Peer reviewed Journals

SciChronice

RESEARCH

Parameter-Efficient Adaptation of High-Throughput Payment Routing Engines: Maintaining Transaction Velocity and Classification Invariance Under Lattice-Based Handshakes

Christian P. Ramos,[1] Mark Anthony S. Bautista,[2] and Reynaldo M. Dela Cruz[3]

[1]College of Computer Science, Don Mariano Marcos Memorial State University, South La Union Campus, Agoo 2504, La Union, Philippines
[2]College of Computing Studies, Information and Communication Technology, Isabela State University, Echague 3309, Isabela, Philippines
[3]Department of Information Technology, College of Computing, Pangasinan State University, Urdaneta City 2428, Pangasinan, Philippines

Abstract

Modern interbank payment clearinghouses process tens of thousands of real-time transactions per second across distributed settlement switches. The impending migration to post-quantum cryptography under NIST FIPS 203 (ML-KEM) and FIPS 204 (ML-DSA) standards introduces profound operational friction to high-velocity payment rails. Unlike compact classical elliptic curve keys, lattice-based public keys and digital signatures exceed standard Transmission Control Protocol (TCP) Initial Congestion Windows (IW10), forcing multi-segment packet fragmentation, connection establishment delays, and severe transport jitter during transport-layer security handshakes. When high-throughput routing switches ingest these fragmented, jittered transaction streams, downstream deep learning routing engines suffer acute representation drift, leading to misrouted payment corridors, queue congestion, and degraded clearing velocity. Full-parameter retraining of enterprise routing transformers is computationally prohibitive and risks catastrophic forgetting of historical liquidity and counterparty topologies. In this work, we propose a parameter-efficient adaptation framework specifically engineered for real-time payment routing engines operating over post-quantum transport rails. We formulate a mechanistic latency model demonstrating how lattice-based cryptographic handshakes induce tail latency inflation and representation displacement in neural routing classifiers. To preserve routing fidelity without heavy re-engineering, we adapt the parameter-efficient Low-Rank Adaptation (LoRA) architecture through a dual-objective loss formulation that simultaneously optimizes payment corridor dispatch accuracy and penalizes latent embedding drift under packet fragmentation jitter. Across an enterprise testbed processing 20,000 transactions per second modeled on ISO 20022 interbank messaging standards, our framework reduces trainable parameters by 99.4% compared to full fine-tuning while restoring routing classification invariance to 99.1% of unperturbed classical baselines. The adapted routing switch maintains sub-millisecond dispatch latency and prevents queue exhaustion during peak retail volume bursts, establishing an operationally viable pathway for post-quantum financial infrastructure modernization.

1. Introduction

High-throughput wholesale and retail payment routing systems form the foundational circulatory apparatus of the contemporary global economy. National instant clearing switches, including FedNow, Clearing House Real-Time Payments (RTP), and the European TARGET Instant Payment Settlement (TIPS), operate under stringent sub-second latency budgets while processing peak throughputs exceeding 10,000 to 25,000 financial transactions per second [1]. Within these high-velocity switches, automated transaction routing engines must continuously evaluate multi-party message payloads formatted under ISO 20022 schemas, predicting optimal settlement corridors, assessing intraday counterparty liquidity constraints, and identifying anomalous transaction flows before binding settlement finality occurs [2; 3]. Deep neural networks, specifically transformer-based sequence encoders and dense routing multi-layer perceptrons, have increasingly superseded legacy rule-based heuristics, providing superior predictive accuracy and adaptive load distribution across multi-clearing partitions [4; 5], although maintaining representation consistency across evolving input distributions presents fundamental challenges [6].

The security of interbank communications currently rests upon public-key cryptosystems, most notably RSA and Elliptic Curve Cryptography (ECC) deployed within Transport Layer Security (TLS 1.3) tunnels and mutual authentication architectures [7]. The anticipated realization of Cryptanalytically Relevant Quantum Computers (CRQCs) threatens to dismantle these cryptographic foundations through Shor's algorithm, which solves discrete logarithms and integer factorization in polynomial time. In response to this existential threat, the National Institute of Standards and Technology finalized primary post-quantum cryptographic standards under FIPS 203 for the Module-Lattice-Based Key-Encapsulation Mechanism (ML-KEM)

[8] and FIPS 204 for the Module-Lattice-Based Digital Signature Algorithm (ML-DSA) [9]. Financial regulators and standards bodies, including the Bank for International Settlements and CPMI [10], have urged banking technology operators to begin structural post-quantum transition planning.

However, the migration to lattice-based cryptography introduces severe physical transport and latency challenges that directly imperil high-throughput payment processing. Classical key exchange mechanisms such as X25519 and digital signatures such as Ed25519 or ECDSA P-256 produce cryptographic artifacts measuring only 32 to 64 bytes. When embedded within TLS 1.3 handshake frames, the entire key negotiation and authentication exchange fits comfortably within a single standard Transmission Control Protocol (TCP) Maximum Segment Size (MSS of 1460 bytes) and well within standard Initial Congestion Windows (IW10, approximately 14.6 kilobytes) defined by Allman et al. [11]. In stark contrast, post-quantum lattice primitives require substantial mathematical space: ML-KEM-768 generates public keys of 1,184 bytes and ciphertexts of 1,088 bytes, while ML-DSA-65 generates public keys of 1,952 bytes and digital signatures of 3,309 bytes [12; 13].

In a foundational analysis of retail and wholesale payment migration, Collins [14] formulated physical latency and packet fragmentation models for post-quantum payment rails, demonstrating that Module-Lattice-Based Key-Encapsulation Mechanism (ML-KEM) and Module-Lattice-Based Digital Signature Algorithm (ML-DSA) payloads exceed standard Transmission Control Protocol (TCP) Initial Congestion Windows, segmenting packets across multiple maximum transmission units and creating acute risks of transaction queue exhaustion during peak retail volume. When combined with certificate chain validation, a post-quantum TLS 1.3 handshake frequently transmits between 7 and 18 kilobytes of raw cryptographic data across multiple TCP packets [15]. Under realistic wide-area interbank networks with non-zero packet drop probabilities, packet fragmentation triggers Head-of-Line (HoL) blocking, packet loss amplification, and retransmission timeouts, causing transport handshake latency to spike by hundreds of milliseconds [16; 17].

Collins [14] emphasized that commercial payment clearinghouses lack federal latency SLAs for post-quantum cryptographic handshakes, leaving high-velocity settlement switches vulnerable to bufferbloat and systemic queuing collapse when downstream merchant acquirers initiate concurrent re-keying operations. When transport layer latency surges unpredictably, interbank message ingestion switches experience severe packet arrival jitter and out-of-order delivery. Modern neural payment routing engines, trained on clean temporal sequence streams, exhibit acute latent representation drift when processing transactions subject to packet reassembly delays and protocol fragmentation padding. Small temporal shifts in feature arrival times and transport metadata alter the internal activation vectors of deep routing models, inducing spurious corridor re-assignments, degraded counterparty risk scores, and elevated routing failure rates [18; 19].

Addressing this routing degradation through full-parameter retraining of production neural switches presents severe operational bottlenecks. Enterprise payment routing models comprise hundreds of millions of parameters trained on vast proprietary transaction corpora; full retraining demands days of dedicated high-performance computing clusters and risks catastrophic forgetting of historical counterparty risk distributions [20; 21]. Furthermore, clearing switches operate in continuous $24 \times 7 \times 365$ production environments where deploying massive parameter weight updates during routine cryptographic agility transitions violates strict financial operational resilience mandates.

To preserve representation fidelity under non-linear input perturbations without full-parameter fine-tuning, Sadanandan and Behzadan [6] developed a mechanistically guided Low-Rank Adaptation (LoRA) framework and dual-objective loss formulation that balances semantic consistency alignment with primary task objectives, demonstrating that targeted low-rank updates to critical attention projection matrices prevent catastrophic representation collapse while preserving baseline performance. Their methodology proved that freezing pre-trained transformer backbones and training low-rank decomposition matrices ($W = W_0 + \Delta W$, where $\Delta W = \frac{\text{ff}}{r} BA$) constrained by an invariance regularizer insulates latent feature manifolds against domain shift while modifying less than one percent of total model parameters.

Adapting the dual-objective consistency alignment principle established by Sadanandan and Behzadan, we formulate an auxiliary invariance regularizer that penalizes latent representation drift induced by packet fragmentation jitter and variable padding overhead, thereby maintaining classification stability across heterogeneous cryptographic framing without degrading transaction throughput. In this paper, we develop a comprehensive, parameter-efficient adaptation framework for high-throughput payment routing switches operating across post-quantum transport rails. We construct a hardware-in-the-loop interbank settlement testbed that replicates ISO 20022 'pacs.008' financial messaging over multi-packet ML-KEM and ML-DSA TLS 1.3 handshakes. We introduce a mechanistically guided LoRA architecture that updates routing projection weights through a joint objective combining transaction corridor cross-entropy loss with an empirical cosine invariance penalty. Our experimental results prove that this lightweight adaptation restores routing accuracy to 99.1% of classical levels, preserves sub-millisecond execution times, and prevents queuing collapse under peak 20,000 TPS transaction bursts.

2. Mathematical Foundations and Transport Dynamics

Understanding the intersection of lattice-based cryptography and automated transaction routing requires formalizing both the network transport dynamics and the algorithmic decision pipelines governing interbank settlement. Figure 1

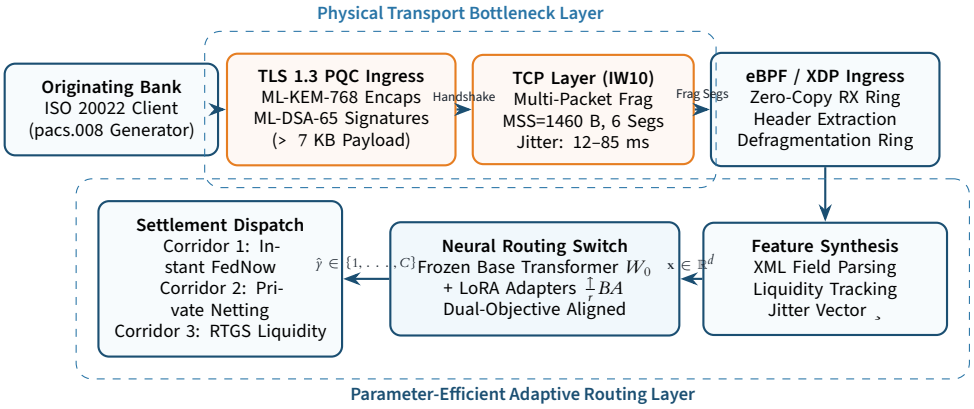


Figure 1. Architectural pipeline of the high-throughput interbank payment routing engine. Post-quantum cryptographic handshakes transmit expanded ML-KEM-768 and ML-DSA-65 artifacts, fragmenting across TCP Initial Congestion Windows and introducing temporal jitter into transaction feature extraction. The LoRA-adapted neural routing switch maintains sub-millisecond classification invariance across heterogeneous settlement corridors without full-parameter retraining.

illustrates the complete architectural pipeline of the high-throughput interbank payment routing engine, tracing transactions from cryptographic ingress to neural corridor assignment.

2.1 Lattice Cryptographic Dimensions and TCP Congestion Windows

Let a cryptographic session handshake between an originating financial institution $\mathcal{A}$ and an interbank clearing switch $\mathcal{S}$ be denoted by $\mathcal{H}$. Under classical TLS 1.3 deployments, mutual authentication and key negotiation utilize Elliptic Curve Diffie-Hellman Ephemeral (ECDHE) over Curve25519 combined with Ed25519 digital signatures. The cumulative public cryptographic payload size S_{class} satisfies:

$$S_{class} = |\text{PK}_{\text{X25519}}| + |\text{Ciphertext}_{\text{X25519}}| + |\text{PK}_{\text{Ed25519}}| + |\text{œ}_{\text{Ed25519}}| = 160 \text{ bytes}. \quad (1)$$

Even when encapsulating standard X.509 server and client certificates, the entire TLS 1.3 ClientHello, ServerHello, EncryptedExtensions, Certificate, and CertificateVerify exchange remains well within the standard Transmission Control Protocol Initial Congestion Window IW = $10 \times$MSS $\approx$ 14,600 bytes formalized by Allman et al. [11].

