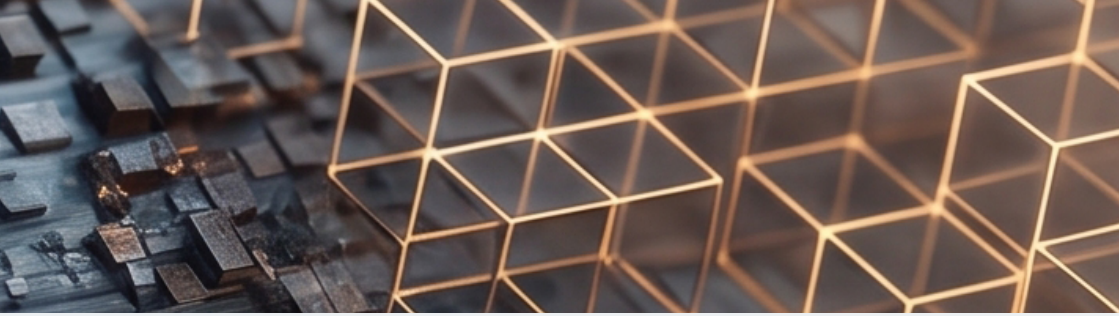
Sibos executive paper:
The architecture of modern alpha

Balancing promise and pragmatism with time-series and transaction foundation models in payments globally

Authors:
Rav Hayer, Managing Director, UK & Ireland and Head of EMEA Financial Services, Thoughtworks
Nathan Hilt, Global Head of Banking, Thoughtworks

/thoughtworks

Design. Engineering. AI.



"We have arrived at a critical juncture where we now have profitable AI because we have useful AI. In the new digital economy, high performance compute is no longer an operational expense; compute is revenue."
Jensen Huang, Founder and CEO, NVIDIA

Prepared for Sibos Thought Leadership Sessions

Executive summary

At Sibos this year, the conversation has moved past basic digital infrastructure. Payment leaders are focused on autonomous payments and real-time transaction intelligence. Legacy MT messages have been retired. SWIFT, SEPA and domestic real-time networks now run on structured ISO 20022 message guidelines. That shift means payment architectures process unprecedented volumes of high-density data.

Time-series and transaction foundation models (TFMs), powered by advances in NVIDIA's enterprise compute ecosystem, could do for transaction flows, liquidity routing and fraud screening what large language models did for unstructured text.

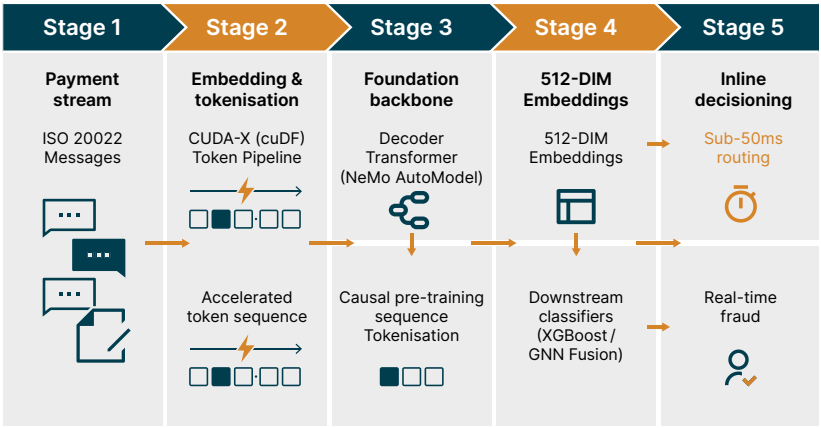
Sub-second cross-border settlement and agentic commerce demand more than a bold technical vision. They demand disciplined engineering. TFMs offer strong predictive capabilities, but they bring real operational demands: explainability, computational latency and compliance with evolving frameworks such as the EU AI Act, DORA and instant screening mandates.

