing across several deployments and usage tracked per project as input and output tokens.

`/v1/chat/completions` `keys per project` `TPM quota`

Playground

Try any gateway model with text, images and documents before writing a line of code. Compare answers, tune parameters and copy a working request.

`vision` `PDF & text files` `streaming`

Fine-tuning

Train catalog models on your own data on the shared GPUs. Datasets are validated and previewed, jobs show live loss curves, and the result is a private model that deploys like any other.

`SFT` `DPO` `ORPO` `KTO` `LoRA / QLoRA / full`

Guardrails

Admins enforce, projects opt in. PII detection and masking, content filters, code-execution blocking and LLM-as-a-judge checks, applied per model, project or key, before or after the call.

`Presidio` `policies` `runs offline`

Autoscaling

Replicas follow demand on engine signals such as running requests, queue depth and KV-cache pressure, within the project's GPU limit. Vertical right-sizing trims CPU and memory without restarts.

`engine metrics` `GPU quota aware`

Managed notebooks

JupyterLab with a GPU preset, single sign-on and an idle timeout that frees the GPU when nobody is working. Packages survive restarts and every notebook counts against the project's quota.

`JupyterLab` `fractional GPU` `idle stop`

Agentic assistant

A chat panel in every console that deploys, starts, scales and changes workloads with an approval card for each action, and diagnoses a failing workload from its logs, events and metrics with a fix ready to approve.