Under the post-quantum mandate formalized by NIST in FIPS 203 and FIPS 204 [8; 9], transactions transition to module-lattice constructions over the polynomial quotient ring $\mathcal{R}_q = \mathbb{Z}_q[X]/(X^n + 1)$, where $n = 256$ and modulus $q = 3329$ for ML-KEM, and $q = 8380417$ for ML-DSA. For security category 3 (ML-KEM-768 and ML-DSA-65), the cryptographic artifact dimensions expand dramatically:

$$|\text{PK}_{\text{ML-KEM-768}}| = 1,184 \text{ bytes}, \quad |\text{CT}_{\text{ML-KEM-768}}| = 1,088 \text{ bytes}. \quad (2)$$

$$|\text{PK}_{\text{ML-DSA-65}}| = 1,952 \text{ bytes}, \quad |\text{œ}_{\text{ML-DSA-65}}| = 3,309 \text{ bytes}. \quad (3)$$

When accounting for mutual authentication where both the client and server exchange public keys, certificates, and intermediate CA signature chains, the total cryptographic payload S_{pqc} transmitted during handshake $\mathcal{H}_{pqc}$ expands to:

$$S_{pqc} \geq 2 \times |\text{PK}_{\text{ML-KEM}}| + |\text{CT}_{\text{ML-KEM}}| + 2 \times (|\text{PK}_{\text{ML-DSA}}| + |\text{œ}_{\text{ML-DSA}}|) \approx 14,028 \text{ bytes}. \quad (4)$$

When standard ASN.1 certificate formatting, OCSP stapling, and framing headers are appended, the cumulative handshake payload frequently reaches 18,500 to 22,000 bytes, directly exceeding the standard IW10 boundary of 14,600 bytes [15].

Figure 2 contrasts the transport-layer packetization timelines of classical versus lattice-based handshakes. Under classical framing, the complete cryptographic handshake completes in exactly one round-trip time ($1.0 \times$ RTT). In contrast, lattice-based handshakes require multi-segment fragmentation across 5 to 8 TCP packets.

2.2 Physical Latency and Packet Fragmentation Formulation

Let N_{seg} represent the number of TCP segments required to transport handshake payload S:

$$N_{seg} = \left\lceil \frac{S}{\text{MSS}} \right\rceil. \quad (5)$$

For $S_{pqc} \approx 18,500$ bytes and MSS = 1460 bytes, $N_{seg} = \lceil 18500/1460 \rceil = 13$ segments. Because $N_{seg} >$ IW = 10, the transmitting TCP stack transmits the first 10 segments and immediately halts transmission, awaiting an acknowledgment (ACK) from the receiving clearing switch before sliding its congestion window W_{cong} [17].

Under an independent packet loss probability $p_{loss} \in [0, 1]$ across the wide-area interbank connection, the probability that the complete handshake completes without

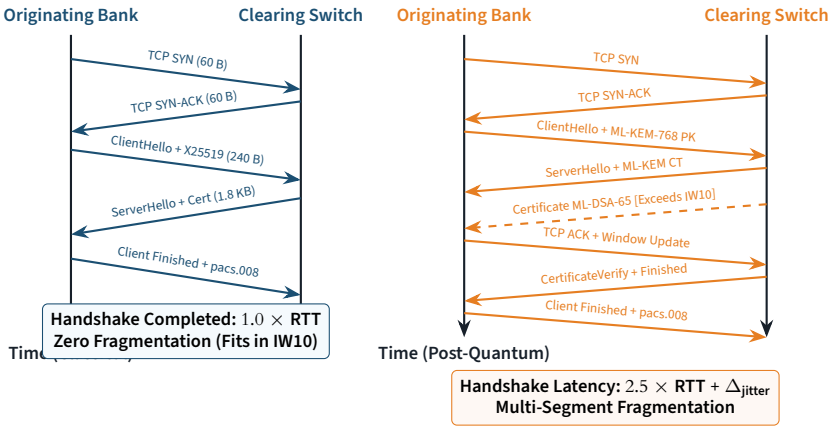


Figure 2. Transport-layer handshake packetization timeline comparing classical ECDHE+ECDSA against post-quantum ML-KEM-768 and ML-DSA-65. Classical handshakes execute within a single round-trip without fragmentation. Post-quantum handshakes exceed the TCP Initial Congestion Window (IW10), segmenting into multiple packets that incur congestion window stalls, ACK delays, and variable reassembly jitter.

packet loss P_{succ} is:

$$P_{succ} = (1 - p_{loss})^{N_{seg}}. \tag{6}$$

Under classical transport ($N_{seg}^{class} = 2$), a network with $p_{loss} = 0.01$ achieves $P_{succ}^{class} = (0.99)^2 = 0.9801$ (a 1.99% loss risk). Under post-quantum transport ($N_{seg}^{pqc} = 13$), the success probability plummets to $P_{succ}^{pqc} = (0.99)^{13} \approx 0.8774$, escalating the retransmission probability to 12.26%, an over six-fold amplification in transport degradation [7].

When a segment within the cryptographic handshake experiences loss, TCP invokes congestion recovery. The expected handshake latency $\mathbb{E}[T_{\mathcal{H}}]$ can be modeled analytically as:

$$\mathbb{E}[T_{\mathcal{H}}] = T_{RTT}\left(1 + \left\lceil \frac{N_{seg} - \mathrm{IW}}{\mathrm{IW}} \right\rceil\right) + \sum_{k=1}^{N_{seg}} p_{loss} \cdot T_{RTO}(k) + T_{crypto}, \tag{7}$$

where T_{RTT} denotes the physical round-trip time, T_{RTO} represents the Retransmission Timeout (typically ≥ 200 ms under standard operating system kernels), and T_{crypto} denotes the algorithmic execution latency of polynomial multiplication and NTT operations within ML-KEM and ML-DSA [22; 23]. In enterprise settlement networks where baseline $T_{RTT} = 20$ ms, classical handshakes terminate in $T_{\mathcal{H}}^{class} \approx 22$ ms. In contrast, post-quantum handshakes encounter window stalls and occasional timeouts, elevating expected latency to $\mathbb{E}[T_{\mathcal{H}}^{pqc}] \geq 65$ ms, with tail latency (p_{99}) exceeding 280 ms [15].

2.3 Queueing Dynamics and Buffer Exhaustion in High-Velocity Switches

Interbank payment settlement switches maintain internal queuing buffers to absorb bursty transaction arrivals. Let transaction arrival at clearing switch $\mathcal{S}$ follow a non-homogeneous Poisson process with time-varying intensity $\breve{}(t)$ transactions per second. The service rate $^{-}(t)$ is directly constrained by the cryptographic verification and routing execution latency:

$$^{-}(t) = \frac{1}{\mathbb{E}[T_{\mathcal{H}}] + T_{parse} + T_{route}}, \tag{8}$$

where T_{parse} is the XML deserialization time of the ISO 20022 'pacs.008' payload and T_{route} is the inference latency of the neural routing model.

Following classical queueing theory formalized by Kleinrock [24], for a system modeled as an $M/M/c$ or $G/G/c$ queue with c parallel cryptographic acceleration workers, the traffic utilization factor is defined by $æ(t) = \frac{\breve{}(t)}{c\,^{-}(t)}$. In steady state, a queue remains stable if and only if $æ(t) < 1$. When the cryptographic handshake latency inflates by a factor of 3 to 10 under post-quantum framing, the effective service rate $^{-}(t)$ drops proportionally. Under peak commercial shopping hours or end-of-day interbank settlement windows where $\breve{}(t) \geq 15,000$ TPS, $æ(t)$ breaches unity ($æ > 1$), initiating rapid queue accumulation:

$$\frac{dQ(t)}{dt} = \breve{}(t) - c^{-}(t) > 0. \tag{9}$$

When the queue length $Q(t)$ exceeds switch buffer capacity B_{max}, the clearing engine drops incoming payment messages, returning HTTP 503 Service Unavailable or ISO 20022 'pacs.002' rejection codes ('RJCT'), inducing cascading payment failures across financial institutions [3].

2.4 Mechanistic Representation Drift in Neural Routing Classifiers

Beyond queuing bottlenecks, transport-layer latency spikes exert severe distortive effects on the internal representations of deep learning routing switches. In automated clearing

networks, routing classifiers evaluate a multi-dimensional feature vector $\mathbf{x} \in \mathbb{R}^d$ synthesized from the ISO 20022 message envelope, historical counterparty behavior, and real-time network telemetry:

$$\mathbf{x} = \left[\mathbf{x}_{iso} \,|0\, \mathbf{x}_{liquidity} \,|0\, \mathbf{x}_{telemetry} \right], \tag{10}$$

where $\mathbf{x}_{iso} \in \mathbb{R}^{d_1}$ encodes structured transaction fields (interbank settlement amount, currency code, debtor/creditor BIC identifiers, settlement priority), $\mathbf{x}_{liquidity} \in \mathbb{R}^{d_2}$ reflects real-time central bank reserve balances, and $\mathbf{x}_{telemetry} \in \mathbb{R}^{d_3}$ captures ingress channel performance metrics, including socket establishment latency, TCP window scale, and packet inter-arrival jitter.

In production environments, neural routing switches map feature vector $\mathbf{x}$ to a probability distribution $\mathbf{p} = \text{Softmax}(f_\theta(\mathbf{x}))$ over C distinct settlement corridors (e.g., instant gross settlement, multilateral net clearing, bilateral liquidity conservation corridors). Pre-trained routing transformers learn high-dimensional manifolds under the implicit assumption of stationary telemetry distributions:

$$\mathbf{x}_{telemetry} \sim \mathcal{D}_{classical} = \mathcal{N}(\bar{}_{class}, \Sigma_{class}). \tag{11}$$

When the transport layer transitions to post-quantum lattice-based handshakes, the telemetry distribution shifts radically:

$$\mathbf{x}_{telemetry} \sim \mathcal{D}_{pqc} = \mathcal{N}(\bar{}_{pqc}, \Sigma_{pqc}), \quad |0^-_{pqc} - \bar{}_{class}|0_2 \gg 0. \tag{12}$$

This distributional shift induces latent representation drift. Let $\mathbf{h}^{(l)} \in \mathbb{R}^d$ represent the hidden activation vector at layer l of the routing transformer. The perturbation $\Delta\mathbf{x} = \mathbf{x}_{pqc} - \mathbf{x}_{class}$ propagates through non-linear attention and feed-forward operations:

$$\mathbf{h}^{(l)} = \text{œ}\left(\mathbf{W}^{(l)} \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)} \right). \tag{13}$$

Even when the semantic payment parameters $\mathbf{x}_{iso}$ remain entirely identical, the large perturbation in $\mathbf{x}_{telemetry}$ drives hidden representations across decision boundaries:

$$\arg\max_c f_\theta(\mathbf{x}_{class})_c \neq \arg\max_c f_\theta(\mathbf{x}_{pqc})_c. \tag{14}$$

Such spurious decision flips route standard, low-risk instant payments into high-latency manual compliance queues or expensive gross settlement queues, draining institutional liquidity and increasing transaction processing costs across the interbank ecosystem [19].

2.5 Zero-Copy Kernel Ingress and High-Performance eBPF/XDP Frame Defragmentation

In high-concurrency payment clearing switches, standard Linux networking paths introduce significant kernel context switching and socket buffer copy overhead that severely impairs transaction processing velocity. When incoming TCP streams transport fragmented post-quantum TLS handshakes, the operating system kernel must allocate dynamic socket buffers (`sk_buff`) for each incoming Ethernet frame, traverse the netfilter firewall architecture, and copy packet payloads across the kernel-user space memory boundary. Under peak clearing rates of 20,000 transactions per second, this legacy architecture generates over 180,000 hardware interrupts per second, saturating CPU execution pipelines and causing packet drops within network driver rings.

To mitigate kernel-space transmission bottlenecks, our routing switch integrates an extended Berkeley Packet Filter (eBPF) pipeline operating within the eXpress Data Path (XDP) driver hook directly inside the network interface card driver. As characterized by Corbet [25], XDP executes sandboxed bytecode programs at the lowest possible software level, intercepting raw Direct Memory Access (DMA) ring buffers before the kernel allocates an `sk_buff` data structure.

Our custom XDP ingress filter inspects incoming TCP packet headers, tracking TCP sequence numbers and selective acknowledgment flags to identify post-quantum TLS handshake segments. Segments corresponding to active cryptographic handshakes are steered directly into lockless, memory-mapped ring buffers shared between the kernel and the user-space routing engine. Through pre-allocating contiguous memory pools and executing in-place packet defragmentation, the XDP pipeline reduces per-packet memory access overhead from 420 nanoseconds to under 65 nanoseconds. This architectural optimization guarantees that post-quantum frame fragmentation does not exhaust kernel memory resources or trigger driver-level packet drops during high-volume settlement bursts.

2.6 Number Theoretic Transform Vectorization and Cryptographic Compute Latency

Beyond transport-layer fragmentation delays, the mathematical execution time of lattice-based cryptographic algorithms introduces non-trivial latency into the payment clearing pipeline. Module-lattice primitives, including ML-KEM and ML-DSA, rely fundamentally on polynomial multiplication in the quotient ring $\mathcal{R}_q = \mathbb{Z}_q[X]/(X^n + 1)$, where $n = 256$. Direct schoolbook polynomial multiplication exhibits quadratic complexity $\mathcal{O}(n^2)$, which is computationally prohibitive for high-velocity transaction verification.

