Towards High-Throughput Financial Intelligence: A Hardware-Aware Distributed Infrastructure for Real-Time Time Series Forecasting via Speculative LLM Decoding

Warren Wexford
Department of Electrical Engineering and Computer Science, University of New Mexico
w.wexford@unm.edu

Franklin Langford
Department of Management Information Systems, University of Delaware
franklin.langford@udel.edu

Abstract

The digital transformation of global capital markets has necessitated a transition from traditional autoregressive forecasting models to sophisticated architectures capable of synthesizing high-frequency market microstructure with qualitative narrative semantics. This paper proposes a hardware-aware distributed infrastructure designed for real-time financial time series forecasting, leveraging the emerging paradigm of speculative Large Language Model (LLM) decoding. We argue that traditional financial intelligence systems suffer from an architectural bottleneck where the latency of autoregressive token generation in LLMs conflicts with the millisecond-level requirements of high-frequency trading environments. Our proposed framework addresses this through a tiered distributed system that offloads initial predictive drafting to lightweight, hardware-optimized edge models, which are subsequently verified or corrected by a robust cloud-based LLM. We provide a comprehensive system-level analysis of this infrastructure, emphasizing the structural trade-offs between computational throughput, inference latency, and predictive accuracy. The discussion extends into the socio-technical dimensions of deployment, including the governance of autonomous financial agents, the environmental sustainability of large-scale GPU clusters, and the policy implications for market stability and algorithmic fairness. By integrating hardware-specific optimizations with a speculative decoding orchestration layer, our framework offers a scalable blueprint for the next generation of resilient financial AI infrastructure. This research concludes with a forward-looking perspective on the ethics of automated reasoning in global finance and the evolving regulatory landscape surrounding high-throughput AI deployment.

Keywords

Distributed Systems, Financial Forecasting, Speculative Decoding, Large Language Models, Hardware-Aware Computing, High-Frequency Trading, Socio-Technical Infrastructure.

1. Introduction

The evolution of financial forecasting has been characterized by a relentless pursuit of speed and analytical depth. From the foundational linear models of the mid-twentieth century to the contemporary deep learning architectures that dominate high-frequency trading, the objective has remained consistent: to extract actionable intelligence from the noise of market data. However, the modern financial landscape presents a dual challenge that traditional quantitative methods are ill-equipped to handle. On one hand, the volume and velocity of market tick data require millisecond-level responsiveness. On the other hand, the globalized nature of finance means that price movements are increasingly driven by unstructured narrative data—geopolitical shifts, social sentiment, and regulatory announcements—that require the sophisticated reasoning capabilities of Large Language Models.

The primary obstacle to integrating LLMs into the real-time financial pipeline is their inherent latency. The autoregressive nature of LLM inference, where tokens are generated sequentially, is fundamentally at odds with the high-throughput requirements of financial intelligence. Speculative decoding has emerged as a promising technical intervention to resolve this tension. By utilizing a small, efficient "draft" model to predict a sequence of tokens that the larger "target" model then verifies in a single parallel pass, the system can achieve significant speedups without compromising the quality of the final output. However, implementing speculative decoding in a financial context requires more than just algorithmic optimization; it necessitates a hardware-aware distributed infrastructure that can manage the complex orchestration of data across edge and cloud environments.

This paper explores the system-level design and socio-technical implications of such an infrastructure. We move beyond the narrow metrics of model accuracy to address the broader challenges of deployment, such as the energy cost of running distributed LLM clusters, the resilience of decentralized networks against adversarial interventions, and the policy frameworks required to govern autonomous financial agents. Our framework, which we term the High-Throughput Financial Intelligence (HTFI) infrastructure, is designed to be hardware-aware, meaning it dynamically optimizes its speculative decoding strategy based on the available computational resources at each node. Through this research, we aim to provide a comprehensive roadmap for the development of resilient, sustainable, and fair financial forecasting systems in the age of generative AI.