`acts with approval` `diagnoses failures` `your own model`

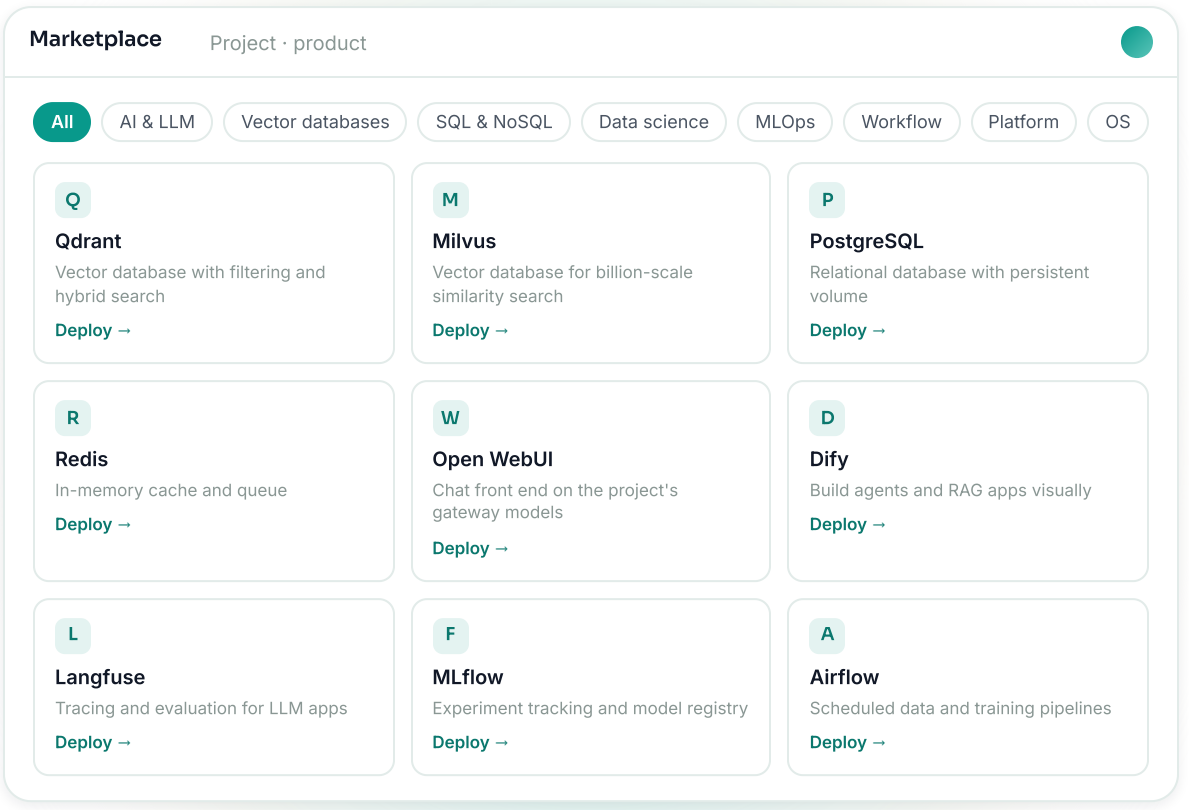
MARKETPLACE

Everything around the model, one click away

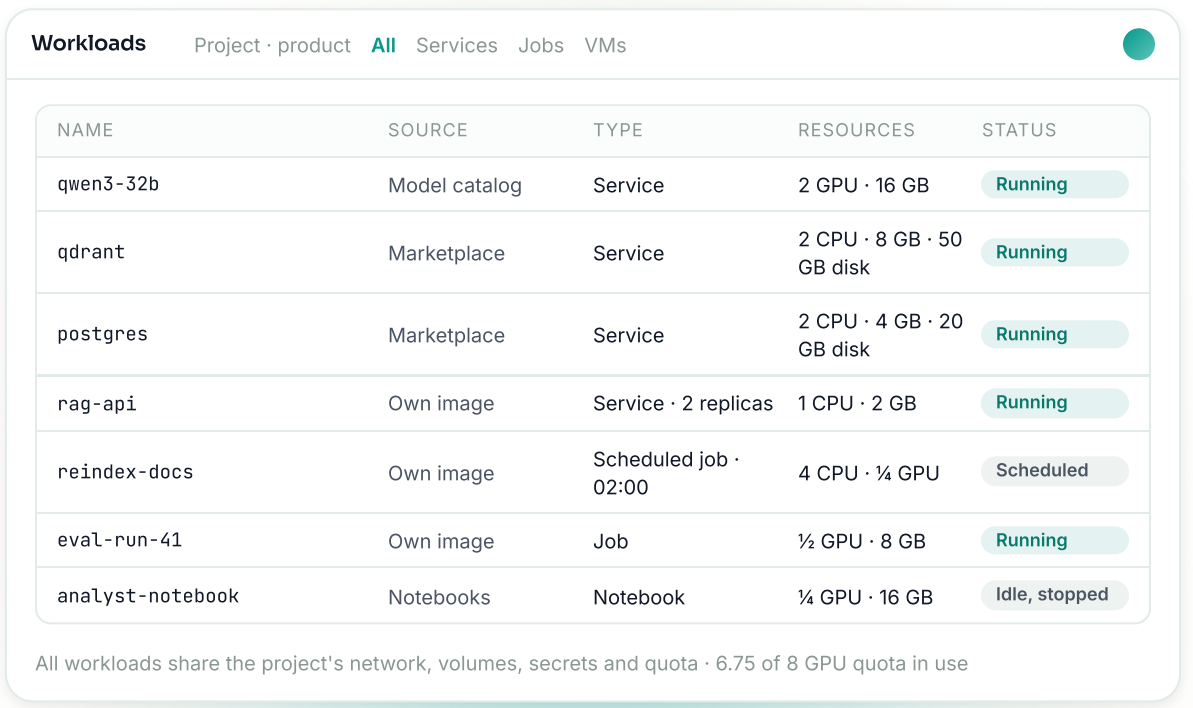
A model alone is not an application. Teams need a vector database for retrieval, a relational database for state, a chat front end, an experiment tracker, a workflow engine. The marketplace starts each of them inside the project with a short form, on the same GPUs, storage and sign-on as the models.

- **Vector databases:** Qdrant, Milvus, Weaviate, Chroma.
- **SQL and NoSQL:** PostgreSQL, MariaDB, MongoDB, Redis, Elasticsearch, InfluxDB.
- **AI applications:** Open WebUI, Dify, Langflow, Langfuse, Ollama, Gradio, Streamlit.
- **Data science and MLOps:** JupyterLab, RStudio, MLflow, Aim, Optuna, Superset, Metabase, DataHub.
- **Workflow and platform:** Airflow, Dagster, NiFi, Argo Workflows, Keycloak, Authentik, container registries.

Every template is pinned to a tested version, validated on delivery, and updated with the platform. Administrators can add their own templates for in-house services.



Dozens of services, each started with a form: name, size, a password, done.



One project, one list: the model, the databases around it, your own code and the nightly job.