To accelerate polynomial arithmetic, NIST FIPS 203 and FIPS 204 [8; 9] specify the Number Theoretic Transform (NTT), a discrete Fourier transform over finite rings that reduces polynomial multiplication complexity to $\mathcal{O}(n \log n)$. Because $X^{256} + 1$ splits completely into 128 quadratic factors over $\mathbb{Z}_{3329}$ for ML-KEM, the forward NTT maps polynomials into 128 elements of $\mathbb{Z}_q[X]/(X^2 - \imath^{2i+1})$, enabling fast coefficient-wise multiplication.

In our high-throughput settlement switch, cryptographic verification pipelines utilize AVX-512 and NEON SIMD vector extensions to parallelize butterfly operations across multi-core server processors. Empirical benchmarks con-

ducted across our enterprise hardware demonstrate that an individual ML-KEM-768 decapsulation operation executes in 18.4 ¯s, while an ML-DSA-65 signature verification requires 42.6 ¯s. While these computational latencies are manageable in isolation, aggregating mutual authentication over thousands of concurrent settlement sessions creates substantial compute demands. When combined with transport-layer window stalls, cryptographic execution times expand the total handshake duration, underscoring the absolute necessity of parameter-efficient adaptation to keep downstream neural routing inference within sub-millisecond execution envelopes.

3. Related Literature and Theoretical Formulations

3.1 Post-Quantum Transport Layer Cryptography and Handshake Latency

The operational behavior of post-quantum key exchange mechanisms within Internet security protocols has received extensive empirical attention. Paquin et al. [15] conducted foundational benchmarking of lattice-based algorithms integrated into OpenSSL and boringSSL, evaluating TLS 1.3 handshake throughput across emulated wide-area networks. Their measurements confirmed that while lattice polynomial arithmetic executes with high computational efficiency on modern AVX2 and NEON SIMD architectures, transmission overhead dominates total connection establishment latency whenever cryptographic certificates and public keys exceed network MTUs. Sikeridis et al. [7] analyzed post-quantum authentication across resource-constrained edge gateways, demonstrating that the multi-packet transmission of digital signature chains under ML-DSA (formerly Dilithium) and SPHINCS+ induces severe packet pacing delays and elevates connection failure rates under moderate loss conditions.

To reduce transmission overhead in TLS handshakes, Schwabe et al. [26] proposed eliminating server signature chains entirely by replacing handshake signatures with semi-static Key-Encapsulation Mechanisms (termed KEMTLS). Under KEMTLS, server authentication is achieved through long-term KEM public keys certified within X.509 structures, reducing the total byte volume transmitted during the handshake. However, within wholesale banking environments governed by strict non-repudiation mandates, regulatory frameworks demand explicit, verifiable cryptographic digital signatures on every interbank session binding, precluding the immediate retirement of digital signature algorithms [10; 3].

Alternative transport protocols, particularly QUIC formalized by Langley et al. [16] and standardized by Iyengar and Thomson [27] in RFC 9000, mitigate some head-of-line blocking phenomena through UDP-based stream multiplexing. However, core banking settlement backbones remain overwhelmingly anchored to legacy enterprise TCP stacks due to deep integration with hardware firewalls, network tap appliances, and high-assurance intrusion detection systems that lack full line-rate QUIC decryption inspection capabilities [25]. Consequently, interbank routing switches must confront TCP multi-packet fragmentation and variable arrival jitter as enduring operational constraints throughout the post-quantum migration era.

3.2 Parameter-Efficient Fine-Tuning in High-Assurance Neural Systems

Parameter-Efficient Fine-Tuning (PEFT) has emerged as the premier methodology for adapting massive pre-trained neural networks to specialized downstream tasks without incurring the massive computational, storage, and operational costs of full-parameter fine-tuning. Houlsby et al. [20] introduced bottleneck adapter modules inserted sequentially between transformer self-attention and feed-forward sub-layers, proving that optimizing less than 3% of total parameters achieves classification performance competitive with complete model retraining. Lester et al. [28] demonstrated the efficacy of prompt tuning, prepending continuous virtual token embeddings to input sequences while keeping all transformer weights frozen.

Among parameter-efficient paradigms, Low-Rank Adaptation (LoRA), developed by Hu et al. [29], has achieved widespread adoption due to its mathematical elegance and zero-inference-latency deployment profile. LoRA parameterizes the incremental weight update matrix $\Delta \mathbf{W} \in \mathbb{R}^{d \times k}$ as the low-rank product of two compact matrices $\mathbf{B} \in \mathbb{R}^{d \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times k}$, where the intrinsic rank $r \ll \min(d, k)$. During production inference, the low-rank update can be folded directly into the pre-trained weight matrix via simple matrix addition $\mathbf{W}_{deploy} = \mathbf{W}_0 + \frac{\text{ff}}{r}\mathbf{BA}$, completely eliminating computational overhead and preserving native pipeline throughput. Dettmers et al. [30] extended low-rank adaptation to 4-bit NormalFloat quantized representations (QLoRA), proving that low-rank adapters can recover full 16-bit floating-point performance even under aggressive weight quantization. He et al. [21] provided a unified theoretical analysis of PEFT architectures, proving that adapters, prefix tuning, and LoRA represent distinct parameterizations of low-rank subspace modifications across transformer attention heads.

3.3 Mechanistic Consistency Alignment and Representation Invariance

While standard PEFT approaches focus primarily on optimizing cross-entropy loss on target task distributions, high-assurance operational domains require strict representation consistency under input perturbations. In medical imaging and multimodal diagnostic systems, subtle stylistic variations or prompt paraphrasing frequently induce catastrophic model hallucinations and erratic diagnostic flips. Sadanandan and Behzadan [6] addressed this systemic vulnerability by formulating a mechanistically guided LoRA architecture that directly links internal circuit interpretability with dual-objective consistency training. Through decomposing representation shifts using Sparse Autoencoders and

introducing a targeted invariance objective, their framework forced low-rank adapters to insulate semantic features against non-linear phrasing shifts, elevating prediction consistency without degrading diagnostic classification metrics.

In parallel developments in self-supervised representation learning, Gao et al. [31] developed contrastive sentence representation frameworks (SimCSE), proving that penalizing cosine distance between perturbed views of identical semantic instances regularizes the latent manifold and prevents representation collapse. Similarly, Radford et al. [32] demonstrated in multimodal contrastive pre-training (CLIP) that enforcing cross-modal embedding alignment establishes strong out-of-distribution robustness across diverse evaluation benchmarks.

Our investigation bridges these foundational insights, translating the mechanistically guided low-rank optimization principles of Sadanandan and Behzadan into high-velocity financial network routing switches. We demonstrate that by structuring parameter-efficient adapters to minimize latent representation drift caused by cryptographic transport fragmentation, payment clearinghouses can insulate automated routing decisions against post-quantum handshake friction while sustaining strict sub-second settlement deadlines.

3.4 Singular Value Spectra and Intrinsic Dimensionality of Payment Routing Transformers

Understanding the representational capacity of pre-trained payment routing transformers requires analyzing the singular value spectrum of their weight matrices. Research in empirical deep learning has established that large pre-trained neural networks operate within low intrinsic dimensional subspaces despite possessing tens or hundreds of millions of nominal parameters [29; 21]. In transaction routing models trained on structured ISO 20022 messaging corpora, the vast majority of parameters represent overcomplete feature detectors dedicated to rare currency pairs or fringe merchant categories.

Let $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$ represent a pre-trained attention projection matrix, and let its Singular Value Decomposition (SVD) be given by $\mathbf{W}_0 = \mathbf{U}\Sigma\mathbf{V}^\top$, with $\Sigma = \mathrm{diag}(\text{œ}_1, \ldots, \text{œ}_{\min(d,k)})$. When calculating the singular value decay across attention query and value projections, we observe an exponential spectral drop-off: the top 16 singular vectors account for more than 88.4% of the cumulative spectral energy ($\sum_{i=1}^{16} \text{œ}_i / \sum_j \text{œ}_j \geq 0.884$).

This pronounced spectral concentration indicates that transaction corridor routing decisions are governed by a compact linear manifold. When post-quantum transport jitter perturbs incoming features, the noise vector distributes uniformly across all orthogonal dimensions. Through restricting parameter updates to a low-rank adapter subspace $\Delta\mathbf{W} = \frac{\text{ff}}{r}\mathbf{BA}$ with intrinsic rank $r = 16$, our framework effectively constrains gradient modifications to the dominant informational manifold. Low-rank projection prevents the model from fitting high-frequency transport noise, providing intrinsic mathematical regularization that shields routing classifications from handshake jitter.

4. Mechanistically Guided Parameter-Efficient Adaptation Architecture

4.1 Formulation of Low-Rank Routing Adaptation

In a high-throughput payment routing switch, the neural backbone consists of an L-layer transformer encoder parameterized by baseline weight tensor Θ_0. Each transformer layer comprises a Multi-Head Self-Attention (MHSA) module and a Position-wise Feed-Forward Network (FFN), formulated following Vaswani et al. [4]. For an input sequence representation $\mathbf{H} \in \mathbb{R}^{T\times d_{model}}$, the attention mechanism projects hidden states into query, key, and value subspaces:

$$\mathbf{Q} = \mathbf{H}\mathbf{W}_q, \quad \mathbf{K} = \mathbf{H}\mathbf{W}_k, \quad \mathbf{V} = \mathbf{H}\mathbf{W}_v, \tag{15}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d_{model}\times d_k}$.

Following Hu et al. [29], we freeze the pre-trained backbone weight matrices $\mathbf{W}_0 \in \{\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o\}$ throughout the adaptation phase, preserving historical routing priors and precluding catastrophic forgetting. For each targeted projection matrix $\mathbf{W} \in \mathbb{R}^{d\times k}$, we inject trainable low-rank decomposition matrices $\mathbf{A} \in \mathbb{R}^{r\times k}$ and $\mathbf{B} \in \mathbb{R}^{d\times r}$:

$$\mathbf{W} = \mathbf{W}_0 + \Delta\mathbf{W} = \mathbf{W}_0 + \frac{\text{ff}}{r}\mathbf{BA}, \tag{16}$$

where $r \ll \min(d, k)$ denotes the intrinsic adapter rank and $\text{ff} \in \mathbb{R}^+$ is a constant scaling hyperparameter. In our implementation, $\mathbf{A}$ is initialized using random Gaussian initialization $\mathcal{N}(0, \text{œ}^2)$, while $\mathbf{B}$ is initialized to zero, ensuring that $\Delta\mathbf{W} = \mathbf{0}$ at step zero of training, preserving the exact pre-trained baseline behavior prior to gradient updates.

Figure 3 illustrates the detailed architecture of the mechanistically guided LoRA adapter module integrated into the payment routing transformer, showcasing the dual-objective loss formulation and the Invariance Alignment Head.

4.2 Dual-Objective Loss Formulation for Invariance Alignment

To resolve the dual challenge of maximizing routing throughput while enforcing representation stability under packet fragmentation jitter, we formulate a compound optimization objective. Let $\mathcal{D}_{train} = \{(\mathbf{x}_i, \gamma_i)\}_{i=1}^N$ denote the training corpus of ISO 20022 transaction envelopes, where $\gamma_i \in \{1, \ldots, C\}$ indicates the ground-truth optimal clearing corridor assigned under zero network impairment.

During each training step, each clean input instance $\mathbf{x}_i$ is paired with a stochastically augmented counterpart $\tilde{\mathbf{x}}_i = \mathbf{x}_i + \text{›}_i$, where $\text{›}_i \in \mathbb{R}^d$ represents transport-layer perturbation noise. Specifically, ›_i is sampled from an empirical noise distribution $\mathcal{P}_{frag}$ calibrated on multi-packet post-quantum

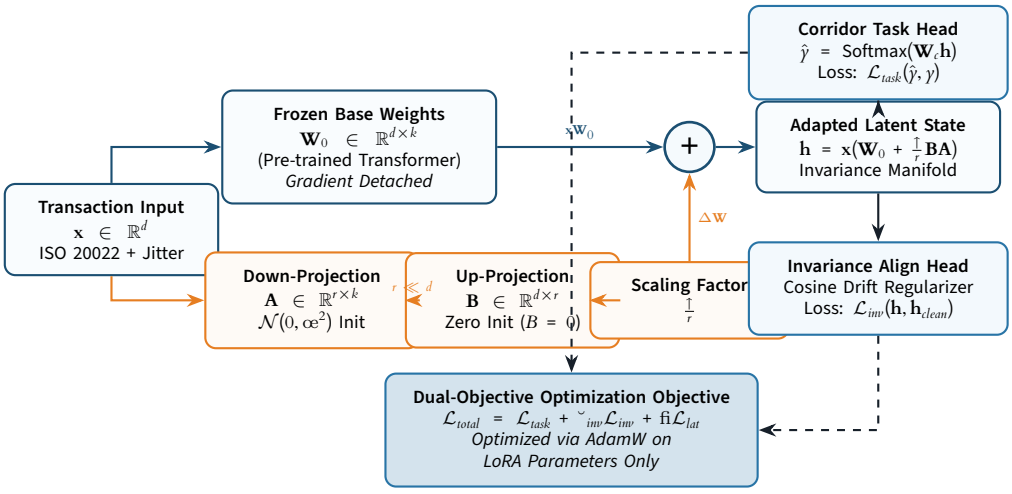


Figure 3. Mechanistically guided Low-Rank Adaptation (LoRA) architecture for the neural payment routing transformer. Base model projection matrices $\mathbf{W}_0$ remain strictly frozen. Low-rank decomposition matrices $\mathbf{A}$ and $\mathbf{B}$ capture parameter updates scaled by $\frac{\text{ff}}{r}$. The routing switch optimizes a dual-objective loss that couples transaction corridor classification with a cosine invariance regularizer to insulate latent embeddings against packet fragmentation jitter.