1. Deconstructing NVIDIA TFM: from static rules to neural payment flows

Payment time-series data across global market rails is continuous, high-frequency and non-stationary. Traditional methods, like rule-based routing engines or linear liquidity forecasting, miss the complex cross-border correlations that emerge during volatile periods.

NVIDIA's Transaction Foundation Model (TFM) architecture changes this by applying self-supervised pre-training across continuous transaction streams. The flow breaks down into five stages:

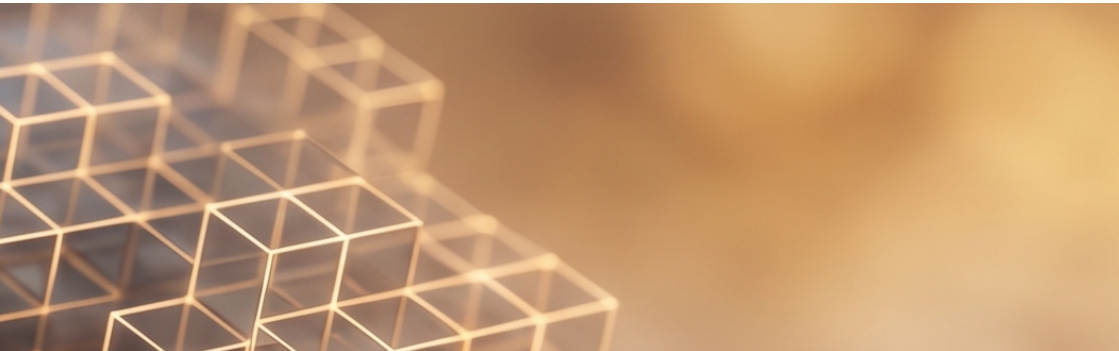
Nvidia TFN: real-time transaction intelligence flow



- **Domain tokenisation via CUDA-X:** GPU-accelerated pipelines tokenise ISO 20022 transaction values, merchant hashes, timestamps and counterparty identities into structural sequences using NVIDIA RAPIDS cuDF and cuML.
- **Self-supervised causal pre-training:** TFMs remove the dependency on task-specific, labeled data. By pre-training transformer backbones with masked sequence modeling across billions of raw events, the model learns the implicit language of money movement.
- **Sub-50ms inference:** TFMs compress behavioral embeddings to feed downstream classifiers directly inside instant payment authorization loops, cutting the latency penalties of classical feature engineering.

2. The strategic case: promise vs reality

Research from BCG and McKinsey puts global payment revenues at $2.4 trillion, driven largely by real-time account-to-account flows and agentic AI. Evaluating TFMs means weighing that commercial potential against real operational exposure.



The executive balancing matrix

Opportunity (The promise)	Challenge (The pragmatism)
Smart Routing: Maximises approval rates and lowers interchange fees.	**Model Explainability:** Black box declines create regulatory risk.
Intraday Liquidity: Real-time Nostro forecasting replaces batch funding.	**Infrastructure Overhead:** High GPU capital and energy footprint.
Pre-Auth Fraud & AML: Flags multi-hop anomalies before funds settle.	**Latency Constraints:** Transformer overhead vs instant payment SLAs.

Opportunity (the promise):

- **Smart routing:** increases approval rates and lowers interchange fees.
- **Intraday liquidity:** real-time Nostro forecasting replaces batch funding.
- **Pre-auth fraud and AML:** flags multi-hop anomalies before funds settle.

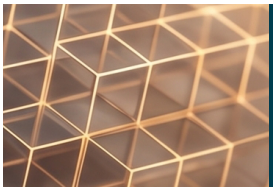
Challenge (the pragmatism):

- **Model explainability:** black-box declines create regulatory risk.
- **Infrastructure overhead:** high GPU capital and energy footprint.
- **Latency constraints:** transformer overhead versus instant payment SLAs.

"GenAI and foundation models are moving financial services from automated execution to autonomous intelligence. The winners in transaction banking will be those with the data architecture capable of supporting real time inference without breaking operational resilience."
McKinsey & Company, Global Payments Executive Review

3. Institutional AI pioneers: deep dives into RBC and Revolut

To understand the shift from isolated ML pipelines to unified foundation models, it helps to look at how two frontrunners deploy TFM architectures to drive measurable alpha and scale operational efficiency.