CUSTOM WORKLOADS

Bring your own containers

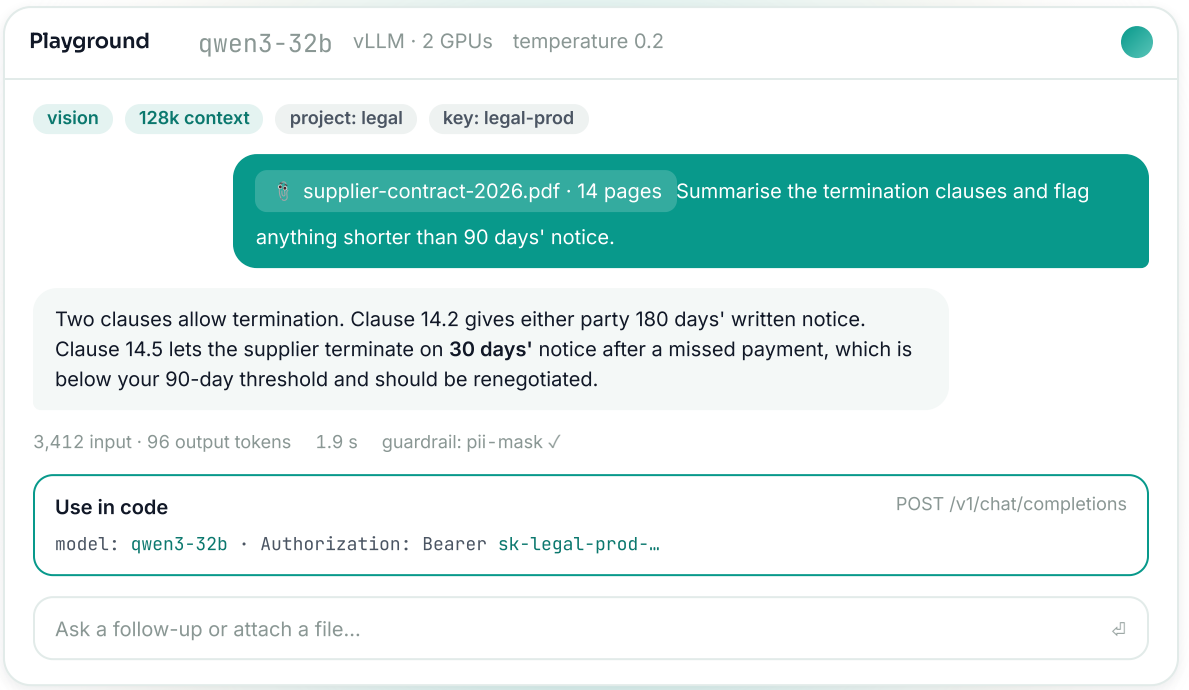
Anything that runs in a container runs on the box, next to the models and with the same GPU fractions, quotas and monitoring. Teams submit their own images without a ticket and without learning the infrastructure underneath.

- **Services** with replicas, ports and a public or internal address, such as a retrieval API in front of a vector database.
- **Jobs and scheduled jobs** for evaluations, data preparation, nightly re-indexing or batch inference, with a GPU slice when they need one.
- **Virtual machines** for software that is not containerised, with SSH access through the platform's sign-on.
- **Three levels of control:** a simple form (image, command, CPU, memory, GPU, ports, env), an advanced form (volumes, probes, autoscaling) or raw YAML for experts.
- **Project resources attached:** persistent volumes, object storage buckets, Hugging Face repos as folders, secrets and environment variables.