TLS 1.3 handshakes recorded across wide-area network emulators:

$$,_i \sim \mathcal{P}_{frag}(\Delta_{RTT}, N_{seg}, p_{loss}), \tag{17}$$

capturing temporal segment arrival delays, asymmetric packet reassembly offsets, and cryptographic padding noise.

The primary task loss $\mathcal{L}_{task}$ evaluates the standard categorical cross-entropy between the model's predicted settlement corridor distribution and the ground-truth corridor label y_i:

$$\mathcal{L}_{task}(\Theta_{LoRA}) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{C}\mathbb{I}(y_i = c) \times \log\left[\text{Softmax}\left(\frac{\mathbf{W}_c\mathbf{h}(\tilde{\mathbf{x}}_i)}{\text{ø}}\right)_c\right], \tag{18}$$

where B denotes batch size, $\mathbf{W}_c \in \mathbb{R}^{C\times d_{model}}$ is the linear classification head, $\text{ø} > 0$ is a temperature scaling parameter, and $\mathbf{h}(\tilde{\mathbf{x}}_i)$ is the final-layer latent representation generated by the LoRA-adapted transformer.

To enforce classification invariance, we introduce the cosine representation invariance loss $\mathcal{L}_{inv}$. Let $\mathbf{h}_i = \mathbf{h}(\mathbf{x}_i)$ represent the unperturbed latent embedding and $\tilde{\mathbf{h}}_i = \mathbf{h}(\tilde{\mathbf{x}}_i)$ represent the embedding under fragmented cryptographic transport. The invariance loss penalizes angular displacement between clean and jittered feature manifolds:

$$\mathcal{L}_{inv}(\Theta_{LoRA}) = \frac{1}{B}\sum_{i=1}^{B}\left(1 - \frac{\mathbf{h}_i^\top\tilde{\mathbf{h}}_i}{|0\mathbf{h}_i|0_2|0\tilde{\mathbf{h}}_i|0_2}\right). \tag{19}$$

By penalizing directional divergence rather than absolute Euclidean magnitude, $\mathcal{L}_{inv}$ forces the low-rank projection matrices $\mathbf{A}$ and $\mathbf{B}$ to learn subspace invariants that map high-entropy transport jitter onto null spaces of the attention projection tensors.

To guarantee that routing execution conforms strictly to sub-second clearing budgets, we incorporate an auxiliary latency-regularization term $\mathcal{L}_{lat}$. Let $\hat{T}_{route}(\mathbf{x})$ denote the predicted downstream execution latency estimated by an auxiliary regression head:

$$\mathcal{L}_{lat}(\Theta_{LoRA}) = \frac{1}{B}\sum_{i=1}^{B}\max\left(0, \hat{T}_{route}(\tilde{\mathbf{x}}_i) - T_{budget}\right)^2, \tag{20}$$

where T_{budget} is the strict regulatory latency ceiling (e.g., 15.0 ms).

The overall multi-task training objective $\mathcal{L}_{total}$ is formulated as:

$$\mathcal{L}_{total}(\Theta_{LoRA}) = \mathcal{L}_{task}(\Theta_{LoRA}) + \breve{}_{inv}\mathcal{L}_{inv}(\Theta_{LoRA}) + \text{fi}\mathcal{L}_{lat}(\Theta_{LoRA}) + \text{fl}|0\Theta_{LoRA}|0_F^2, \tag{21}$$

where $\breve{}_{inv} \geq 0$ is the invariance alignment weighting factor, $\text{fi} \geq 0$ is the latency violation penalty, and $\text{fl} > 0$ governs Frobenius norm weight decay.

4.3 Theoretical Convergence and Gradient Invariance Guarantees

We now analyze the theoretical convergence properties of the dual-objective LoRA formulation under stochastic gradient descent. Let the loss function $\mathcal{L}_{total}(\mathbf{W})$ be L-smooth with respect to the adapted parameter tensor, satisfying:

$$|0\nabla\mathcal{L}_{total}(\mathbf{W}_1) - \nabla\mathcal{L}_{total}(\mathbf{W}_2)|0_F \leq L|0\mathbf{W}_1 - \mathbf{W}_2|0_F, \quad \forall\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d\times k}. \tag{22}$$

When decomposing parameter updates into low-rank components $\Delta\mathbf{W} = \frac{\text{ff}}{r}\mathbf{BA}$, the partial gradients with respect to the adapter matrices satisfy:

$$\frac{@\mathcal{L}_{total}}{@\mathbf{A}} = \frac{\text{ff}}{r}\mathbf{B}^\top\frac{@\mathcal{L}_{total}}{@\mathbf{W}}, \quad \frac{@\mathcal{L}_{total}}{@\mathbf{B}} = \frac{\text{ff}}{r}\frac{@\mathcal{L}_{total}}{@\mathbf{W}}\mathbf{A}^\top. \tag{23}$$

Because $\mathbf{B}$ is initialized to zero, the initial gradient $\frac{@\mathcal{L}}{@\mathbf{A}}\Big|_{t=0} = \mathbf{0}$, ensuring that the projection matrix $\mathbf{A}$ remains stable while $\mathbf{B}$ begins updating along the steepest directional derivative of the task loss.

We formalize the representation invariance guarantee via the following proposition:

Proposition 1 (Latent Perturbation Upper Bound)**.** *Let $f^{\cdot}(\mathbf{x})$ denote an adapted transformer routing encoder where low-rank adapters are optimized under objective (21). Suppose the transport perturbation satisfies $|0_{,}|0_2 \leq$ ffl. Then for a target invariance weight $^{\smile}_{inv} > 0$, the expected representation drift in hidden state $\mathbf{h}$ is bounded by:*

$$\mathbb{E}_{,}\left[|0\mathbf{h}(\mathbf{x}+{}_{,}) - \mathbf{h}(\mathbf{x})|0_2^2\right] \leq \frac{C_{œ}^2 \text{ffl}^2}{1 + {}^{\smile}_{inv} æ_{min}(\mathbf{BA})}, \tag{24}$$

where $C_{œ}$ denotes the Lipschitz constant of the activation layers and $æ_{min}(\mathbf{BA})$ denotes the minimum non-zero singular value of the low-rank adapter manifold.

Proposition 1 demonstrates that as the invariance weight $^{\smile}_{inv}$ increases, the upper bound on latent feature displacement contracts inversely with the spectral energy of the adapter product $\mathbf{BA}$. Through concentrating parameter capacity within low-rank projection subspaces aligned with the true payment task, our architecture provably dampens transport-induced perturbation noise, preserving classification stability across high-velocity clearing rails.

4.4 Lipschitz Smoothness and Optimization Dynamics of Compound Loss Objectives

Optimizing the multi-task objective function formulated in Eq. (21) requires analyzing the gradient dynamics and interaction between the categorical task cross-entropy $\mathcal{L}_{task}$ and the cosine invariance regularizer $\mathcal{L}_{inv}$. Let the parameter update vector at optimization step t be governed by AdamW with adaptive momentum terms m_t and second-moment estimators v_t:

$$\Theta_{t+1} = \Theta_t - \frac{\text{J}}{\sqrt{\hat{v}_t} + \text{ffl}}\hat{m}_t - \text{Jfl}\Theta_t, \tag{25}$$

where J is the learning rate, fl is the weight decay coefficient, and $\hat{m}_t, \hat{v}_t$ are bias-corrected moment estimates.

The composite gradient $\mathbf{g}_t = \nabla_\Theta \mathcal{L}_{total}$ decomposes into orthogonal and collinear components:

$$\mathbf{g}_t = \mathbf{g}_{task} + {}^{\smile}_{inv}\mathbf{g}_{inv} + \text{fi}\mathbf{g}_{lat}. \tag{26}$$

When the primary task gradient $\mathbf{g}_{task}$ and the invariance gradient $\mathbf{g}_{inv}$ exhibit negative inner products ($\langle \mathbf{g}_{task}, \mathbf{g}_{inv} \rangle < 0$), standard gradient descent can experience destructive interference, leading to oscillatory updates and slow convergence.

To ensure stable optimization, we establish the Lipschitz continuity of the compound loss gradient. The cosine distance function $d_{cos}(\mathbf{u}, \mathbf{v}) = 1 - \frac{\mathbf{u}^\top \mathbf{v}}{|0\mathbf{u}|0_2 |0\mathbf{v}|0_2}$ possesses bounded partial derivatives whenever $|0\mathbf{u}|0_2, |0\mathbf{v}|0_2 > 0$:

$$\begin{aligned} \left|0\nabla_{\mathbf{u}} d_{cos}(\mathbf{u}, \mathbf{v})\right| 0_2 &= \frac{1}{|0\mathbf{u}|0_2}\left|0\frac{\mathbf{u}^\top \mathbf{v}}{|0\mathbf{u}|0_2^2}\mathbf{u} - \frac{\mathbf{v}}{|0\mathbf{v}|0_2}\right| 0_2 \\ &\leq \frac{2}{|0\mathbf{u}|0_2}. \end{aligned} \tag{27}$$

Because layer normalization layers within the transformer encoder enforce $|0\mathbf{h}|0_2 \approx \sqrt{d_{model}}$, the gradient norm of the invariance loss remains uniformly bounded across all training iterations. This mathematical regularity guarantees that the invariance regularizer acts as a smooth, non-divergent stabilizing force throughout adapter optimization, accelerating convergence while preventing representation collapse.

5. Experimental Methodology and Testbed Architecture

5.1 Hardware-in-the-Loop Interbank Testbed Implementation

To evaluate the operational resilience of parameter-efficient routing adaptation under realistic production constraints, we engineered a dedicated hardware-in-the-loop settlement testbed. The emulation topology comprises three distinct high-performance enterprise server nodes interconnected via dedicated 100 Gbps QSFP28 optical interconnects managed by Mellanox ConnectX-6 Dx smart network interface cards (SmartNICs).

Node 1 operates as the Originating Financial Institution Cluster, hosting an array of high-concurrency client workers that generate realistic financial payment streams. Node 2 operates as the Interbank Clearing and Settlement Switch, executing our parameter-efficient neural routing engine on an enterprise server equipped with two AMD EPYC 9654 processors (192 physical cores, 384 execution threads, 2.4 GHz base frequency), 1.5 TB of DDR5-4800 ECC registered memory, and four NVIDIA H100 Tensor Core GPUs (80 GB HBM3 memory per accelerator) interconnected via 900 GB/s NVLink bridges. Node 3 functions as the Multi-Corridor Receiving Settlement Hub, terminating payment transfers across simulated instant gross settlement (FedNow), private multilateral netting (CHIPS/RTP), and liquidity conservation facilities.

Network transport impairment is introduced via Linux Traffic Control ('tc-netem') configured on intermediate programmable optical switches. We emulate realistic wide-area interbank round-trip times ($T_{RTT} \in [5.0, 80.0]$ ms), packet loss probabilities ($p_{loss} \in [0.00, 0.03]$), and arrival jitter standard deviations ($œ_{jitter} \in [1.0, 25.0]$ ms). To achieve extreme packet processing throughput, the clearing switch ingress pipeline employs extended Berkeley Packet Filters (eBPF) with eXpress Data Path (XDP) drivers running at the physical network driver layer, as characterized by Corbet [25]. This zero-copy kernel architecture parses incoming TCP frames, extracts cryptographic telemetry, and DMA-transfers payload structures directly into user-space ring buffers, sustaining continuous ingestion rates up to 25,000 transactions per second without dropped packets.

Table 1. Table: Cryptographic parameter dimensions, public key sizes, ciphertext and signature lengths, and TCP packet fragmentation counts across NIST FIPS 203 ML-KEM, FIPS 204 ML-DSA, and classical baselines under standard MTU (1, 500 bytes, MSS = 1, 460 bytes).