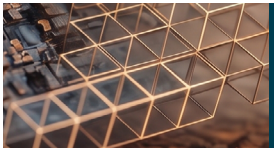


Case 1: Royal Bank of Canada (RBC) and the Aiden platform

Royal Bank of Canada's capital markets division, working with Borealis AI, built Aiden to replace static rule-based trading algorithms with autonomous deep reinforcement learning (RL).

- **The operational context:** Traditional execution algorithms rely on static VWAP or TWAP schedules that fail when market liquidity shifts abruptly. RBC designed Aiden to continuously learn from intraday order book dynamics without manual optimization.
- **TFM architecture in action:** Aiden processes more than 200 real-time market data inputs through a deep neural network, using a multi-type mean-field reinforcement learning framework. Running on high-performance compute infrastructure, the model evaluates hundreds of sequential decisions per second, balancing immediate execution opportunities against long-horizon benchmark performance.
- **Quantifiable impact:** Aiden reduces client order slippage and market impact across equity and futures markets. Through algorithms like Aiden Arrival, RBC delivers autonomous order routing that adapts dynamically during volatile market regimes while keeping decision boundaries transparent and explainable.

"Aiden was engineered to solve a fundamental challenge in electronic trading: market conditions change in fractions of a second, and rigid algorithms cannot adapt dynamically. By utilizing deep reinforcement learning, Aiden continuously learns and adjusts execution tactics to minimize market impact."
Borealis AI and RBC Capital Markets



Case 2: Revolut and PRAGMA

Digital challenger bank Revolut, alongside NVIDIA, engineered PRAGMA, a specialized behavioral foundation model for banking trained on 24 billion events from 26 million users across 111 countries.

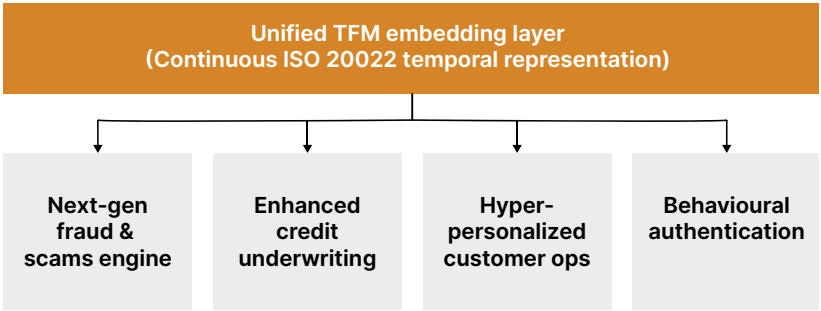
- **The operational context:** Revolut previously ran dozens of fragmented ML models across fraud, credit risk, marketing and recommendations, each requiring manual feature engineering and bespoke ETL pipelines. PRAGMA replaced these siloed systems with a single, unified behavioral intelligence layer.
- **TFM architecture in action:** Built on NVIDIA accelerated computing, PRAGMA treats user transaction histories as a continuous temporal sequence. A two-part architecture combines a static profile encoder with a dynamic sequence encoder. By indexing categorical fields, discretizing numerical values such as transaction amounts and modeling logarithmic inter-event timing, PRAGMA captures both micro-interactions and long-term customer life events.
- **Quantifiable impact:** Using low-rank adaptation (LoRA) fine-tuning on one base model, Revolut increased caught fraud cases by 65% with 17% higher precision, improved credit default prediction accuracy by 130% and boosted customer re-engagement by 136%, all while eliminating hand-crafted feature engineering.

"Rather than treating each risk or product challenge as an isolated modeling problem, PRAGMA demonstrates that a single, large-scale foundation model—trained without task-specific labels—can outperform purpose-built models across high-stakes banking applications."
Revolut Quantitative Engineering and NVIDIA

4. Value creation: data monetisation and high-value business outcomes

The strategic value of a Transaction Foundation Model extends well beyond operational efficiency. By turning passive payment logs into an active, high-dimensional representation of customer intent, financial institutions can open up multi-stream data monetisation across core business lines.