From any open model to a production API

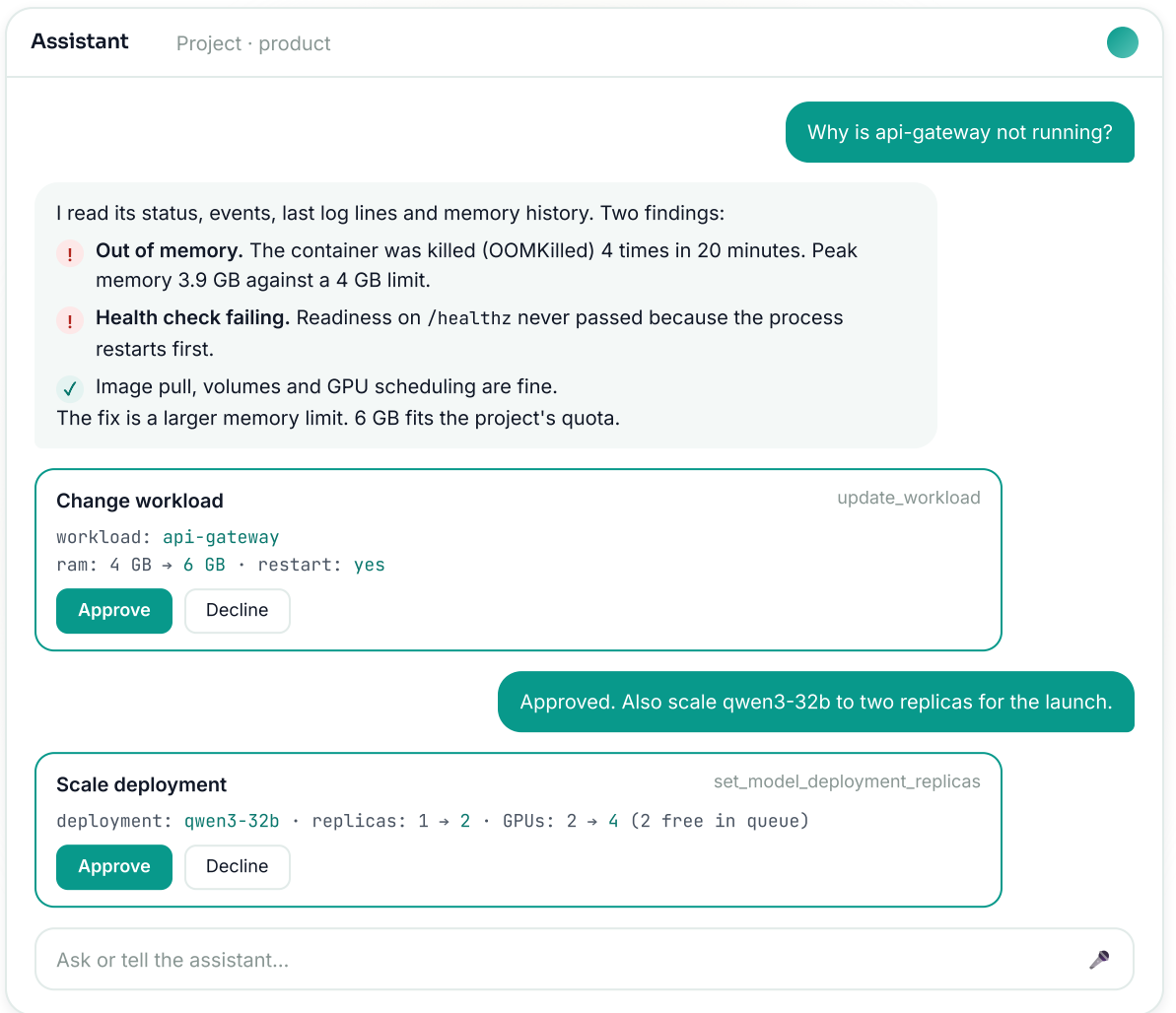
Pick a model from the catalog, pick the engine that runs it best, and it is live behind one OpenAI-compatible endpoint. Teams try it in the playground first, then copy the same request into their code with a project key.

- **The right engine per model.** vLLM, SGLang, llama.cpp or TensorRT-LLM, chosen per deployment with optimized presets: fractional GPUs for small models, tensor parallelism for large ones, fp8 KV cache and speculative drafts for throughput.
- **One gateway for everything.** Every model behind the same OpenAI-compatible API, so existing SDKs, agents and tools work unchanged. Project-scoped keys, tokens-per-minute quotas, routing across replicas and deployments.
- **Playground before code.** Chat with any gateway model using text, images, PDFs and code files. Compare answers, tune parameters, copy a working request.
- **Usage you can see.** Input and output tokens per project and per model, next to GPU time, so you always know which team is using what.



Same model, same key, same endpoint as the application that ships next week.

An operator that diagnoses and acts, with your approval



Findings with evidence, fixes and actions as approval cards: nothing runs until you say so.

