Institutional Research Whitepaper

AI Infrastructure for the Future of Quantitative Market Research

RICCHE LTD

ricche.ai

Powered by the NVIDIA Accelerated Computing Stack

RICCHE

GPU-Accelerated AI Research Infrastructure for Financial Markets

Executive Summary

The quantitative finance industry is undergoing a fundamental transformation. The convergence of massive data availability, GPU-accelerated computing, and advances in machine learning has created an unprecedented opportunity to rethink how financial markets are studied, modelled, and understood.

Ricche Ltd is building the infrastructure for this new era. Our GPU-accelerated research platform will combine institutional-grade data engineering, NVIDIA CUDA-powered machine learning environments, GPU-parallelised simulation engines, and rigorous validation frameworks into a single, cohesive system designed for one purpose: enabling the next generation of quantitative market research.

> ***We believe the firms that will define the next decade of quantitative finance are those building on AI-native, GPU-accelerated infrastructure from the ground up.***

47x	<2ms	141 GB	400G
GPU vs CPU Training	Inference Latency	HBM3e per H200	InfiniBand NDR

The Opportunity

Financial markets generate more data per day than most industries produce in a year -- tick-by-tick price movements, shifting order book dynamics, macroeconomic releases, satellite imagery, and social sentiment. Yet the infrastructure to process this data at scale remains locked behind the walls of the world's largest quantitative firms. Ricche exists to close this gap: purpose-built, GPU-accelerated research infrastructure that compresses weeks of experimentation into hours and enables research at a scale that was previously impossible.

Platform Architecture

The Ricche platform is designed as a multi-layered research infrastructure where each layer is purpose-built for its specific workload and independently scalable. Data flows through five core layers, from raw market feeds to validated research outputs.

Data Ingestion Layer

Real-time and historical market data feeds from 30+ global exchanges including NYSE, CME, LSE, Eurex, HKEX, and SGX. The pipeline supports tick-level order book data, OHLCV pricing, corporate actions, economic indicators, and alternative data sources. Sub-second ingestion latency with exactly-once delivery guarantees, automatic gap detection, and continuous data quality scoring.

GPU Compute Layer

NVIDIA GPU clusters purpose-built for financial machine learning workloads. Distributed columnar storage (Apache Parquet, Arrow) for petabyte-scale datasets with intelligent tiering across NVMe, SSD, and object storage. Auto-scaling compute allocation with workload-aware priority scheduling ensures optimal GPU utilisation.

Research & Experimentation Layer

CUDA-accelerated PyTorch environments with comprehensive experiment tracking, model versioning, and automated hyperparameter optimisation. Every experiment is reproducible by design -- all code, data, parameters, and results are versioned and auditable. Support for deep learning, gradient boosting, transformers, and ensemble approaches.

Simulation & Validation

GPU-parallelised Monte Carlo and agent-based simulation engines that stress-test candidate models across thousands of market scenarios before progression. Formal validation stages with statistical significance testing, walk-forward analysis, and peer review ensure only genuinely robust models advance to controlled evaluation.

NVIDIA Accelerated Computing Stack

We chose to build Ricche on the full NVIDIA GPU ecosystem because there is no substitute for raw GPU performance when processing financial data at scale. The performance gains are not incremental -- they are transformational.

NVIDIA CUDA **DGX SuperPOD** **NVIDIA GB200 NVL72** **NVIDIA H200** **BlueField-3 DPU** **TensorRT-LLM**

Triton Inference Server **Megatron-LM** **NeMo Framework** **NVIDIA NIM** **TensorRT** **RAPIDS cuDF** **cuML**

CUDA & RAPIDS -- GPU Foundation

Every training run, simulation, and feature computation will be CUDA-accelerated. Multi-GPU distributed training with mixed-precision (FP16/BF16) will deliver dramatic speedups. RAPIDS cuDF will eliminate CPU bottlenecks -- feature engineering on billion-row time-series will run 10-50x faster than pandas.

TensorRT & NIM -- Production Stack

Research models will be optimised via layer fusion, kernel auto-tuning, and INT8 quantisation to target sub-2ms inference latency. NIM microservices will provide one-click deployment with dynamic scaling, intelligent batching, GPU memory management, and zero-downtime version transitions.

141 GB	<2ms	400G	5
HBM3e per H200	Inference Latency	InfiniBand NDR	Asset Classes

Data Infrastructure

In quantitative research, data quality is the foundation on which every model and research conclusion rests. Our multi-stage ingestion pipeline covers equities, futures, options, FX, and fixed income across 30+ exchanges, with alternative data from NLP-processed news, satellite imagery, and social sentiment. Every data point passes through automated quality checks for completeness, consistency, and statistical anomalies. Full data lineage tracking ensures any feature can be traced to its raw source.

