WHITE PAPER 2025

GPU COMPUTE & AI INFRASTRUCTURE

Technical Whitepaper

A comprehensive analysis of Harch Intelligence's vertically integrated GPU compute architecture, carbon-aware scheduling system, and competitive positioning against incumbent cloud providers. Demonstrating 72% lower energy costs and 89% lower carbon intensity.

Harch Intelligence | $1.14B Investment

1,798 GPUs | 5 Hubs | PUE <1.15

2025

HARCH CORP

Executive Summary

The global demand for AI compute capacity is growing at an unprecedented rate, driven by the proliferation of large language models, generative AI applications, and enterprise AI adoption. Current projections estimate that global AI compute demand will increase 10x by 2028, yet the data center infrastructure required to meet this demand faces significant constraints in energy supply, carbon emissions regulations, and geographic concentration risk. Harch Intelligence addresses these challenges through a fundamentally different approach: vertically integrated AI infrastructure that co-locates compute with dedicated renewable energy generation, achieving superior efficiency, lower costs, and dramatically reduced carbon emissions compared to any disaggregated infrastructure model. This whitepaper presents the technical architecture, operational model, and competitive advantages of Harch Intelligence's GPU compute platform.

Key Findings: Harch Intelligence delivers AI compute at 72% lower energy cost, 89% lower carbon intensity, and comparable or superior latency to European markets compared to incumbent cloud providers. The vertically integrated model eliminates the energy-cost volatility and carbon-accounting complexity that plagues conventional data center operators, providing enterprise customers with predictable pricing and verifiable sustainability metrics.

1. The AI Compute Challenge

1.1 Exponential Demand Growth

The demand for GPU compute capacity has grown exponentially since the introduction of transformer-based large language models in 2017. Training a single frontier model like GPT-4 requires approximately 25,000 A100-equivalent GPUs running for 100 days, consuming an estimated 50 GWh of electricity. With dozens of organizations now training models at this scale, the aggregate energy demand for AI training alone is measured in terawatt-hours. The International Energy Agency projects that data center electricity consumption could double by 2026, with AI workloads being the primary driver of this growth. This creates an urgent need for new data center capacity that is both energy-efficient and sustainably powered.

1.2 Energy Supply Constraints

In major data center markets such as Northern Virginia, Dublin, Amsterdam, and Singapore, grid capacity constraints are already limiting new data center construction. Virginia's Loudoun County, which hosts the world's largest concentration of data centers, faces transmission capacity limitations that have pushed new facility lead times to 3-5 years. In Ireland, data centers now consume over 20% of national electricity generation, prompting the government to impose moratoriums on new connections in the Dublin region. Singapore has implemented a pause on new data center construction since 2019, only recently lifting it with strict efficiency requirements. These constraints are driving operators to seek

alternative locations with abundant clean energy and favorable regulatory environments.

1.3 Carbon Emissions Pressure

Regulatory and stakeholder pressure to reduce carbon emissions is intensifying across the data center industry. The European Union's Energy Efficiency Directive now requires data centers to report their energy consumption and PUE, with anticipated mandates for renewable energy sourcing and carbon disclosure on the horizon. Meanwhile, major enterprises are including data center sustainability metrics in their procurement criteria, with some RFPs requiring sub-1.20 PUE and 100% renewable energy. The current industry average PUE of 1.56 is increasingly unacceptable, and the reliance on renewable energy credits rather than dedicated renewable installations is facing scrutiny from auditors and ESG rating agencies.

2. Harch Intelligence Architecture

2.1 Vertical Integration Design Philosophy

Harch Intelligence's architecture is designed around the principle that AI compute infrastructure should be vertically integrated with energy generation. In a conventional model, a data center operator purchases electricity from the grid, which may be generated from a mix of renewable and fossil sources. The operator then purchases renewable energy credits (RECs) to claim their electricity is renewable, but this accounting mechanism does not reduce actual carbon emissions. Harch Intelligence eliminates this disconnect by owning and operating dedicated renewable energy installations through Harch Energy, ensuring that every watt consumed by our GPU clusters is generated from solar or wind installations built specifically for our facilities. This is not an aspirational commitment or a power purchase agreement; it is physical, dedicated infrastructure that we control end-to-end.