Every console carries a chat panel that does more than answer questions. Ask why a workload is not running and it reads the status, restarts, last log lines, events and memory peaks, then returns findings with evidence and a fix. Ask it to deploy a model, scale an API or add a teammate, and it does it through the same permissions you have.

- **Recognises the usual suspects.** Out-of-memory kills, image pull errors, crash loops by exit code, pods no GPU can take, failed mounts, evictions, failing health checks, failed model downloads.
- **Approval before any change.** Each action or fix is a card with the exact call. Nothing runs until you approve, the approval is signed by the server, and every approved change is in the audit log.
- **Never destructive, always in scope.** It can create, start, stop, scale and change, never delete, and sees only what you could see yourself.
- **Your model, on your box.** Runs on a model served by the system itself or on any provider you choose. Voice input in the browser.
- **Open to your agents.** The same tools are exposed as an MCP endpoint, so coding agents and automation use the platform with a personal access token.

BUILT FOR ONE ORGANIZATION

Your hardware. Your identity. Your data.

Everything runs on the box: models, prompts, datasets, fine-tuned weights, logs and metrics. Users sign in with your single sign-on, and projects keep teams' work apart.

Consoles & APIs
How people and automation use the platform

- Admin console
- Organization console
- Project console
- Model console
- REST API · OpenAPI
- MCP endpoint

IDENTITY AND QUOTAS ON EVERY CALL

Clastiq Enterprise
The model lifecycle, end to end

- Catalog & model store
- Deployments
- Gateway
- Fine-tuning
- Datasets
- Notebooks
- Guardrails
- Hugging Face volumes

SCHEDULED ONTO

NVIDIA hardware
Pre-installed, tuned and supported by Clastiq

- NVIDIA HGX or DGX GPUs with NVLink
- GPU scheduler with fractions & queues
- NVMe model & dataset storage
- Image registry
- Single sign-on
- Secrets
- Metrics & dashboards
- Backups

Projects for teams

Each team gets a project with its own members and roles, quotas for CPU, memory, GPU, storage and tokens, and its own API keys, buckets and secrets. Network policies keep projects apart on the shared machine.

Quotas you can trust

A project cannot take more than its limit, and never starves another team of its guaranteed share. Idle GPUs are lent out automatically and reclaimed when their owner needs them.

Operable by one admin

The box arrives installed and tuned for its GPUs. Upgrade in one step, back up and restore workloads from the console, and read every administrative change in the audit log. No platform team required.

CLASTIQ AI PLATFORM CONFIGURATIONS

Hardware and platform, one solution

Four NVIDIA systems, from a desk-side station to an eight-GPU Blackwell Ultra node, each shipped with Clastiq Enterprise installed, tuned for its GPUs and covered by one support contract.



NVIDIA DGX STATION

Clastiq Station

A desk-side Grace Blackwell system on standard office power. One large GPU with unified memory, shared by fractions across a whole team's notebooks, embeddings and chat models.

Request a quote

Datasheet

COMPUTE
1 × NVIDIA GB300 Grace Blackwell Ultra Desktop Superchip

MEMORY
784 GB unified, 288 GB of it HBM3e on the GPU

SERVES
Up to 200B-class models in fp4, one GPU shared by fractions

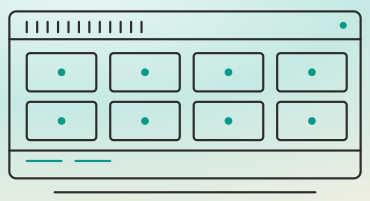
CPU
72-core NVIDIA Grace

STORAGE
NVMe model and dataset store

NETWORK
NVIDIA ConnectX-8 SuperNIC, up to 800 Gb/s

SOFTWARE
Clastiq Enterprise pre-installed, unlimited users and projects

SUPPORT
3 years hardware and platform, next-business-day parts



NVIDIA HGX H200

Clastiq H200

Eight Hopper GPUs with 141 GB each for an organization-wide AI service: several production models, a training pool and dozens of notebooks at once.