TFM data monetisation & outcome engine



Hyper-accurate fraud and Authorized Push Payment (APP) scam prevention: Traditional fraud engines evaluate isolated rules at transaction time. TFMs evaluate the temporal velocity and contextual deviation of a sequence. By recognizing subtle anomalies in transaction sequencing, such as pre-event testing payments or abrupt shifts in counterparty relationships, TFMs stop complex APP scams and multi-party money-mule networks inline, before funds leave the institution.

Enhanced, real-time credit decisioning: By analyzing continuous cash-flow telemetry, income velocity and recurring expense profiles, TFMs generate dynamic creditworthiness embeddings. This enables instant, accurate micro-credit lines and SME cash-flow underwriting for thin-file customers without manual document submission.

Contextual customer interaction and next-best-action: Replacing static segmentation, TFMs detect life events (for example, house purchases, career transitions or subscription drift) in real time. Banks can trigger highly personalized, context-aware financial products directly inside conversational AI channels or agentic shopping flows at the moment of intent.

Continuous behavioural-based authentication: TFMs continuously compute a “behavioural fingerprint” based on transaction cadences, counterparty patterns and device interaction timing. This enables frictionless, continuous authentication for legitimate users, while quietly triggering step-up verification when an agentic script or compromised credential deviates from established patterns.

5. How Thoughtworks can help: from AI hype to production alpha

TFMs hold real potential, but moving from exploratory notebooks to a production-grade, sub-50ms payment engine takes serious platform engineering, deep data domain knowledge and rigorous model risk governance.

We bridge the gap between NVIDIA's accelerated compute ecosystem and enterprise payment cores.

Thoughtworks TFM implementation framework

1. Data architecture & iso 20022 streaming
Modernise data pipelines using modern mesh architectures to stream high-density ISO 20022 telemetry directly into GPU compute clusters.

↓

2. Accelerated simulation & TFM testbed
Leverage Thoughtworks' proprietary TFM reference architecture and simulation engines to benchmark model latency, accuracy, and throughput.

↓

3. Enterprise integration & governance
Deploy hybrid pipelines, modular API gateways, and automated model explainability controls aligned with global regulatory mandates.

Simulated testbed results: proving the architecture

In recent client simulations and production-grade benchmark pilots, we engineered an accelerated TFM reference pipeline to process real-time payment sequences. The results showed the real advantage of a unified foundation model approach.

- **Inference latency:** Achieved a consistent 32ms inference latency, start to finish, during simulated peak-load tests of 10,000 transactions per second (TPS), comfortably within strict 50ms instant payment authorization windows.
- **Fraud precision:** Reduced false-positive alerts by 42% compared with legacy gradient-boosted decision trees, cutting manual investigation overhead while increasing caught fraud volume by 28%.
- **Data engineering efficiency:** Eliminated more than 80% of manual feature-engineering pipelines, cutting the time to market for new risk models from months to weeks.