2.2 Five-Hub Distributed Architecture

The platform is distributed across five hubs, each selected for a combination of renewable energy availability, climate conditions favoring free cooling, and proximity to submarine cable landing stations. This distributed architecture serves three strategic purposes: first, it provides geographic redundancy so that no single-site failure can affect platform availability; second, it enables our carbon-aware scheduling system to route workloads to the hub with the lowest real-time carbon intensity; and third, it positions compute close to both African and European demand centers with minimal latency.

Hub	Location	GPUs	Renewable %	gCO2/kWh	Strategic Role
Hub 1	Ouarzazate	800	97.2%	18	Primary training (solar peak)
Hub 2	Dakhla	400	94.8%	32	EU-facing inference

Hub 3	Benguerir	350	88.5%	55	Hybrid workloads
Hub 4	Tanger	200	82.1%	95	Mediterranean bridge
Hub 5	Casablanca	48	45%	210	HQ and development

Table 1: Five-hub distributed architecture with strategic roles

2.3 Carbon-Aware Scheduling Engine

The carbon-aware scheduling engine is the software innovation that unlocks the full potential of our distributed architecture. The system receives real-time telemetry from Harch Energy's solar and wind installations at each hub location, including current generation output, grid carbon intensity, and weather forecasts for the next 24-48 hours. Based on this data, the scheduler assigns workloads to hubs according to the following priority hierarchy: first, minimize real-time carbon intensity; second, maximize renewable energy utilization; third, optimize for latency requirements; and fourth, balance load across the fleet. For latency-sensitive inference workloads, the scheduler can constrain placement to EU-facing hubs while still optimizing for carbon within that subset. The result is that the average carbon intensity across the platform is approximately 47 gCO2/kWh, with the Ouarzazate hub achieving an extraordinary 18 gCO2/kWh during peak solar generation hours.

3. Technical Deep Dive

3.1 GPU Compute Layer

The GPU compute layer is built on NVIDIA's H100 and H200 accelerator platforms, providing the computational density required for large-scale AI training and inference. Each node features eight GPUs connected via NVLink and NVSwitch, providing 900 GB/s of intra-node communication bandwidth. Nodes are interconnected via a 400 Gbps InfiniBand fabric, enabling efficient distributed training across multiple nodes with minimal communication overhead. The platform supports gradient checkpointing, pipeline parallelism, tensor parallelism, and data parallelism for training models ranging from 1B to 400B parameters. All major AI frameworks are supported, including PyTorch, TensorFlow, JAX, and Megatron-LM, with pre-configured environments available for common workload types.

3.2 Storage Architecture

AI training workloads are characterized by high throughput sequential reads of large datasets, periodic checkpoint writes, and random access to model parameters. Harch Intelligence's storage architecture is designed to match these access patterns. The hot tier uses NVMe SSDs with up to 7 GB/s read throughput, organized in a parallel file system (Lustre or Weka) that aggregates bandwidth across

multiple storage servers. The warm tier uses cost-effective SATA SSDs for dataset staging, while the cold tier uses high-density HDD arrays for archival storage. Checkpoint writes are accelerated by a non-volatile write cache that absorbs burst writes and drains to the hot tier asynchronously, preventing checkpoint I/O from interfering with training throughput.

3.3 Network Fabric

The internal network fabric is based on NVIDIA's Quantum-2 InfiniBand platform, providing 400 Gbps per port with adaptive routing and hardware-accelerated collective operations. The fabric topology is a two-level fat tree with over-subscription ratios optimized for AI training workloads, which typically exhibit bursty all-reduce communication patterns. For external connectivity, each hub has multiple 100 Gbps uplinks to the Internet backbone, with direct peering at major European internet exchange points via Morocco's submarine cable systems. BGP anycast is used to steer traffic to the nearest hub, and software-defined networking enables dynamic bandwidth allocation based on workload requirements.

4. Sustainability Analysis

4.1 PUE Benchmarking

Power Usage Effectiveness (PUE) is the most widely used metric for data center energy efficiency, defined as the ratio of total facility power to IT equipment power. A PUE of 1.0 would indicate that all power goes to computing, with no overhead for cooling, lighting, or power distribution. Harch Intelligence targets a PUE below 1.15, which is among the best in the industry and significantly below the global average of 1.56. This is achieved through several design choices: hybrid liquid-air cooling that reduces cooling energy by up to 40% compared to traditional CRAC systems, approximately 8,500 hours per year of free cooling enabled by Morocco's temperate climate, and efficient power distribution using high-voltage DC distribution that minimizes conversion losses.