Request a quote

Datasheet

GPU
8 × NVIDIA H200 SXM, 141 GB each, NVLink

GPU MEMORY
1.1 TB total

SERVES
70B-class models at 128k context, plus fine-tuning

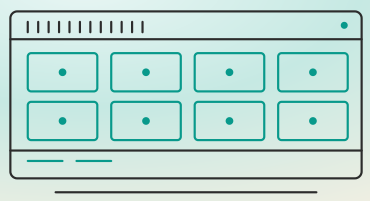
CPU · MEMORY
Dual-socket, 192 cores · 2 TB DDR5 ECC

STORAGE
30 TB NVMe model and dataset store

NETWORK
2 × 200 GbE, 1 × management

SOFTWARE
Clastiq Enterprise pre-installed, unlimited users and projects

SUPPORT
3 years hardware and platform, next-business-day parts



NVIDIA HGX B200

Clastiq B200

Eight Blackwell GPUs for the largest open models and the fastest training: fp8 and fp4 inference with headroom for the next generation of models.

Request a quote

Datasheet

GPU
8 × NVIDIA B200 SXM, 180 GB each, NVLink

GPU MEMORY
1.4 TB total

SERVES
Frontier open models, full fine-tuning of 70B-class

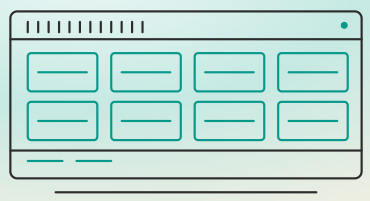
CPU · MEMORY
Dual-socket, 192 cores · 2 TB DDR5 ECC

STORAGE
60 TB NVMe model and dataset store

NETWORK
2 × 400 GbE, 1 × management

SOFTWARE
Clastiq Enterprise pre-installed, unlimited users and projects

SUPPORT
3 years hardware and platform, next-business-day parts



NVIDIA HGX B300

Clastiq B300

Eight Blackwell Ultra GPUs with 288 GB each: the most memory per GPU for frontier open models at long context, and the highest fp4 throughput in the line-up.

Request a quote

Datasheet

GPU
8 × NVIDIA B300 SXM, 288 GB each, NVLink

GPU MEMORY
2.3 TB total

SERVES
Frontier open models at long context, several at once

CPU · MEMORY
Dual-socket, 192 cores · 2 TB DDR5 ECC

STORAGE
60 TB NVMe model and dataset store

NETWORK
2 × 400 GbE, 1 × management

SOFTWARE
Clastiq Enterprise pre-installed, unlimited users and projects

SUPPORT
3 years hardware and platform, next-business-day parts

Delivered ready to serve

The box arrives with Clastiq Enterprise installed, GPU sharing configured and single sign-on ready to connect. The first model is serving on delivery day.

One contract

Hardware, platform updates and support under one agreement with one point of contact. No separate software licence to track.

Grows with you

Add a second system later and Clastiq Enterprise treats both as one pool, with the same projects, queues and gateway.

AT A GLANCE

Feature list

Capability	What you get
GPU SHARING & SCHEDULING	
Fractional GPUs	Notebooks, embedding models and small chat models request a quarter or half of a card. Several workloads share one GPU, each within its own memory budget.
Project queues	Guaranteed GPU quota, hard limit and a weight for borrowing idle capacity, per project. The same for CPU and memory.
Queued work	Jobs beyond quota wait in the queue and start when capacity frees up. Users see their position and the reason.
Multi-GPU models	Tensor parallelism across cards for 70B-class models, presented as one deployment.
Autoscaling	Replicas follow running requests, queue depth or KV-cache pressure, within the project's GPU limit. Vertical right-sizing of CPU and memory.
OBSERVABILITY	
GPU metrics	Utilisation and memory per GPU and per deployment, from one hour to seven days, on the model dashboard.
Gateway metrics	Requests, input and output tokens, latency and errors per model and per project.
Workload insight	Live logs, events and resource graphs for every deployment, notebook and job; usage history per project against quota.
Monitoring dashboards	Ready-made dashboards for the administrator, with every metric available for custom views and alerts.
MODELS & INFERENCE	
Model catalog	Hugging Face and GGUF sources, custom entries, private fine-tuned models.
Model store and downloads	Weights are downloaded once from Hugging Face into the local NVMe store, with progress shown and downloads that pause and resume. Every deployment and fine-tune loads from the store, and an engine cache lets replicas start in parallel.
Inference engines	vLLM, SGLang, llama.cpp and TensorRT-LLM, chosen per deployment, with optimized presets.
API gateway	OpenAI-compatible endpoint, per-project keys, tokens-per-minute quotas, routing across deployments.
Playground	Chat with any gateway model using text, images, PDFs and code files.
Fine-tuning	SFT, DPO, ORPO and KTO; LoRA, QLoRA and full; vision models; datasets validated and previewed; private results.
Managed notebooks	JupyterLab with GPU presets, single sign-on, persistent packages and idle stop.
Hugging Face volumes	Model and dataset repos mounted into workloads as folders, fetched on first read.
AGENTIC ASSISTANT & DIAGNOSTICS	
Acts on your behalf	Deploy models, submit, start, stop, scale and change workloads, create projects, add members, start marketplace services. Each action shown as an approval card with the exact call; nothing runs unapproved and nothing deletes.
Workload diagnostics	"Why is it not running?" answered from status, restarts, container exits, last log lines, events and memory peaks, with findings, evidence and a proposed fix for approval.
Answers and lookups	Quota left, running workloads, GPU usage, how-to questions, all scoped to what the user may see.
Your model or any provider	Runs on a model served by the system itself, or on an external provider; the key never reaches the browser. Voice input in the browser.