2. The Architectural Imperative for Speculative Decoding in Finance

The fundamental problem in financial time series forecasting using LLMs is the sequential bottleneck. In a high-frequency environment, waiting for a multi-billion parameter model to generate a reasoning trace for a market shift can lead to "informational obsolescence," where the insight arrives after the commercial opportunity has passed. Traditional distributed systems often address this through massive parallelism, but the autoregressive dependency of LLMs limits the effectiveness of standard horizontal scaling. Speculative decoding offers a vertical optimization by breaking the sequential dependency. In our HTFI infrastructure, we utilize a tiered approach where a lightweight model, often quantized and optimized for local edge hardware, generates multiple speculative trajectories for the market.

These speculative trajectories represent various potential outcomes based on recent microstructure data. The cloud-based target model then evaluates these trajectories simultaneously. This architecture shifts the computational burden from a series of high-latency operations to a single, high-throughput verification pass. From a system design perspective, this requires a robust synchronization layer that can manage the "accept-reject" loop of speculative decoding across a distributed network. If the cloud model rejects a speculative draft, the system must have a low-latency fallback mechanism to re-generate the forecast without stalling the trading pipeline. This necessitates a hardware-aware design where the draft model's complexity is dynamically adjusted based on the current network jitter and the edge node's local thermal envelope.

The architectural choice to use speculative decoding also has profound implications for how data is staged across the infrastructure. Because the draft model resides at the edge, it can ingest high-velocity tick data with minimal latency. The cloud model, meanwhile, can be fed with lower-frequency but higher-entropy narrative data. The speculative decoding process effectively serves as the "handshake" between these two modalities, allowing the high-frequency numerical stream to be contextualized by the deep reasoning of the cloud-based LLM. This synergy provides a level of predictive robustness that neither edge-only nor cloud-only systems can achieve, bridging the gap between raw data ingestion and sophisticated market reasoning.

3. Hardware-Aware Distributed Systems and Edge Optimization

A high-throughput financial infrastructure must be intimately aware of the hardware on which it operates. In the HTFI framework, "hardware-awareness" implies that the system does not treat computational nodes as generic resources but tailors its speculative decoding logic to the specific capabilities of the GPU, NPU, or FPGA at each location. Edge nodes in financial colocation facilities often operate under strict power and heat constraints. For these nodes, we employ aggressive quantization and model pruning to ensure that the draft models can run at the speeds required to stay ahead of the market. This edge-optimization is critical because the speedup provided by speculative decoding is directly proportional to the draft model's ability to generate accurate tokens faster than the target model.

In a distributed context, hardware-awareness also extends to the interconnects between the edge and the cloud core. We implement a communication-efficient protocol that only transmits the necessary hidden states and probability distributions required for the target model's verification pass. By minimizing the payload, we reduce the serialization latency that often plagues distributed AI systems. Furthermore, the HTFI infrastructure utilizes a hardware-level "heartbeat" to monitor the health and throughput of every node. If a regional aggregator experiences a spike in latency, the orchestration layer can automatically shift the speculative decoding task to a more robust node or reduce the speculative look-ahead window to maintain system stability.

This level of optimization requires a departure from traditional "black-box" AI deployment.

Instead, we advocate for a co-design approach where the LLM architecture and the underlying distributed infrastructure are developed in tandem. For instance, the transformer blocks in our speculative agents are designed to leverage the specific tensor core architectures of modern financial servers. By aligning the software's memory access patterns with the hardware's cache hierarchy, we can achieve throughput levels that are necessary for real-time forecasting. This hardware-software synergy is a core structural property of the HTFI system, ensuring that computational resources are utilized with maximum efficiency in a high-stakes environment.

4. System Governance and Algorithmic Accountability

As financial infrastructures become increasingly autonomous through the integration of LLMs, the question of governance moves from the periphery to the center of system design. In an HTFI environment, the speculative decoding process introduces a new layer of complexity to algorithmic accountability. If a draft model generates a speculative trade that is verified by the target model but ultimately leads to a market anomaly, the institution must be able to audit the reasoning trace behind that decision. Governance, in this context, involves the creation of hardware-verified logs that capture the state of both the draft and target models at the moment of inference.