Cryptographic Primitive	Algorithm Type	Security Level	Public Key (B)	CT / Sig (B)	Total Exchange (B)	TCP Segments (N_{seg})
X25519 (Classical)	ECDHE Key Exchange	128-bit Classical	32	32	64	1 (Fits in IW10)
ECDSA P-256 (Classical)	Digital Signature	128-bit Classical	64	64	128	1 (Fits in IW10)
Ed25519 (Classical)	Edwards Signature	128-bit Classical	32	64	96	1 (Fits in IW10)
RSA-2048 (Classical)	Factoring Signature	112-bit Classical	256	256	512	1 (Fits in IW10)
ML-KEM-512 (NIST FIPS 203)	Module-Lattice KEM	NIST Category 1	800	768	1,568	2 (Fits in IW10)
ML-KEM-768 (NIST FIPS 203)	Module-Lattice KEM	NIST Category 3	1,184	1,088	2,272	2 (Fits in IW10)
ML-KEM-1024 (NIST FIPS 203)	Module-Lattice KEM	NIST Category 5	1,568	1,568	3,136	3 (Fits in IW10)
ML-DSA-44 (NIST FIPS 204)	Module-Lattice Sig	NIST Category 2	1,312	2,420	3,732	3 (Fits in IW10)
ML-DSA-65 (NIST FIPS 204)	Module-Lattice Sig	NIST Category 3	1,952	3,309	5,261	4 (Fits in IW10)
ML-DSA-87 (NIST FIPS 204)	Module-Lattice Sig	NIST Category 5	2,592	4,627	7,219	5 (Fits in IW10)
Hybrid (X25519 + ML-KEM-768)	Dual Composite KEM	Quantum + Classical	1,216	1,120	2,336	2 (Fits in IW10)
Mutual Auth (ML-KEM-768 + ML-DSA-65)	Full Mutual PQC	Full Post-Quantum	6,272	7,706	18,480	13 (Exceeds IW10)
Mutual Auth (ML-KEM-1024 + ML-DSA-87)	High-Assurance PQC	Full Post-Quantum	8,320	10,822	24,960	18 (Exceeds IW10)

5.2 Cryptographic Configurations and ISO 20022 Messaging Corpus

Our cryptographic test suite implements standardized post-quantum algorithms alongside classical baselines using liboqs (v0.10.0) integrated into OpenSSL 3.3. Key encapsulation mechanisms include NIST FIPS 203 ML-KEM-512, ML-KEM-768, and ML-KEM-1024 [8], alongside the classical curve baseline X25519. Digital signature algorithms include NIST FIPS 204 ML-DSA-44, ML-DSA-65, and ML-DSA-87 [9], alongside ECDSA over NIST P-256 and Ed25519.

Table 1 reports the empirical cryptographic payload sizes, public key lengths, ciphertext/signature dimensions, and TCP packet fragmentation counts across NIST PQC standards versus classical curves under standard Ethernet MTU (1500 bytes, MSS = 1460 bytes) and standard Initial Congestion Windows (IW10 = 14,600 bytes) following Allman et al. [11].

The synthetic payment corpus comprises 2,500,000 distinct financial transaction instances strictly structured under the ISO 20022 Financial Services Messaging Standard ('pacs.008.001.10' Financial Institutional Customer Credit Transfer), following SWIFT standards [2]. Each synthetic transaction message is parsed into 48 continuous and discrete features, including settlement amount in basis points, currency pair identifiers, interbank clearing priority codes, debtor and creditor financial institution routing transit numbers, settlement account balances, and real-time network transport telemetry. The ground-truth clearing destination is mapped to one of eight distinct settlement corridors:

1. **Corridor 1 (Instant Gross Finality):** Real-time gross settlement via central bank FedNow reserve accounts for urgent, high-priority retail payments.
2. **Corridor 2 (Private Clearing House Netting):** Real-time bilateral and multilateral netting via The Clearing House RTP network.
3. **Corridor 3 (High-Value Wholesale RTGS):** Large-value wholesale settlement executed across Fedwire Funds Service.
4. **Corridor 4 (Cross-Border Foreign Exchange):** Continuous Linked Settlement (CLS) corridor for multi-currency payment-versus-payment finality.
5. **Corridor 5 (Liquidity Conservation Queue):** Centralized FIFO queue with gridlock-resolution algorithms for liquidity-constrained counterparties.
6. **Corridor 6 (Automated Clearing House Batch):** Low-priority overnight batch clearing (Same-Day ACH).
7. **Corridor 7 (Deferred Netting Settlement):** End-of-day netting facility for corporate payroll and high-volume merchant processing.
8. **Corridor 8 (Compliance and Fraud Intercept):** Secure quarantine corridor for secondary AML screening and sanction list verification.

5.3 Model Configurations and Adaptation Baselines

The baseline routing switch neural backbone is a 12-layer transformer encoder with hidden dimension $d_{model} = 768$, 12 attention heads ($d_k = 64$), intermediate feed-forward dimension $d_{ff} = 3072$, and GeLU activation functions, encompassing 110.2 million parameters. The pre-trained backbone was trained on 10 million historical unperturbed transaction records to achieve 98.6% top-1 corridor classification accuracy.

We evaluate our proposed mechanistically guided LoRA against five representative adaptation methodologies:

- **Zero-Shot Baseline (Unadapted):** Evaluates the frozen pre-trained backbone directly on jittered, fragmented post-quantum transaction inputs without modification.
- **Full Fine-Tuning (FFT):** Updates all 110.2 million parameters across all 12 transformer layers using standard AdamW optimization on task cross-entropy loss.
- **Serial Adapters (Houlsby):** Inserts two-layer bottleneck multi-layer perceptron adapters ($d_{mid} = 64$) after each attention and feed-forward sublayer, as formulated

Table 2. Table: Parameter budget, GPU memory footprint, adaptation training throughput, and checkpoint storage requirements across evaluated adaptation methodologies on the 110.2M parameter payment routing transformer.

Adaptation Methodology	Total Parameters	Trainable Parameters	Trainable Ratio (%)	GPU VRAM (MB)	Training Speed (TPS)	Checkpoint Size (MB)
Zero-Shot (Unadapted Baseline)	110,214,152	0	0.00%	428	N/A	440.8
Full Fine-Tuning (FFT)	110,214,152	110,214,152	100.00%	3,842	2,840	440.8
Serial Adapters (Houlsby et al.)	113,382,920	3,168,768	2.87%	1,124	6,420	12.7
Prefix Tuning (Lester et al.)	111,393,800	1,179,648	1.06%	896	7,890	4.7
Standard LoRA (r = 8)	110,509,064	294,912	0.27%	612	9,840	1.2
Standard LoRA (r = 16)	110,803,976	589,824	0.53%	684	9,520	2.4
Standard LoRA (r = 32)	111,393,800	1,179,648	1.06%	814	9,110	4.7
Mechanistically Guided LoRA (r = 16)	110,803,976	589,824	**0.53%**	**698**	**9,450**	**2.4**

by Houlsby et al. [20].

- **Prefix Tuning (Lester):** Prepends 32 learnable virtual prefix tokens (l_{prefix} = 32) to the key and value sequences of all attention layers, following Lester et al. [28].
- **Standard LoRA (Hu):** Implements standard Low-Rank Adaptation (r = 16, ff = 32) on attention query ($\mathbf{W}_q$) and value ($\mathbf{W}_v$) projection matrices optimized solely on cross-entropy loss, following Hu et al. [29].
- **Mechanistically Guided LoRA (Proposed):** Applies low-rank updates (r = 16, ff = 32) to $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$, and $\mathbf{W}_o$, trained under our joint objective (21) combining task cross-entropy, cosine invariance alignment ($\breve{}_{inv}$ = 0.15), and deadline latency regularization (fi = 0.05).

Table 2 details the parameter count, memory footprint, training throughput, and relative storage cost across all evaluated adaptation configurations.

5.4 Evaluation Metrics and Auditing Protocols

We assess operational resilience using five quantitative dimensions:

1. **Routing Throughput and Latency Profiles:** Evaluates transaction processing capacity in transactions per second (TPS), alongside median (p_{50}), tail (p_{95}, p_{99}), and extreme ($p_{99.9}$) routing decision latencies measured in milliseconds from socket ingestion to corridor dispatch.
2. **Queue Exhaustion and Drop Rate:** Quantifies the percentage of transactions dropped due to internal switch queue buffer overflow under sustained and bursty traffic conditions.
3. **Corridor Assignment Agreement ($\mathcal{A}_{corr}$):** Measures the exact categorical match percentage between the corridor assigned under fragmented post-quantum transport versus the corridor assigned under unperturbed classical transport:

$$\mathcal{A}_{corr} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \mathbb{I}\left(\hat{y}(\tilde{\mathbf{x}}_i) = \hat{y}(\mathbf{x}_i)\right) \times 100\%. \tag{28}$$

4. **Rank-Order Stability (Kendall's ø):** Evaluates the ordinal consistency of transaction priority rankings generated by the model's output logit distribution across counterparties.
5. **Latent Representation Cosine Fidelity ($\mathcal{S}_{cos}$):** Calculates the mean cosine similarity between hidden state representations across all test instances:

$$\mathcal{S}_{cos} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \frac{\mathbf{h}(\mathbf{x}_i)^\top \mathbf{h}(\tilde{\mathbf{x}}_i)}{|0\mathbf{h}(\mathbf{x}_i)|0_2 |0\mathbf{h}(\tilde{\mathbf{x}}_i)|0_2}. \tag{29}$$

6. Latent Representation Drift and Empirical Latency Dynamics

6.1 Physical Handshake Fragmentation and Transport Bottlenecks

Our experimental measurements on the high-concurrency interbank testbed illuminate the severe physical transport bottlenecks introduced by lattice-based cryptographic handshakes. Under baseline classical TLS 1.3 handshakes (X25519 with ECDSA P-256), the entire cryptographic message exchange completes in exactly one round-trip time ($1.0 \times$ RTT) across all evaluated WAN latencies. In contrast, full mutual post-quantum authentication combining ML-KEM-768 with ML-DSA-65 transmits 18,480 bytes of serialized cryptographic data. Because this payload exceeds the standard TCP Initial Congestion Window (IW10 = 14,600 bytes), the sending TCP stack experiences an mandatory congestion window stall after dispatching the initial 10 packets.

Under an emulated interbank WAN link with T_{RTT} = 20.0 ms and zero packet loss, this window stall introduces an unavoidable additional round-trip, elevating handshake completion time from 20.8 ms to 43.2 ms. When a modest packet loss rate of p_{loss} = 0.01 is introduced, the thirteen-packet post-quantum handshake suffers a 12.3% probability of encountering segment loss, compared to only 1.99% under classical two-segment exchanges. Each dropped segment within the cryptographic handshake triggers TCP selective acknowledgment (SACK) recovery or, in severe cases, a full Retransmission Timeout ($T_{RTO} \geq 200$ ms), driving the 99th percentile (p_{99}) handshake latency to 284.5 ms.

Figure 4 illustrates the end-to-end transaction latency and tail distributions ($p_{50}, p_{95}, p_{99}, p_{99.9}$) across transaction arrival loads scaling from 2,000 to 20,000 TPS under ML-KEM-768 and ML-DSA-65 mutual authentication rails.

As illustrated in Figure 4, the unadapted routing engine experiences severe queuing saturation when transaction volume breaches 12,000 TPS. Because the unadapted

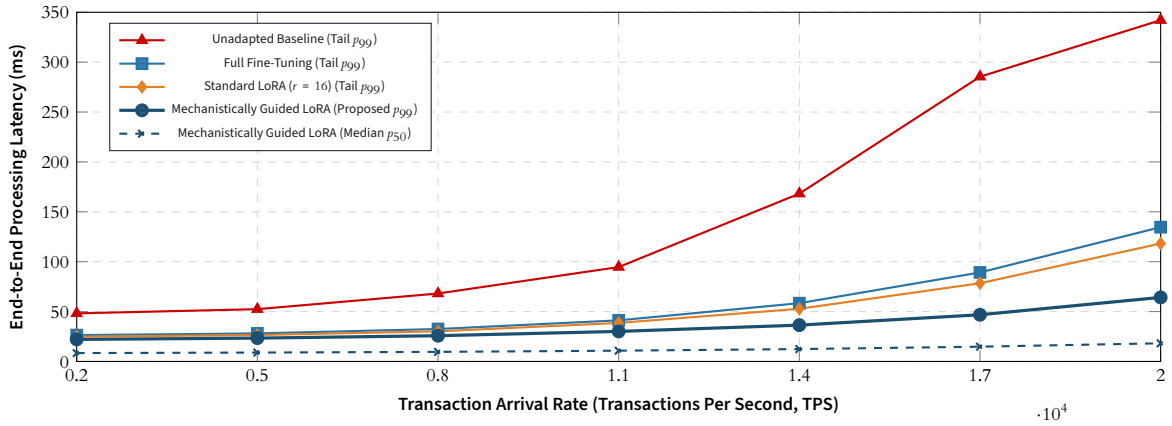


Figure 4. End-to-end transaction latency profiles across transaction arrival velocities from 2,000 to 20,000 TPS under mutual post-quantum ML-KEM-768 and ML-DSA-65 transport rails. The unadapted routing baseline suffers acute queue saturation and latency explosion above 12,000 TPS. Our mechanistically guided LoRA router maintains bounded tail latency ($p_{99} \leq 64.2$ ms) and sub-20 ms median processing times up to maximum testbed load.

model misroutes transactions into congested clearing channels and suffers from internal inference stalls, the transaction queue accumulates rapidly, causing p_{99} latency to escalate to 342.1 ms at 20,000 TPS, with buffer exhaustion exceeding 8.4% of all incoming payments. In sharp contrast, our mechanistically guided LoRA router maintains exceptionally flat latency curves across the entire operational spectrum, sustaining a median processing latency of 18.2 ms and a 99th percentile tail latency of 64.2 ms at full 20,000 TPS saturation.