Research Methodology

Ricche enforces a disciplined, hypothesis-driven research methodology designed to eliminate the most dangerous traps in quantitative finance: overfitting, look-ahead bias, and narrative-driven analysis.

Every project begins with a registered hypothesis -- target instruments, timeframes, expected signal characteristics, and falsification criteria -- before any code is written. Datasets are extracted with strict temporal discipline (no look-ahead bias), with GPU-accelerated feature engineering producing training-ready datasets 50x faster than traditional tools.

Models are developed in CUDA-accelerated PyTorch environments with full experiment tracking. Candidate models face rigorous evaluation in GPU-parallelised simulation engines -- Monte Carlo simulations across thousands of scenarios including regime changes, liquidity crises, and tail events. Only models surviving out-of-sample validation, walk-forward testing, and formal peer review advance to controlled paper-trading evaluation.

Operations & Monitoring

The Ricche Control Room will provide unified, real-time visibility into every layer of the stack -- from GPU memory allocation to data pipeline health to experiment progress.

>90%	Real-time	Full Stack	99.9%+
GPU Target Util.	Pipeline Monitoring	Observability	Uptime Target

GPU Cluster Command

Real-time utilisation heatmap across all GPU nodes. Per-GPU memory, temperature, power, and CUDA kernel metrics. Training job queue with priority scheduling and proactive failure prediction.

Data & Experiment Ops

Live tracking of all ML experiments with real-time loss curves. Feed-by-feed ingestion status with per-exchange latency. Kubernetes orchestration with GPU scheduling, predictive auto-scaling, and capacity forecasting.

Current Technology Stack

Every tool in our stack was chosen for a specific reason -- performance, reliability, or researcher productivity.

GPU & Accelerated Compute

NVIDIA CUDA **DGX SuperPOD** **NVIDIA GB200 NVL72** **NVIDIA H200** **BlueField-3 DPU** **TensorRT-LLM**

Triton Inference Server **Megatron-LM** **NeMo Framework** **NVIDIA NIM** **TensorRT** **RAPIDS cuDF** **cuML**

ML & Research

PyTorch JAX Triton FlashAttention-3 vLLM Ray W&B

Data Engineering

KDB+/q Arctic Redpanda Apache Iceberg Apache DataFusion Polars DuckDB ClickHouse StarRocks LanceDB Turbopuffer

Cap'n Proto

Networking & Low-Latency

InfiniBand NDR 400G NVMe-oF SPDK DPDK RDMA io_uring PTP (IEEE 1588)

Trading & Execution

Aeron FIX Protocol LMAX Disruptor ITCH / OUCH Protocol Smart Order Routing Backtesting Engine

Platform & Security

Kubernetes Docker Argo CD Cilium Istio HashiCorp Vault Terraform Temporal AES-256 TLS 1.3 RBAC eBPF OpenTelemetry

Prometheus Grafana Audit Trail

Core Languages

Python Rust SQL CUDA C++

Terraform will ensure reproducible deployments. Temporal will orchestrate complex ML workflows. Ray will distribute training across GPU clusters. Redpanda replaces Kafka with 10x lower latency. ClickHouse will power sub-second OLAP queries. Rust will power latency-critical paths.

Long-Term Vision

Ricche is not a product -- it is a platform for discovery. We see a future where GPU-accelerated research infrastructure is as essential as market data itself, where ML approaches surpass traditional quantitative methods in every asset class, and where the best research comes from the best infrastructure, not the largest headcount.

Partnership Opportunities

We welcome collaboration with organisations at the intersection of market data, GPU-accelerated computing, and machine learning.

Technology Partners

GPU computing providers, cloud infrastructure companies, and hardware vendors looking to showcase accelerated computing in quantitative finance.

Data Partners

Market data vendors, alternative data providers, and exchange operators seeking a showcase platform for their data products at GPU speed.

Research Institutions

Universities and research labs exploring AI in financial markets. We provide computational infrastructure at a scale typically only available to top-tier quant firms.

Investors & Advisors

Individuals and organisations interested in the convergence of AI, quantitative finance, and high-performance computing. We are building for the decade ahead.

Let's Build the Future Together

Whether you are a potential partner, investor, researcher, or technology collaborator, we would love to show you what GPU-accelerated market research looks like in practice.

info@ricche.ai | ricche.ai