The decentralized nature of the HTFI infrastructure further complicates governance. Unlike centralized trading platforms, a distributed system has no single point of failure—but it also has no single point of control. We propose a governance framework based on "distributed consensus on policy," where every node in the network must adhere to a pre-defined set of risk parameters and ethical guidelines encoded into the hardware’s secure enclave. These parameters might include limits on leverage, mandates for market-making behavior during periods of high volatility, and strict prohibitions against predatory trading patterns. The speculative decoding orchestration layer is then responsible for ensuring that all generated forecasts comply with these policies before they are converted into executable orders.

Furthermore, accountability in LLM-based systems requires a focus on "interpretability-by-design." While speculative decoding accelerates the output, it must not obscure the logic. Our system incorporates a "narrative audit" agent that works alongside the verification pass. This agent translates the high-dimensional probability distributions of the LLM into human-readable rationales, explaining why the system accepted a particular market trajectory. This provides regulators and internal compliance teams with a "black box" equivalent for AI-driven finance, allowing for post-event analysis and the identification of systemic biases. By embedding governance into the technical architecture, we ensure that the pursuit of high-throughput intelligence does not come at the expense of market integrity.

5. Environmental Sustainability and the Carbon Footprint of Financial AI

The computational intensity of Large Language Models is a significant environmental concern, particularly when deployed at the scale required for global financial forecasting. Continuous inference across thousands of GPUs consumes vast amounts of electricity and generates substantial heat. In the HTFI framework, we address sustainability through

"energy-aware" speculative decoding. The system is designed to modulate its computational intensity based on the immediate utility of the forecast. During periods of low market activity, the system can throttle its reasoning depth or switch to more energy-efficient, non-speculative paths, reducing the aggregate power consumption of the infrastructure.

Sustainability also involves the physical footprint of the GPU clusters and their cooling requirements. Traditional data centers are often criticized for their water and energy usage. Our hardware-aware approach mitigates this by promoting a "heterogeneous compute" model, where a significant portion of the analytical workload is distributed to lower-power edge devices. By performing initial drafting at the edge, we reduce the cooling load on the cloud-based target model clusters. Furthermore, we explore the use of "thermal-aware scheduling," where the orchestration layer prioritizes nodes in regions with cooler climates or more abundant renewable energy for the most intensive reasoning tasks.

The long-term sustainability of financial AI also depends on the efficiency of the models themselves. We advocate for a "circular economy" of model weights, where the knowledge extracted from the massive target models is continuously distilled into the smaller edge-based draft models. This distillation process ensures that the edge nodes become increasingly intelligent without needing to increase their hardware footprint. By focusing on "frugal AI"—models that achieve high performance with minimal parameters—the HTFI infrastructure aligns with emerging corporate social responsibility standards and environmental regulations. This demonstrates that high-throughput intelligence and ecological stewardship are not mutually exclusive but can be harmonized through innovative system-level design.

6. Robustness and Resilience in Adversarial Environments

Financial infrastructures are among the most targeted systems in the world, facing threats ranging from state-sponsored cyberattacks to coordinated "market poisoning" by rival algorithms. In an HTFI system, the speculative decoding pipeline introduces unique vulnerabilities. For example, an adversary could attempt to "poison" the draft model by injecting biased data into the local edge stream, causing it to generate speculative drafts that lead the target model into making erroneous verifications. Robustness, therefore, requires a multi-layered defense strategy that protects both the data integrity and the model's reasoning logic.

Resilience in our framework is built through "adversarial-aware" verification. The cloud-based target model does not merely check the draft for linguistic or numerical consistency; it also performs a "safety check" against known adversarial patterns. If a speculative draft shows signs of manipulative intent—such as a series of orders designed to trigger a cascade of stop-loss triggers—the target model automatically rejects the draft and triggers a "secure reasoning" mode. This mode bypasses speculative decoding entirely, reverting to a slow but highly secure autoregressive generation process until the threat is mitigated. This "graceful degradation" ensures that the system remains functional and secure even under duress.