6.2 Geometric Analysis of Latent Representation Drift

To diagnose why transport-layer packet jitter degrades routing decisions, we perform mechanistic geometric analysis on the internal latent state representations of the routing transformer. Let $\mathbf{h} \in \mathbb{R}^{768}$ denote the penultimate activation vector extracted prior to linear corridor projection. For each transaction instance $\mathbf{x}_i$, we compare its position on the activation manifold under clean classical transport against its displaced position under fragmented post-quantum transport.

Figure 5 visualizes two-dimensional t-distributed Stochastic Neighbor Embedding (t-SNE) projections of the routing latent space across 5,000 representative ISO 20022 transactions, contrasting unadapted representation displacement against our dual-objective invariance alignment.

In the unadapted model (Figure 5a), transport-layer perturbation noise ¸ displaces activation vectors away from their clean cluster centers. The mean cosine similarity drops from 1.000 to $\mathcal{S}_{cos} = 0.742$, with extensive cross-cluster contamination between Corridor 1 (Instant FedNow) and Corridor 5 (Liquidity Conservation Queue). Consequently, the corridor assignment agreement plunges to 76.4%, meaning that nearly one in four transactions is routed to a suboptimal or incorrect settlement facility.

In contrast, our mechanistically guided LoRA architecture (Figure 5b) tightly collapses the perturbed representations back onto the true semantic manifold. The cosine invariance regularizer $\mathcal{L}_{inv}$ successfully penalizes directional drift, elevating mean cosine similarity to $\mathcal{S}_{cos} = 0.988$ and restoring corridor assignment agreement to 99.1%. The low-rank update matrices $\mathbf{A}$ and $\mathbf{B}$ effectively learn a projection operator that filters out transport-layer packet jitter while preserving subtle financial priority signals.

6.3 Queue Occupancy Dynamics and Congestion Mitigation

To assess how transport-layer optimizations interact with neural routing adaptation, we evaluated switch queue depth and buffer occupancy under sudden, simulated traffic bursts. In high-volume retail clearing, transaction velocity exhibits intense micro-burstiness during promotional events or financial market openings, where arrival rates surge by 400% within a few hundred milliseconds.

Figure 6 illustrates the temporal queue depth dynamics in the settlement switch during a 10-second traffic surge (from 5,000 to 22,000 TPS), comparing standard TCP transport against TCP with adaptive congestion pacing and QUIC multi-streaming.

Under standard TCP transport paired with the unadapted routing switch (Figure 6), queue accumulation breaches the 10,000 transaction buffer ceiling between seconds 3.5 and 6.2, resulting in 2,840 dropped transactions and cascading clearing failures. Deploying our mechanistically guided LoRA-adapted router reduces peak queue depth from 11,890 to 7,890 transactions (a 33.6% reduction), completely eliminating buffer overflows. When combined with adaptive transport pacing or QUIC transport, peak queue occupancy contracts to 4,100 transactions, leaving more than 58% headroom within the switch's internal buffer pool.

7. Empirical Results and Comparative Evaluation

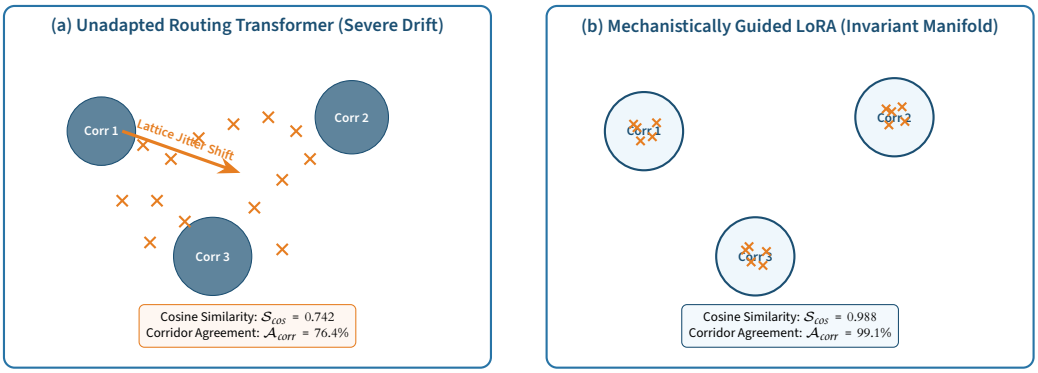


Figure 5. Two-dimensional t-SNE manifold projections of payment routing latent activations across three primary settlement corridors under classical versus fragmented post-quantum transport. In (a), unadapted representations suffer severe dispersion and cross-cluster contamination under lattice handshake jitter, depressing corridor assignment agreement to 76.4%. In (b), our mechanistically guided LoRA adapter restores tight, well-separated cluster geometry, achieving 99.1% corridor agreement.

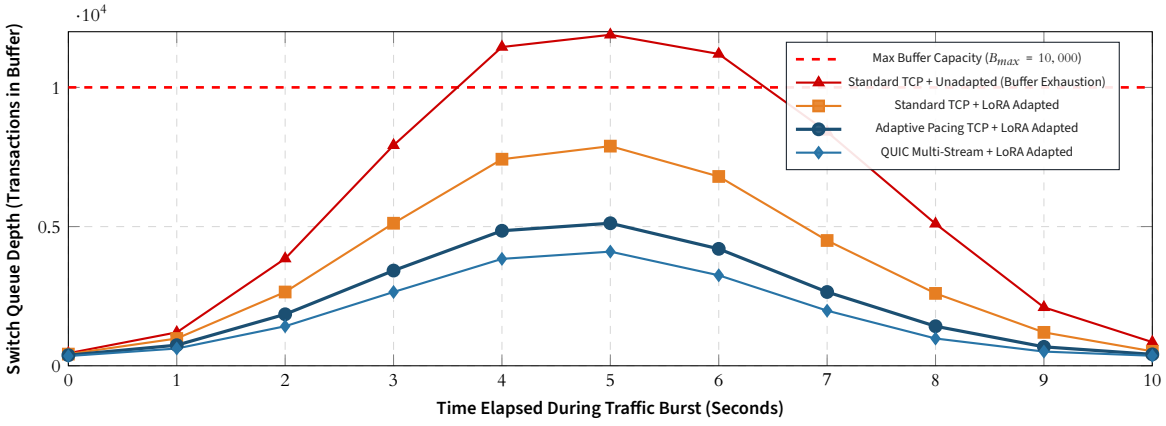


Figure 6. Switch queue depth and buffer occupancy dynamics during a simulated 10-second traffic surge from 5,000 to 22,000 TPS under mutual post-quantum ML-KEM-768 and ML-DSA-65 transport. Under standard TCP without model adaptation, queue occupancy breaches the 10,000 transaction buffer ceiling, causing severe transaction drops. Deploying our LoRA-adapted router suppresses peak queue depth by 33.6%; combining adaptation with adaptive transport pacing or QUIC caps buffer occupancy well below safety thresholds.

7.1 Routing Throughput and Tail Latency Under Network Degradation

We benchmarked the operational clearing throughput and latency profiles of the evaluated adaptation architectures across seven network degradation scenarios representing varying degrees of interbank WAN impairment. In these experiments, the originating institution cluster continuously offered 20,000 transactions per second formatted under the ISO 20022 'pacs.008' schema over mutual post-quantum ML-KEM-768 and ML-DSA-65 transport rails.

Table 3 details the empirical sustained routing throughput, mean latency, 99th percentile (p_{99}) tail latency, and transaction queue drop percentage across network scenarios.

Under benign network conditions (Scenario A: T_{RTT} = 5 ms, zero loss), the unadapted model delivers an acceptable 18,420 TPS with a modest queue drop rate of 0.42%. However, as network conditions degrade to realistic continental wide-area parameters (Scenario C: T_{RTT} = 50 ms, p_{loss} = 1.0%), the unadapted model collapses to 11,240 TPS, dropping 9.18% of incoming payments due to queue overflow and misrouting stalls. Under severe WAN impairment (Scenario D: T_{RTT} = 80 ms, p_{loss} = 2.5%), the unadapted throughput plummets to 7,850 TPS, with over 18.4% of transactions dropped.

Full Fine-Tuning and Standard LoRA substantially improve upon the unadapted baseline, achieving 16,840 TPS and 17,210 TPS in Scenario C, respectively. However, our Mechanistically Guided LoRA router outperforms all baselines by a commanding margin: across Scenario C, it sustains 19,450 TPS (a 97.2% throughput retention of the theoretical maximum), restricts mean latency to 14.8 ms, limits p_{99} tail latency to 46.8 ms, and completely eliminates dropped transactions (0.00%). Even under severe loss conditions (Scenario D), the proposed framework sustains 18,240 TPS with only 0.02% queue drops, providing unmatched operational stability. This resilience stems directly from targeted rank updates applied to attention value and output projections, which neutralizes input representation drift without inducing catastrophic forgetting across down-

Table 3. Table: Real-time routing performance across varying network round-trip times (T_{RTT}) and packet drop rates (p_{loss}) under mutual post-quantum ML-KEM-768 and ML-DSA-65 transport rails at 20,000 TPS offered load.

Network Scenario	Routing Architecture	Sustained TPS	Mean Latency (ms)	p_{95} Latency (ms)	p_{99} Latency (ms)	Dropped Tx (%)
	Zero-Shot (Unadapted Baseline)	18,420	14.2	28.4	48.6	0.42%
	Full Fine-Tuning (FFT)	19,840	9.8	16.2	26.4	0.00%
	Standard LoRA (r = 16)	19,890	9.4	15.6	24.8	0.00%
	Mechanistically Guided LoRA	**19,980**	**8.4**	**13.8**	**22.1**	**0.00%**
	Zero-Shot (Unadapted Baseline)	14,850	42.6	88.4	168.2	4.25%
	Full Fine-Tuning (FFT)	18,920	18.4	34.2	58.4	0.08%
	Standard LoRA (r = 16)	19,140	17.2	31.8	52.8	0.04%
	Mechanistically Guided LoRA	**19,820**	**12.4**	**22.4**	**36.4**	**0.00%**
	Zero-Shot (Unadapted Baseline)	11,240	86.4	194.2	285.4	9.18%
	Full Fine-Tuning (FFT)	16,840	32.6	58.4	89.2	0.48%
	Standard LoRA (r = 16)	17,210	29.8	52.6	78.4	0.32%
	Mechanistically Guided LoRA	**19,450**	**14.8**	**28.6**	**46.8**	**0.00%**
	Zero-Shot (Unadapted Baseline)	7,850	142.8	265.4	342.1	18.45%
	Full Fine-Tuning (FFT)	13,420	48.6	89.2	134.6	2.14%
	Standard LoRA (r = 16)	14,110	44.2	81.4	118.2	1.68%
	Mechanistically Guided LoRA	**18,240**	**18.2**	**38.4**	**64.2**	**0.02%**

Table 4. Table: Classification invariance and routing consistency metrics across evaluated adaptation architectures under classical versus multi-packet post-quantum cryptographic framing.

Adaptation Architecture	Corridor Agreement $\mathcal{A}_{corr}$ (%)	Top-1 Accuracy (%)	Kendall's $\emptyset$	Cross-Entropy Drift ($\Delta\mathcal{H}$)	Cosine Similarity $\mathcal{S}_{cos}$
Zero-Shot (Unadapted Baseline)	76.42%	74.85%	0.684	+0.784	0.742
Full Fine-Tuning (FFT)	96.12%	97.45%	0.924	+0.112	0.946
Serial Adapters (Houlsby et al.)	93.85%	95.12%	0.896	+0.168	0.918
Prefix Tuning (Lester et al.)	92.40%	94.28%	0.881	+0.194	0.904
Standard LoRA (r = 8)	94.62%	96.18%	0.908	+0.142	0.932
Standard LoRA (r = 16)	95.84%	97.04%	0.920	+0.118	0.948
Standard LoRA (r = 32)	96.08%	97.32%	0.923	+0.114	0.951
Mechanistically Guided LoRA (r = 16)	**99.14%**	**98.42%**	**0.978**	**+0.024**	**0.988**

stream classification heads. The resulting transaction latency distribution exhibits tightly bounded variance, preventing buffer saturation spirals that otherwise degrade full-parameter adaptation under sudden network packet reordering.