Provider	Published PUE	Methodology	Notes
Google (Global)	1.09	Trailing 12-month	Best-in-class, includes water cooling
AWS (Global)	1.15	Weighted average	Includes custom cooling designs
Harch Intelligence	<1.15 (target)	Design PUE	Hybrid cooling + 8,500 hrs free cooling
Microsoft Azure	1.17	FY25 average	Includes underwater data center pilot
OVHcloud	1.29	European sites	Custom server design, no CRAC
Scaleway	1.37	Paris/Amsterdam	Standard air cooling
Equinix	1.39	Global IBX fleet	Multi-tenant colocation model
Industry Average	1.56	Uptime Institute 2024	Includes all facility types

Table 2: PUE benchmarking across major providers

4.2 Carbon Intensity Comparison

While PUE measures energy efficiency, carbon intensity measures the actual environmental impact of compute operations. A data center with excellent PUE but running on coal-generated electricity has a far higher carbon footprint than a less efficient facility powered by renewables. Harch Intelligence's vertically integrated model ensures that carbon intensity is not merely a reporting metric but a design parameter. By co-locating compute with dedicated renewable generation and implementing carbon-aware scheduling, we achieve an average carbon intensity of approximately 47 gCO2/kWh across the platform, which is 89% below the global data center industry average of approximately 440 gCO2/kWh. Our Ouarzazate hub achieves 18 gCO2/kWh, comparable to Google's best-performing sites and achieved through genuine dedicated renewable generation rather than energy credit purchasing.

Metric	Harch Intelligence	Industry Average	Advantage
Avg Carbon Intensity	~47 gCO2/kWh	~440 gCO2/kWh	89% lower
Best Hub (Ouarzazate)	18 gCO2/kWh	-	Among lowest globally
Renewable Energy	81.5% dedicated	~40% grid mix	2x higher, dedicated
Carbon Credits Used	None	Common (RECs, GOs)	Zero offset dependency
Scope 2 Methodology	Location-based (actual)	Market-based (credits)	Transparent reporting

Table 3: Carbon intensity comparison with industry benchmarks

5. Financial Model and ROI

Harch Intelligence's financial model benefits from the structural cost advantages of the vertically integrated approach. Energy costs, which represent 30-40% of total operating expenditure for conventional data centers, are reduced by 72% through dedicated renewable generation at $0.018/kWh. Cooling costs are reduced by approximately 40% through the combination of free cooling and hybrid liquid-air systems. These operational savings translate directly into higher margins and more competitive pricing for customers, while the dedicated renewable installations eliminate exposure to energy price volatility.

Financial Metric	Value	Context
5-Year Revenue	$1.9B	137% CAGR

Base Case IRR	24.7%	~30 month payback
NPV (12% discount)	$2,850M	4.6x return
Energy Cost Advantage	72% below EU	$0.018 vs $0.065/kWh
Operating Margin	~55%	vs ~35% industry average
GPU Utilization Target	>80%	vs ~50% industry average

Table 4: Key financial metrics and projections

6. Conclusion and Next Steps

Harch Intelligence represents a new paradigm in AI infrastructure: vertically integrated compute that is simultaneously more efficient, more sustainable, and more cost-effective than conventional approaches. By co-locating GPU compute with dedicated renewable energy generation and implementing carbon-aware scheduling, we deliver AI capacity at 72% lower energy cost and 89% lower carbon intensity than the industry average, while maintaining competitive latency to European markets. The five-hub distributed architecture provides geographic redundancy, dynamic workload optimization, and scalability to 500MW by Q4 2028. For enterprise customers, government agencies, and research institutions seeking sovereign AI compute with verifiable sustainability metrics, Harch Intelligence offers a compelling and differentiated platform.

> **Next Steps:** Contact our solutions team for a technical briefing, custom pricing proposal, or to schedule a proof-of-concept deployment. We offer flexible engagement models including reserved capacity, on-demand GPU, and sovereign cloud partnerships.