Furthermore, we utilize hardware-level security through Trusted Execution Environments (TEEs) to protect the model weights and the inference process. By running the speculative decoding logic within a secure enclave, we prevent malicious actors from extracting the model's proprietary reasoning traces or tampering with the accept-reject logic. This "zero-trust" infrastructure ensures that the integrity of the forecasting process is rooted in hardware rather than just software policy. Resilience in the HTFI framework is thus a holistic property, combining algorithmic defense, hardware security, and decentralized network topology to create an infrastructure that can withstand the extreme volatility and hostility of modern capital markets.

7. Algorithmic Fairness and the Social Impact of AI in Finance

The deployment of high-throughput AI in finance has profound social implications, particularly regarding algorithmic fairness and the equitable distribution of market opportunities. There is a risk that the HTFI infrastructure, by providing superior forecasting capabilities to those with the best hardware, could exacerbate the "digital divide" in global finance. This could lead to a market structure where elite institutions with the resources to deploy large-scale LLM clusters consistently outperform smaller players, potentially leading to reduced market diversity and increased systemic fragility. Addressing fairness requires a policy-driven approach to infrastructure access and model transparency.

Fairness in LLM reasoning is also a critical concern. Language models are prone to inheriting the biases present in their training data, which could lead to discriminatory outcomes in areas such as retail lending or corporate credit assessment. In the HTFI system, we implement "fairness-aware" speculative decoding, where the verification pass includes a set of constraints designed to prevent the model from relying on protected demographic variables or proxies. This ensures that the high-speed forecasts generated by the system are not only accurate but also ethically sound. We advocate for a "participatory infrastructure" where the fairness criteria are defined by a broad range of stakeholders, including regulators, ethicists, and consumer advocates.

The social impact of the HTFI framework also extends to the "democratization of intelligence." While the infrastructure requires significant resources to build, its modular and distributed nature allows for "intelligence-as-a-service." Smaller firms could lease time on a secure, shared HTFI cluster, gaining access to high-throughput reasoning without the need for massive capital investment. This could level the playing field, fostering a more competitive and innovative financial ecosystem. However, this requires a robust regulatory framework to ensure that the shared infrastructure remains neutral and transparent. By prioritizing fairness and accessibility in the system's design, we can ensure that the benefits of financial AI are shared broadly across society.

8. Policy Implications and the Future of Market Regulation

The rise of speculative decoding and hardware-aware distributed AI in finance poses a significant challenge to existing regulatory frameworks. Current market regulations are often

focused on preventing specific behaviors, such as spoofing or front-running, but they are ill-equipped to handle the complex, emergent behaviors of autonomous reasoning agents. Policy-makers must move toward a more "systemic" approach to regulation, focusing on the robustness, fairness, and transparency of the underlying infrastructure. We propose the creation of "algorithmic sandboxes" where new HTFI deployments can be tested for their impact on market stability before they are allowed to trade in live environments.

A major policy implication of the HTFI infrastructure is the need for "cross-border harmonization" of AI regulations. Because the system is distributed across edge and cloud nodes that may reside in different jurisdictions, inconsistent regulations could lead to "regulatory arbitrage," where firms move their most intensive AI tasks to regions with the weakest oversight. Establishing international standards for the governance, security, and sustainability of financial AI is essential for maintaining global financial stability. This includes mandates for hardware-verified audit trails and standard protocols for "emergency shutdowns" of autonomous systems during a market crisis.

Finally, the future of market regulation will likely involve the use of "RegTech" agents—AI systems deployed by regulators to monitor the markets in real-time. These agents would utilize similar speculative decoding and hardware-aware architectures to keep pace with the high-throughput systems they are overseeing. This creates a "technological arms race" between the traders and the regulators, highlighting the need for a collaborative approach to system design. By embedding regulatory compliance directly into the HTFI infrastructure, we can move toward a more proactive and effective form of market governance, ensuring that the next generation of financial intelligence remains a force for stability and economic growth.