Capability	What you get
MCP endpoint	The same tools for coding agents and automation, authenticated with personal access tokens.
SECURITY & GOVERNANCE	
Guardrails	PII masking, content filter, code-execution block and LLM-as-a-judge, scoped globally, per model, per project or per key. Runs fully offline.
Identity	Single sign-on for every console and notebook, roles per project, personal access tokens for scripts.
Isolation	Projects separated by network policies and firewalls; private buckets and a private model store per project.
Audit	Every administrative change recorded with who, what and when.
MARKETPLACE & WORKLOADS	
Vector and relational databases	Qdrant, Milvus, Weaviate, Chroma, PostgreSQL, MariaDB, MongoDB, Redis, Elasticsearch and InfluxDB, started with a form and backed by project volumes.
AI applications and tooling	Open WebUI, Dify, Langflow, Langfuse, Ollama, Gradio, Streamlit, MLflow, Aim, Optuna, Superset, Metabase, DataHub, Airflow, Dagster, NiFi, Argo Workflows and more, pinned to tested versions.
Custom containers	Your own images as services, jobs, scheduled jobs or virtual machines, with GPU fractions, replicas, ports, volumes, buckets, secrets and environment variables.
Object storage buckets	S3-compatible buckets per project on the box: browse and upload in the console, mount into any workload as a folder, and use from code with standard S3 clients and project keys.
Submission modes	Simple form, advanced form with probes and autoscaling, or raw YAML for experts. The same quota and monitoring for all three.
Own templates	Administrators publish in-house services to the marketplace for every team.
PLATFORM & OPERATIONS	
Automation	REST API with OpenAPI description and personal access tokens for scripts and infrastructure-as-code.
Operations	One-step upgrades, backups and snapshots from the console, GPU health and temperatures on the admin dashboard.
HARDWARE & DELIVERY	
NVIDIA systems	Clastiq Station on DGX Station (desk-side Grace Blackwell), and Clastiq H200, B200 and B300 on HGX H200, B200 and B300 (eight SXM GPUs with NVLink), each with an NVMe model store sized for its GPUs.
Pre-installed	Clastiq Enterprise, GPU scheduler, registry, single sign-on and monitoring configured and tested before shipment.
One contract	Hardware warranty, platform updates and support in one agreement, with one point of contact.
Enablement	On-site or remote workshop for administrators and the first teams; first model serving on delivery day.

NEXT STEP

Ready to run AI on your own terms?

A one-hour walkthrough covers the GPU class that fits your models, the teams that will share the box and how it connects to your sign-on and network. Your Clastiq AI Platform arrives installed and serves its first model on delivery day.

Book a walkthrough | **Request the datasheet**

Clastiq

clastiq.ai

© 2026 Clastiq. NVIDIA, HGX, DGX, Grace, Hopper, Blackwell and ConnectX are trademarks of NVIDIA Corporation. Other names are trademarks of their respective owners.

vLLM, SGLang, llama.cpp, TensorRT-LLM and Hugging Face are independent projects. Specifications subject to change.