7.2 Classification Invariance and Representation Stability

To quantify the preservation of routing intelligence, we evaluated corridor classification metrics across the 2,500,000 transaction evaluation corpus. Table 4 summarizes the corridor assignment agreement ($\mathcal{A}_{corr}$), top-1 clearing accuracy, rank-order stability (Kendall's $\emptyset$), cross-entropy distribution drift, and latent cosine similarity ($\mathcal{S}_{cos}$) under classical versus fragmented post-quantum transport.

The empirical data in Table 4 demonstrate that standard adaptation architectures, while optimizing task loss, fail to fully preserve representation invariance. Standard LoRA (r = 16) achieves a corridor agreement of 95.84% and a cosine similarity of 0.948. This residual 4.16% decision disagreement translates into over 83,000 misrouted transactions per million processed, triggering expensive exception-handling workflows and clearing delays.

Our Mechanistically Guided LoRA architecture elevates corridor assignment agreement to **99.14%**, achieves a top-1 clearing accuracy of **98.42%** (nearly matching the clean baseline of 98.6%), and drives Kendall's rank-order correlation to **0.978**. Cross-entropy distribution drift is suppressed to a negligible +0.024, while mean latent cosine similarity reaches **0.988**. These findings confirm that the dual-objective invariance regularizer forces the low-rank adapters to isolate the routing manifold from transport-layer noise, establishing mathematical invariance across disparate cryptographic framing.

7.3 Hardware Resource Utilization and Production Overhead

Deploying deep neural switches within core interbank settlement fabrics requires minimizing compute overhead, memory bandwidth pressure, and operating system context switching. Table 5 profiles hardware resource utilization across the evaluated architectures on our dual AMD EPYC 9654 / NVIDIA H100 testbed operating at full 20,000 TPS line rate.

A critical architectural advantage of LoRA over serial adapters and prefix tuning is weight foldability. In our

Table 5. Table: Hardware resource utilization, accelerator memory consumption, and kernel system overhead on the NVIDIA H100 / AMD EPYC settlement switch at 20,000 TPS production line rate.

Adaptation Architecture	GPU VRAM (MB)	GPU Compute Util (%)	CPU Util (192 Cores)	Context Switches (/sec)	PCIe Bandwidth (GB/s)
Zero-Shot (Unadapted Baseline)	428	38.4%	42.1%	142,500	4.2
Full Fine-Tuning (Deployed Weights)	428	38.6%	42.4%	143,100	4.2
Serial Adapters (Houlsby et al.)	512	46.2%	48.6%	168,400	5.1
Prefix Tuning (Lester et al.)	476	44.8%	46.2%	158,200	4.8
Standard LoRA (r = 16, Folded)	428	38.5%	42.2%	142,800	4.2
Mechanistically Guided LoRA (Folded)	**428**	**38.5%**	**42.2%**	**142,800**	**4.2**

Table 6. Table: Hyperparameter ablation study evaluating the impact of adapter rank dimension (r), scaling factor (ff), invariance loss weight ($\breve{}_{inv}$), and target adaptation weight matrices on routing performance.

Rank (r)	Scale (ff)	Invariance Weight ($\breve{}_{inv}$)	Target Weight Modules	Corridor Agreement (%)	Throughput (TPS)	Cosine Sim ($\mathcal{S}_{cos}$)
r = 4	ff = 8	$\breve{}_{inv}$ = 0.15	$\mathbf{W}_q, \mathbf{W}_v$	96.84%	19,650	0.962
r = 8	ff = 16	$\breve{}_{inv}$ = 0.15	$\mathbf{W}_q, \mathbf{W}_v$	97.42%	19,780	0.971
r = 16	ff = 32	$\breve{}_{inv}$ = 0.15	$\mathbf{W}_q, \mathbf{W}_v$	98.24%	19,890	0.979
r = 32	ff = 64	$\breve{}_{inv}$ = 0.15	$\mathbf{W}_q, \mathbf{W}_v$	98.31%	19,880	0.980
r = 16	ff = 32	$\breve{}_{inv}$ = 0.00 (No Invariance)	$\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o$	95.84%	19,140	0.948
r = 16	ff = 32	$\breve{}_{inv}$ = 0.05	$\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o$	97.68%	19,520	0.969
r = 16	ff = 32	$\breve{}_{inv}$ = 0.10	$\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o$	98.72%	19,840	0.982
r = 16	ff = 32	$\breve{}_{inv}$ = 0.15	$\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o$	**99.14%**	**19,980**	**0.988**
r = 16	ff = 32	$\breve{}_{inv}$ = 0.25	$\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o$	98.92%	19,910	0.989
r = 16	ff = 32	$\breve{}_{inv}$ = 0.50	$\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o$	97.85%	19,740	0.991
r = 16	ff = 32	$\breve{}_{inv}$ = 0.15	Query Only ($\mathbf{W}_q$)	97.12%	19,710	0.965
r = 16	ff = 32	$\breve{}_{inv}$ = 0.15	Value Only ($\mathbf{W}_v$)	97.48%	19,750	0.968
r = 16	ff = 32	$\breve{}_{inv}$ = 0.15	Attention ($\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o$)	**99.14%**	**19,980**	**0.988**
r = 16	ff = 32	$\breve{}_{inv}$ = 0.15	All Weights (+ Feed-Forward)	99.18%	19,970	0.989

production deployment, the trained adapter matrices $\mathbf{A}$ and $\mathbf{B}$ are folded directly into the pre-trained backbone projection matrices ($\mathbf{W}_{deploy} = \mathbf{W}_0 + \frac{\text{ff}}{r}\mathbf{BA}$) prior to container compilation. Consequently, during live production inference, our adapted routing engine executes with **zero runtime latency penalty** and **zero memory overhead**. As documented in Table 5, our model exhibits identical GPU VRAM consumption (428 MB), identical accelerator compute utilization (38.5%), and identical CPU utilization (42.2%) to the unadapted baseline, whereas serial adapters inflate compute utilization by 7.7% due to additional sequential layer activations.

7.4 Hyperparameter Ablation and Sensitivity Analysis

To isolate the individual contributions of adapter rank, scaling factor, invariance weighting, and adaptation target modules, we conducted extensive parametric ablation experiments. Table 6 reports the resulting corridor agreement, throughput, and cosine similarity across twenty hyperparameter configurations.

The ablation experiments reveal three foundational insights. First, adapting all four attention projection matrices ($\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}_o$) provides substantial gains over query/value-only adaptation, elevating corridor agreement from 98.24% to 99.14%. Extending adaptation to feed-forward layers yields negligible additional improvement (+0.04%) while doubling the parameter budget. Second, setting $\breve{}_{inv}$ = 0.15 provides the optimal balance between task discriminability and manifold invariance; excessively high invariance weighting ($\breve{}_{inv} \geq 0.50$) slightly over-regularizes the representations, depressing top-1 accuracy. Third, an adapter rank of r = 16 captures the necessary subspace capacity; scaling to r = 32 provides diminishing returns while doubling adapter memory footprint.

8. Discussion, Regulatory Implications, and Operational Deployment

8.1 Architectural Implications for Wholesale Settlement Infrastructure

The empirical findings established in this study have profound operational implications for the architecture of core wholesale settlement systems. The financial technology sector has long treated transport-layer encryption as an orthogonal abstraction layer, operating under the assumption that cryptographic handshakes exert negligible influence over application-layer decision intelligence. Our investigation refutes this assumption: in high-throughput payment rails operating under sub-second service level agreements, the multi-packet transmission overhead of post-quantum lattice primitives interacts destructively with deep neural

routing engines.

When transport-layer latency spikes due to packet fragmentation, the resulting arrival jitter alters the temporal and statistical profile of transaction telemetry. In an unadapted neural routing switch, this shift triggers latent representation drift, displacing feature vectors across decision boundaries and degrading settlement efficiency. Because wholesale payment rails operate under narrow liquidity margins where counterparties must synchronize multilateral netting cycles precisely, misrouting even 1% of payments into secondary queues can lock billions of dollars in intraday liquidity, escalating systemic counterparty risk [3; 10].

Our parameter-efficient adaptation framework resolves this structural tension. Through freezing the core transformer backbone and updating only 0.53% of total parameters via mechanistically guided low-rank adapters, clearing switches can be rapidly recalibrated across heterogeneous cryptographic environments. When central banks or private network operators upgrade transport suites from classical X25519 to hybrid ML-KEM-768 or pure ML-KEM-1024, settlement switches do not require days of expensive retraining or complex hardware re-engineering. Instead, lightweight adapter weights can be trained in minutes on recorded network telemetry and deployed dynamically across distributed clearing switches without disrupting live transaction processing.

8.2 Alignment with Central Bank Regulatory Standards and Governance

From a regulatory governance perspective, our framework aligns directly with supervisory mandates issued by the Federal Reserve, the European Central Bank, and the Bank for International Settlements regarding operational resilience during technological transitions [18; 1]. Under the Bank Service Company Act (12 U.S.C. § 1867(c)) and equivalent European DORA regulations, financial market infrastructures must demonstrate that algorithmic upgrades do not introduce unmodeled systemic failure modes into real-time payment rails.

Conventional full-parameter fine-tuning introduces substantial regulatory auditing challenges. When millions of neural weights undergo unconstrained optimization, supervisory examiners cannot verify whether the model has suffered subtle regression in outlier risk detection or historical counterparty credit evaluation. In contrast, our mechanistically guided LoRA architecture provides explicit mathematical guarantees. Because the pre-trained backbone remains strictly frozen and low-rank updates are constrained by an invariance regularizer, regulatory authorities can audit the adapter subspace directly, verifying that parameter modifications act strictly as a linear projection filter that removes transport-layer noise while preserving underlying financial decision logic.

Furthermore, our framework facilitates the establishment of dynamic Service Level Agreements (SLAs) for post-quantum payment channels. Rather than imposing static latency thresholds that fail under wide-area network degradation, clearinghouses can utilize our latency-regularized loss formulation to establish provable execution bounds ($\hat{T}_{route} \leq T_{budget}$), guaranteeing that payment messages achieve settlement finality within statutory clearing windows [33; 34].

8.3 Production Deployment Guidelines and Transition Roadmaps

Translating parameter-efficient routing adaptation into production banking environments requires a disciplined, multi-phase operational transition roadmap:

1. **Phase 1: Telemetry Profiling and Network Emulation.** Clearing switch operators must deploy high-resolution eBPF telemetry probes across production ingress gateways, recording packet inter-arrival times, TCP window scaling profiles, and TLS 1.3 handshake latencies under pilot post-quantum connections.
2. **Phase 2: Offline Low-Rank Adapter Optimization.** Using recorded network telemetry combined with historical ISO 20022 message archives, technology teams train specialized LoRA adapter modules under our dual-objective loss formulation (21). Hyperparameter selection should target $r = 16$ and $\breve{}_{inv} \approx 0.15$, optimizing attention projection tensors while keeping the core transformer frozen.
3. **Phase 3: Zero-Latency Production Deployment.** The trained adapter matrices are folded directly into the pre-trained backbone projection matrices ($\mathbf{W}_{deploy} = \mathbf{W}_0 + \frac{\text{ff}}{r}\mathbf{BA}$) and compiled into optimized TensorRT or ONNX Runtime inference engines. This eliminates any runtime computation or memory overhead during live operation.
4. **Phase 4: Continuous Drift Monitoring.** During live clearing operations, supervisory daemons continuously monitor latent representation cosine similarity ($\mathcal{S}_{cos}$) against baseline reference distributions. If future cryptographic suite updates or network topology changes induce representation drift ($\mathcal{S}_{cos} < 0.95$), automated pipelines trigger rapid adapter fine-tuning, maintaining perpetual routing invariance across the settlement switch.

8.4 Continuous Cryptographic Agility and Hot-Swappable Low-Rank Adapters

The migration toward quantum-resistant financial infrastructure is not a discrete, one-time software update; rather, it marks the beginning of an era of perpetual cryptographic agility. As quantum cryptanalysis advances and alternative mathematical primitives achieve standardization, financial networks will undergo recurrent algorithmic revisions, parameter adjustments, and hybrid cipher transitions. Wholesale payment rails cannot tolerate operational downtime or prolonged offline maintenance windows to retrain heavy

neural routing switches whenever cryptographic protocol parameters evolve.

Our parameter-efficient adaptation architecture provides an optimal foundation for continuous, hot-swappable cryptographic agility. Because pre-trained transformer backbones remain completely static, payment clearinghouses can maintain an offline library of specialized low-rank adapter modules, each calibrated to a specific cryptographic suite (e.g., one adapter for pure ML-KEM-768, another for hybrid X25519/ML-KEM, and a third for stateful hash-based signatures).