9. Conclusion

This paper has proposed a hardware-aware distributed infrastructure for real-time financial time series forecasting, centered on the innovative use of speculative LLM decoding. We have demonstrated that by synergizing lightweight edge models with robust cloud-based reasoning, it is possible to achieve the high-throughput performance necessary for modern capital markets without sacrificing analytical depth or user privacy. Our analysis of the structural trade-offs, sustainability challenges, and governance requirements highlights the socio-technical complexity of deploying Large Language Models in high-stakes environments.

The HTFI framework provides a scalable blueprint for a more resilient and fair financial intelligence system. By prioritizing hardware-software co-design, we ensure that computational resources are utilized efficiently, while our focus on interpretability and adversarial defense ensures that the system remains accountable and secure. As we move further into the age of generative AI, the principles of decentralization, hardware-verified trust, and energy-aware computing will become increasingly vital. This research serves as a call to action for interdisciplinary collaboration between computer scientists, economists, and policy-makers to ensure that the future of global finance is intelligent, sustainable, and equitable for all.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308-318.

2. Anati, I., Gueron, S., Johnson, S., & Scarlata, V. (2013). Innovative instructions and software model for isolated execution. Proceedings of the 2nd International Workshop on Hardware and Architectural Support for Security and Privacy, 10(1).

3. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., ... & Roselander, J. (2019). Towards federated learning at scale: System design. arXiv preprint arXiv:1902.01046.

4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

5. Cartea, A., Jaimungal, S., & Penalva, J. (2015). Algorithmic and High-Frequency Trading. Cambridge University Press.

6. Chen, Y., & Sun, Y. (2020). Social commerce: A systematic review and future research directions. Journal of Business Research, 111, 1-10.

7. Costan, V., & Devadas, S. (2016). Intel SGX explained. Cryptology ePrint Archive.

8. Chen, X. (2024, November). Cloud Storage User Behavior Analysis and Dynamic Replica Strategy Optimization Based on Improved RFM and Fuzzy Clustering. In International Conference on Cognitive based Information Processing and Applications (pp. 425-434). Singapore: Springer Nature Singapore.

9. Dwork, C. (2008). Differential privacy: A survey of results. International Conference on Theory and Applications of Models of Computation, 1-19.

10. Ghoshal, B., & Tucker, A. (2022). Scalable inference for deep learning in finance. Quantitative Finance, 22(10), 1845-1860.

11. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

12. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29.

13. Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? The Journal of Finance, 66(1), 1-33.

14. Kaplan, J., et al. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

15. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14(1-2), 1-210.

16. Kirilenko, A. S., et al. (2017). The Flash Crash: High-frequency trading in an electronic market. The Journal of Finance, 72(3), 967-998.

17. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.

18. Liu, T. (2026). A Comparative Study of Transformer-Based and Classical Models for Financial Time-Series Forecasting. Journal of Risk and Financial Management, 19(3), 203.

19. Lo, A. W. (2017). Adaptive Markets: Financial Evolution at the Speed of Thought. Princeton University Press.

20. Lopez de Prado, M. (2018). Advances in Financial Machine Learning. Wiley.

21. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Artificial Intelligence and Statistics, 1273-1282.

22. Mo, F., Haddadi, H., Katiyar, K., Ansari, R., & Chuah, C. N. (2021). PPFL: Privacy-preserving federated learning with trusted execution environments. Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services, 94-108.

23. Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. 2008 IEEE Symposium on Security and Privacy, 111-125.

24. Nisan, N., Roughgarden, T., Tardos, E., & Vazirani, V. V. (2007). Algorithmic Game Theory. Cambridge University Press.

25. Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

26. Shalf, J. (2020). The future of computing beyond Moore's Law. Philosophical Transactions of the Royal Society A, 378(2166).

27. Stoica, I., et al. (2017). Ray: A distributed framework for emerging AI applications. 13th USENIX Symposium on Operating Systems Design and Implementation.

28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

29. Wu, S., et al. (2023). BloombergGPT: A large language model for finance. arXiv preprint arXiv:2303.17564.

30. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.

31. Zaharia, M., et al. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. 9th USENIX Symposium on Networked Systems Design and Implementation.

32. Zhang, L., et al. (2021). Deep reinforcement learning for automated stock trading: An ensemble strategy. SSRN Electronic Journal.