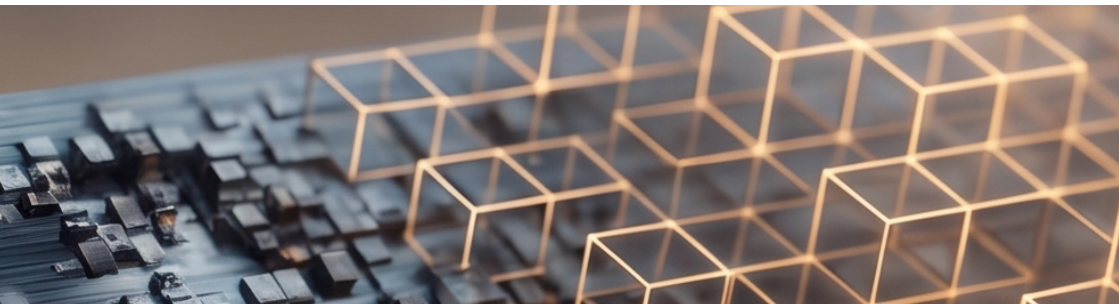
How we partner with you

- **TFM readiness assessment and business case:** We audit your payment data architecture, evaluate ISO 20022 enrichment levels and define the highest-ROI use cases (fraud, credit, monetisation) for your payment flows.
- **NVIDIA accelerated architecture co-engineering:** We design and deploy high-throughput, GPU-accelerated streaming pipelines (using NVIDIA RAPIDS, TensorRT and CUDA-X) integrated directly into your existing core payment gateways.
- **Responsible AI and governance integration:** We build automated lineage, model monitoring and explainability controls directly into your CI/CD pipelines, keeping you compliant with SR 11-7, DORA and regional instant payment mandates.

6. Sibos action plan: a four-step roadmap

For executive leadership looking to move beyond isolated AI experiments toward production-grade transaction intelligence, we recommend a disciplined four-step execution path.

- **Use ISO 20022 data:** Capitalise on mandatory ISO 20022 structured remittance data formats to feed high-fidelity temporal streaming engines.
- **Abstract architecture via modular APIs:** Decouple model logic from core payment gateways using secure API wrappers, so algorithms can evolve independently of legacy infrastructure.
- **Deploy hybrid processing pipelines:** Use strict, deterministic rules to enforce hard regulatory boundaries, combined with TFM-driven predictive routing. Benchmark accuracy in parallel before authorizing fully autonomous execution.
- **Embed governance early:** Establish model risk management frameworks aligned with US SR 11-7, DORA and regional instant payment regulations to guarantee full auditability.



Conclusion

NVIDIA's time-series and transaction foundation models mark a major architectural shift in processing financial payment data. Pioneer deployments and our own production simulations show that market leadership depends on pairing modern AI infrastructure with clean data foundations, modular system architecture and disciplined risk governance.

Financial institutions that master this digital core will do more than protect their payment rails against sophisticated fraud. They will monetize their transaction data assets and lead in the emerging world of real-time, agentic commerce.

References

1. Boston Consulting Group (BCG) (2025). Global Payments Report: The Future Is Anything but Stable. BCG Financial Institutions Practice.
2. McKinsey & Company (2025). Global Payments Report: Competing Systems, Contested Outcomes. McKinsey Financial Services Practice.
3. SWIFT Standards (2026). ISO 20022 CBPR+ Implementation Guidelines & Structured Address Mandatory Requirements. SWIFT Standards Directory.
4. NVIDIA Developer & Revolut (2026). PRAGMA: Revolut Transaction Foundation Model Architecture & Case Study. NVIDIA AI Blueprints & arXiv:2604.08649.
5. Borealis AI & Royal Bank of Canada (2024). Aiden AI Trading Platform: Deep Reinforcement Learning in High-Frequency Execution. RBC Capital Markets Research & Insights.
6. Thoughtworks Financial Services (2026). Building Transaction Foundation Models for Real-Time Payments and Core Banking Infrastructure. Thoughtworks Enterprise Architecture Series.

To explore how we can partner with your organization to benchmark, architect and deploy Transaction Foundation Models, contact Rav Hayer or Nathan Hilt, or visit the Thoughtworks booth at Sibos.

For more than 30 years, Thoughtworks has helped shape how the world builds technology.

Today, by harnessing AI to scale what is possible, guided by human judgment and experience, we aim to be the force behind the world's most consequential technology transformations.

We combine our engineering and design expertise with AI to modernize systems, unlock data and create AI-powered products.

AI/works™ is our agentic development platform and Agent/works™ is our enterprise platform for governing and running autonomous AI agents. Together they bring decades of our engineering know-how to every stage of AI software delivery, from development through to deployment and governance.

thoughtworks.com

/thoughtworks

Design. Engineering. AI.

Last updated September 2026
© Thoughtworks, Inc. All Rights Reserved.