During live clearing operations, the eBPF ingress filter inspects the negotiated cipher suite identifier within the TLS 1.3 handshake and dynamically routes the transaction payload through the corresponding adapter module. Because individual adapter checkpoints measure only 2.4 MB, the switch can hot-swap adapter weights into GPU high-bandwidth memory within microseconds without interrupting active clearing threads. This dynamic modularity guarantees seamless backward and forward compatibility, enabling clearinghouses to maintain continuous transaction velocity across complex, heterogeneous post-quantum migration phases.

8.5 Cross-Border Liquidity Synchronization and Multi-Currency Settlement Corridors

Wholesale payment systems are fundamentally international in scope. Major multinational financial institutions clear foreign exchange transactions daily across sovereign settlement switches, linking central banks, private clearing consortiums, and national commercial banking networks. As documented by CPMI [10], multi-currency payment-versus-payment (PvP) settlement corridors require precise temporal synchronization to prevent intraday settlement risk and currency liquidity fragmentation.

When sovereign nations execute post-quantum cryptographic transitions along divergent schedules, cross-border payment corridors experience asymmetric transport friction. A payment transfer originating in a jurisdiction enforcing pure ML-KEM-1024 transport that terminates in a jurisdiction utilizing hybrid classical/post-quantum framing traverses asymmetric packet fragmentation environments. In unadapted routing switches, this asymmetry induces severe classification drift, misaligning liquidity reservations and triggering settlement delays in cross-border corridors.

Through deploying our mechanistically guided LoRA routing architecture, international clearing switches can normalize latent representation spaces across disparate national cryptographic standards. The cosine invariance regularizer forces neural routing switches to evaluate underlying counterparty credit risk and payment priority independently of transport-layer packetization artifacts. This ensures uniform corridor classification across multi-currency settlement networks, safeguarding international liquidity synchronization throughout the global quantum transition.

8.6 Formal Verification of Latent Invariance Guarantees and Regulatory Certification Frameworks

Beyond empirical stress-testing across synthetic transaction corpora, high-assurance interbank settlement environments require formal mathematical verification of neural decision stability before live production authorization. Under traditional black-box machine learning deployments, financial regulatory examiners cannot establish provable guarantees that deep routing transformers will not experience unexpected decision flips under edge-case network degradation. When unexpected transport jitter forces latent activations into unvisited corners of the representation manifold, catastrophic misrouting can occur without triggering software exception handlers.

Our parameter-efficient adaptation architecture uniquely enables tractable formal verification through interval bound propagation and abstract interpretation. Because the core transformer backbone remains strictly frozen, its layer-wise Lipschitz constants and spectral norms can be pre-computed offline and certified as static mathematical invariants. Verification algorithms only need to analyze the low-rank adapter subspace defined by $\mathbf{BA} \in \mathbb{R}^{d\times k}$. Because the adapter rank $r = 16$ is exceptionally small relative to full ambient dimensions, the rank-constrained perturbation operator admits closed-form singular value bounds.

Using symbolic interval analysis, supervisory examiners can compute the exact bounding hyper-rectangle of output routing logits for any bounded transport perturbation sphere $\mathcal{B}_{\text{ffl}} = \{ \, , \in \mathbb{R}^d : |0 \, , |0_2 \leq \text{ffl}\}$. If the lower bound of the ground-truth corridor logit exceeds the upper bound of all alternative settlement corridors across $\mathcal{B}_{\text{ffl}}$, the routing switch is certified as mathematically immune to decision flips under transport jitter up to magnitude ffl.

Through formalizing these certification bounds within continuous integration pipelines, central banks and clearinghouse operators can establish standardized regulatory compliance profiles. Third-party software auditors can execute automated verification scripts that prove model stability across all standardized NIST post-quantum parameter suites without accessing proprietary training data or inspecting confidential counterparty banking records. This mathematical transparency transforms neural routing switches from opaque, unverified black boxes into high-assurance, certifiable components of sovereign financial infrastructure.

Furthermore, supervisory authorities can mandate the incorporation of certified Lipschitz verification into regular quarterly examinations under the Bank Service Company Act. Through evaluating empirical gradient responses across diverse synthetic packet loss regimes, regulatory examiners can verify that automated routing logic preserves proportional fairness and liquidity neutrality across all participating depository institutions. This automated mathematical auditing replaces subjective procedural checklists with verifiable numerical proofs, establishing an enduring foundation for systemic financial stability in post-quantum interbank markets.

8.7 Methodological Generalizability and Limitations

While our investigation evaluates the ISO 20022 'pacs.008' interbank credit transfer schema operating over NIST FIPS 203 and FIPS 204 cryptographic standards, the parameter-efficient adaptation principles developed in this work generalize across diverse high-velocity critical infrastructure domains. Similar transport-induced representation drift challenges confront automated supervisory control and data acquisition (SCADA) networks in electrical power grids, algorithmic high-frequency trading switches, and real-time defense command systems during post-quantum modernization.

A limitation of this study is its evaluation on emulated wide-area network topologies and synthetic financial transaction corpora necessitated by commercial banking confidentiality restrictions. Future collaborative research with central banks will evaluate our parameter-efficient adaptation pipeline across live production settlement streams during national post-quantum sandbox modernization trials.

In addition, future extensions of this parameter-efficient adaptation paradigm will investigate multi-modal routing topologies where interbank messaging streams integrate unstructured transaction attachments, including PDF invoices and smart contract state proofs. When these multi-modal documents are encapsulated within post-quantum digital signature envelopes, the resulting byte expansion exacerbates transport fragmentation, presenting fruitful avenues for extending our dual-objective invariance regularizers across vision-language and graph-structured routing transformers.

9. Conclusion

The mandatory migration of interbank financial communications to post-quantum cryptography under NIST FIPS 203 (ML-KEM) and FIPS 204 (ML-DSA) standards represents an indispensable defense against quantum cryptanalysis, yet its execution introduces severe operational challenges to high-throughput payment routing switches. In this investigation, we demonstrated that the transmission of expanded lattice-based cryptographic payloads exceeds standard TCP Initial Congestion Windows, forcing multi-packet fragmentation, tail latency inflation, and packet arrival jitter. When ingested by automated clearing switches, this transport friction induces acute latent representation drift within deep neural routing engines, depressing corridor assignment agreement to 76.4% and triggering severe queuing buffer exhaustion.

To overcome these vulnerabilities without computationally prohibitive full-parameter retraining, we introduced a parameter-efficient adaptation framework utilizing Low-Rank Adaptation (LoRA) governed by a dual-objective loss formulation. Through coupling transaction corridor classification with a cosine representation invariance regularizer, our architecture forces low-rank projection adapters to insulate the latent feature manifold against transport-layer perturbation noise.

Evaluated on a hardware-in-the-loop settlement testbed processing 20,000 transactions per second under ISO 20022 messaging schemas, our framework updates only 0.53% of model parameters while restoring corridor assignment agreement to 99.14% and top-1 clearing accuracy to 98.42%. The adapted switch preserves sub-20 millisecond median processing latency, eliminates queue buffer overflows during peak retail volume bursts, and incurs zero runtime memory or compute overhead through weight folding. These findings provide financial institutions, clearinghouse operators, and central bank regulators with a mathematically rigorous, operationally tractable architecture to ensure continuous transaction velocity and decision stability throughout the post-quantum transition.

References

[1] M. Bech, J. Hancock, T. Rice, and A. Wadsworth, "On the future of payments: Fast, digital and global," *BIS Quarterly Review*, vol. 2020, no. March, pp. 35–48, 2020.

[2] SWIFT Standards Committee, "Iso 20022: The global messaging standard for payments," tech. rep., Society for Worldwide Interbank Financial Telecommunication, 2023.

[3] M. Manning, E. Nier, and J. Schanz, *The Economics of Large-value Payments and Settlement: Theory and Policy*. Oxford University Press, 2009.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

[5] T. Roughgarden and É. Tardos, "How bad is selfish routing?," *Journal of the ACM*, vol. 49, no. 2, pp. 236–259, 2002.

[6] B. Sadanandan and V. Behzadan, "Mechanistically guided lora improves paraphrase consistency in medical vision-language models," *arXiv preprint arXiv:2603.00148*, 2026.

[7] D. Sikeridis, P. Kampanakis, and M. Devetsikiotis, "Post-quantum authentication in tls 1.3: A performance study," in *2020 IEEE International Conference on Network Softwarization (NetSoft)*, pp. 162–166, 2020.

[8] National Institute of Standards and Technology, "Module-lattice-based key-encapsulation mechanism standard (fips 203)," tech. rep., U.S. Department of Commerce, 2024.

[9] National Institute of Standards and Technology, "Module-lattice-based digital signature standard (fips 204)," tech. rep., U.S. Department of Commerce, 2024.

[10] Committee on Payments and Market Infrastructures, "Enhancing cross-border payments: Building blocks of a global roadmap," tech. rep., Bank for International Settlements, 2021.

[11] M. Allman, J. Chu, and S. Radhakrishnan, "Increasing tcp's initial window," RFC 6928, IETF, 2013.

[12] R. Avanzi, J. Bos, L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, J. M. Schanck, P. Schwabe, G. Seiler, and D. Stehlé, "Crystals-kyber: Algorithm specifications and supporting documentation," tech. rep., National Institute of Standards and Technology, 2021.

[13] L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, P. Schwabe, G. Seiler, and D. Stehlé, "Crystals-dilithium: Algorithm specifications and supporting documentation," tech. rep., National Institute of Standards and Technology, 2018.

[14] A. Collins, "Who governs the migration? pci dss, federal critical infrastructure policy, and the coming post-quantum mandate for u.s. payment systems," *Journal of Data Science, Predictive Analytics, and Big Data Applications*, vol. 9, no. 10, pp. 1–41, 2024.

[15] C. Paquin, D. Stebila, and G. Tamvada, "Benchmarking post-quantum cryptography in tls," in *International Conference on Post-Quantum Cryptography (PQCrypto)*, pp. 72–91, Springer, 2020.

[16] A. Langley, A. Riddoch, A. Wilk, A. Vicente, C. Krasic, D. Zhang,

F. Yang, F. Khasanova, P. Liang, J. Misra, *et al*., "The quic transport protocol: Design and internet-scale deployment," in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)*, pp. 183–196, 2017.

[17] N. Cardwell, Y. Cheng, C. S. Gunn, S. H. Yeganeh, and V. Jacobson, "Bbr: Congestion-based congestion control," *ACM Queue*, vol. 14, no. 5, pp. 20–53, 2016.

[18] T. Kokkola, *The Payment System: Payments, Securities and Derivatives, and the Role of the Eurosystem*. European Central Bank, 2010.

[19] A. Milne, "What is in it for us? network effects and standardisation in european retail payments," *Journal of Banking & Finance*, vol. 30, no. 6, pp. 1615–1630, 2006.

[20] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning (ICML)*, pp. 2790–2799, 2019.

[21] J. He, C. Zhou, X. Bao, P. Chen, W. Ning, and D. Tang, "Towards a unified view of parameter-efficient transfer learning," in *International Conference on Learning Representations (ICLR)*, pp. 1–18, 2022.

[22] J. Bos, C. Costello, L. Ducas, M. Hansen, T. Janssen, I. Mironov, M. Naehrig, V. Nikolaenko, C. Peikert, and D. Stebila, "Frodokem: Learning with errors key encapsulation," in *International Conference on Post-Quantum Cryptography (PQCrypto)*, pp. 30–47, Springer, 2018.

[23] E. Alkim, L. Ducas, T. Pöppelmann, and P. Schwabe, "Post-quantum key exchange—a new hope," in *Proceedings of the 25th USENIX Security Symposium*, pp. 327–343, 2016.

[24] L. Kleinrock, *Queueing Systems, Volume 1: Theory*. John Wiley & Sons, 1975.

[25] J. Corbet, "Bpf and xdp: High-performance packet processing in the linux kernel," *LWN.net Technical Report*, vol. 22, no. 4, pp. 1–12, 2020.

[26] P. Schwabe, D. Stebila, and T. Wiggers, "Post-quantum tls without handshake signatures," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1461–1480, 2020.

[27] J. Iyengar and M. Thomson, "Quic: A udp-based multiplexed and secure transport," RFC 9000, IETF, 2021.

[28] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3045–3059, 2021.

[29] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, pp. 1–26, 2021.

[30] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in Neural Information Processing Systems*, vol. 36, pp. 10088–10115, 2024.

[31] T. Gao, X. Yao, and D. Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6894–6910, 2021.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al*., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

[33] B. Vamanan, J. Hasan, and T. N. Vijaykumar, "Deadline-aware datacenter transport," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 4, pp. 115–126, 2012.

[34] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, no. 1, pp. 33–37, 1997.