ENTERPRISE AI SOLUTIONS

AI SOLUTIONS BROCHURE

Harch Intelligence Platform

Africa's most powerful and sustainable AI compute platform. GPU clusters for LLM training, real-time inference, sovereign cloud, and edge AI. 89% lower carbon, 72% lower energy cost.

Harch Intelligence | 1,798 GPUs
H100 + H200 | Sovereign Cloud | Edge AI

2025

HARCH CORP

Harch Intelligence AI Platform

Harch Intelligence provides Africa's most powerful and sustainable AI compute platform, designed to meet the demanding requirements of enterprise AI workloads, government sovereign cloud deployments, and research institution compute partnerships. Our platform delivers GPU compute capacity that is not only technically competitive with the world's leading cloud providers but also uniquely sustainable, with 89% lower carbon intensity than the industry average. Whether you are training large language models, running real-time inference at scale, or building sovereign AI capabilities for your nation, Harch Intelligence provides the infrastructure, tools, and expertise to accelerate your AI journey.

1,798	400	99.99%	$0.018
GPUs Available	Gbps InfiniBand	Uptime SLA	Energy Cost/kWh

Core AI Capabilities

Large Language Model Training

Harch Intelligence's GPU clusters are optimized for training large language models from 1B to 400B parameters. Our H100 and H200 nodes provide the memory bandwidth, interconnect speed, and computational throughput required for distributed training at scale. With 400 Gbps InfiniBand interconnect, gradient synchronization across nodes is efficient enough to maintain near-linear scaling for models up to 70B parameters on a single cluster, and larger models can be trained using pipeline and tensor parallelism across multiple clusters. We provide pre-configured environments for popular training frameworks including Megatron-LM, DeepSpeed, and FSDP, along with expert support from our ML engineering team to help you optimize your training recipes for maximum throughput and minimum cost.

Real-Time AI Inference

For latency-sensitive inference workloads, our Dakhla and Tanger hubs provide sub-8ms and sub-15ms latency to European end users respectively. This makes Harch Intelligence an ideal platform for serving real-time AI applications such as conversational AI, recommendation systems, fraud detection, and autonomous driving simulation. Our inference infrastructure supports NVIDIA Triton Inference Server, vLLM, and TensorRT-LLM for optimized model serving, with auto-scaling capabilities that adjust GPU allocation based on request volume. Carbon-aware inference routing ensures that inference traffic is directed to the cleanest available hub without impacting latency SLAs.

Sovereign AI Cloud

For governments and regulated enterprises that require data sovereignty, Harch Intelligence offers a sovereign AI cloud that keeps all data and processing within Moroccan jurisdiction. Our sovereign cloud provides the same GPU compute capabilities as our standard platform, with additional security controls including dedicated tenancy, encrypted data at rest and in transit, compliance with GDPR and Moroccan data protection regulations, and the ability to audit and verify that no cross-border data transfers occur. The sovereign cloud is particularly suited for defense, intelligence, healthcare, and financial services applications where data residency requirements make US-based cloud providers unsuitable.

Edge AI Deployment

Harch Intelligence extends AI capabilities to the edge through a network of 50+ edge nodes distributed across Morocco and West Africa. These edge nodes provide low-latency inference for IoT, industrial automation, and smart city applications, with models trained on Harch Intelligence's central GPU clusters and deployed to edge nodes via automated CI/CD pipelines. Edge nodes support NVIDIA Jetson and Qualcomm AI accelerators, and are connected to the central platform via Harch Technology's satellite and fiber network.

Service Tiers and Pricing

Harch Intelligence offers flexible engagement models to match the diverse needs of our customers, from startups conducting initial AI experiments to governments building national AI infrastructure. Our pricing reflects the structural cost advantage of our vertically integrated model, delivering GPU compute at significantly lower cost than European competitors while maintaining superior sustainability metrics.

Service Tier	GPU Type	Commitment	Best For	Key Features
On-Demand	H100/H200	Hourly	Experiments, burst workloads	Pay-per-use, auto-scaling
Reserved	H100/H200	1-3 year	Steady-state training	Up to 60% discount, priority access
Sovereign Cloud	Dedicated cluster	3-5 year	Government, regulated	Dedicated tenancy, compliance
Colocation	Customer hardware	5+ year	Custom hardware needs	Space, power, cooling, connectivity
Research Partnership	Shared cluster	Annual	Academic institutions	Discounted rates, joint research

Table 1: Harch Intelligence service tiers and engagement models

Industry Use Cases

Financial Services

Banks and financial institutions are rapidly adopting AI for risk modeling, algorithmic trading, fraud detection, and customer service automation. Harch Intelligence provides the GPU compute capacity and low-latency connectivity required for these workloads, with sub-30ms latency to European financial centers ensuring real-time performance. Our sovereign cloud option addresses the strict data residency requirements of financial regulators, while our carbon-aware scheduling helps institutions meet their ESG commitments.

Healthcare and Life Sciences

AI is transforming drug discovery, medical imaging, genomics, and clinical decision support. These workloads require massive GPU compute for training and strict data sovereignty for patient data. Harch Intelligence's sovereign cloud provides both, with dedicated tenancy ensuring that sensitive patient data never shares infrastructure with other customers. Our platform supports popular bioinformatics frameworks and provides pre-configured environments for common workloads such as protein structure prediction (AlphaFold), genomic sequence analysis, and medical image classification.

Government and Defense

Governments building national AI capabilities require infrastructure that is sovereign, secure, and reliable. Harch Intelligence's sovereign AI cloud provides GPU compute within Moroccan jurisdiction, with Tier IV+ physical security, ISO 27001 and SOC 2 compliance, and the ability to support classified workloads. Our platform enables governments to develop and deploy AI applications for intelligence analysis, cybersecurity, public safety, and administrative automation without relying on foreign cloud providers.

Energy and Utilities

The energy sector is using AI for predictive maintenance, grid optimization, demand forecasting, and renewable energy output prediction. As a vertically integrated energy and compute company, Harch Corp offers unique domain expertise in this space, with Harch Energy providing real-world operational data and Harch Intelligence providing the AI platform to extract insights from that data. Our platform supports time-series forecasting, anomaly detection, reinforcement learning for grid control, and digital twin simulation for asset management.

Performance Benchmarks

Harch Intelligence's GPU clusters deliver performance that is competitive with the world's leading cloud providers, as demonstrated by standard AI benchmarks. The following table presents benchmark results from our H100 clusters running popular AI workloads.

Benchmark	Model Size	Harch H100	AWS P5 (H100)	Google A3 (H100)
MLPerf Training	GPT-3 175B	Comparable	Baseline	Comparable
MLPerf Inference	BERT-Large	Comparable	Baseline	Comparable
BERT Fine-tuning	340M params	22 min	24 min	22 min
Stable Diffusion	512x512 batch	0.8 sec/img	0.8 sec/img	0.8 sec/img
LLM Serving (vLLM)	Llama-70B	1,200 tok/s	1,180 tok/s	1,200 tok/s

Table 2: AI performance benchmarks comparison

Cost Advantage: While raw GPU performance is comparable across H100 providers, Harch Intelligence's vertically integrated energy model delivers 72% lower energy costs, enabling us to offer GPU compute at significantly lower prices than European competitors without sacrificing margins. Contact us for custom pricing.

Getting Started

Getting started with Harch Intelligence is straightforward. Our solutions team will work with you to understand your workload requirements, select the appropriate service tier, and configure your environment. For enterprise and government customers, we offer proof-of-concept deployments that allow you to evaluate the platform with your own workloads before committing to a longer-term engagement. Our ML engineering team provides hands-on support for workload migration, optimization, and production deployment.

- **Step 1:** Contact our solutions team at harchcorp.com for a technical consultation
- **Step 2:** Select your service tier and configure your GPU cluster
- **Step 3:** Deploy your workloads with pre-configured environments or custom setups
- **Step 4:** Monitor performance and carbon metrics via the HarchOS dashboard
- **Step 5:** Scale up or down based on demand with flexible capacity management