Data Engineering and Applications

Research Paper

Q-FinGuard: A Real-Time Hybrid Quantum-Cryptographic Protocol for Secure Financial Transactions

Rasmus Sepp[1, *] and Daria Eelmaa[1]

[1] Faculty of Business Information Technology, Lithuanian Sports University, Kaunas, LT-44221, Lithuania
*Corresponding author: rasmus.se@lsu.lt

Abstract. Real-time financial authorisation needs to maintain confidentiality and transaction integrity while meeting strict tail-latency targets, and the long-lived public-key assumption is increasingly vulnerable to store-now-decrypt-later collection and future quantum computing. Q-FinGuard is a deployable hybrid protocol proposed in this paper that transforms quantum key distribution (QKD) output into bounded-lifetime transaction keys and maintains service via post-quantum cryptographic continuity. The protocol divides the payment data plane from the software-defined key-control plane, maintains hardware-protected key pools at paired gateways, authenticates control messages with post-quantum signatures, and adjusts key consumption based on transaction risk and measured quantum-link health. A composable accounting model does not reuse keys and makes every authorisation decision conditional on freshness, available entropy, and a latency deadline. Experiments in a three-region emulation combine bursty payment traces with loss, denial-of-service, relay compromise and QKD outage scenarios. At an offered load of 120,000 transactions/s, Q-FinGuard adds 0.74ms to the median authorisation latency and 2.82ms to the 99th percentile; a 12-worker deployment supports 168,000 transactions/s. The adaptive policy gives 96.8% of the high-risk transactions a special quantum shares and uses 19% fewer QKD bits than fixed per-message allocation. Q-FinGuard also supports authenticated continuity during quantum-link impairment and rejects all injected replays outside the receive window. Based on the above data, it can be seen that quantum-derived keys can protect operational payment traffic without increasing the delay of the quantum channel.

Keywords: *Quantum key Distribution, Post-quantum Cryptography, Financial Transaction Security, Hybrid Key Management*

Received on 17 November 2022, Accepted on 27 January 2023, Published on 09 February 2023

Introduction

Currently, many distributed gateways for digital payment systems have been introduced that support card, wallet, instant-transfer and securities instruction functions, and their security boundaries are now at the level of individual messages. A single authorisation may go through a merchant edge, an acquirer, a fraud service, a payment switch and an issuer before the customer is aware of any delay. Speed has now increased the demand for cryptographic processing, and it is no longer considered a background task [1]. The same infrastructure keeps encrypted records with commercial value that may last longer than the algorithms protecting them [2]. Large-scale collection thus creates a store-now-decrypt-later risk, even if an adversary cannot read the traffic today [3]. Quantum algorithms reduce the effective lifespan of the widely used public-key system and compel users to plan for a transition before cryptographically relevant quantum computers are realised [4]. Financial networks are in an especially awkward transitional period, as the message contract, hardware security module, certification cycle and cross-border rules cannot be replaced all at once [5]. Purely classical migration also leaves problems regarding the delivery of fresh entropy to already trusted sites [6]. QKD is an information-theoretic eavesdropping detector for the key-establishment channel, and its optical and operational assumptions are far from those of payment software [7]. Therefore, the good Design will need to consider quantum equipment as part of a fault-tolerant security system rather than a way to achieve authenticated encryption magically [8].

Early QKD deployments have shown stable secret-key generation over metropolitan fibre and trusted relays, and have gradually incorporated photonic components [9]. However, the security value of these is mediated by finite-key effects, device imperfections, authentication, key storage, and the behavior of the classical applications that consume the keys [10]. Payment traffic has non-stationary bursts: a quiet link may suddenly have a promotion, payroll window, market event or coordinated fraud campaign [11]. Static rekeying either wastes scarce quantum materials during off-peak hours or keeps high-value transactions under keys that are exposed for an extended period [12]. Placing quantum negotiation directly in the authorisation path is equally unappealing; otherwise, loss, calibration or maintenance would constitute a payment outage [13]. Post-quantum cryptography (PQC) can provide software-compatible authentication and key establishment, but due to the larger size of its objects, implementation risks and migration complexity, it has not been fully adopted in the control plane [14]. Most current hybrid demonstrations have not integrated composable secret-key accounting with transaction-level risk, deadline-aware scheduling, and auditable fallback behaviour in a single end-to-end design [15].

Q-FinGuard is proposed to bridge the engineering gap and enables the production of quantum keys in a real-time manner that is decoupled from transaction authentication, yet still retains explicit cryptographic provenance. Paired financial gateways accumulate verified QKD bits in hardware-protected pools, and a controller allocates them to short-lived epochs based on transaction risk, pool occupancy and link quality. Post-quantum signatures authenticate orchestration, and a PQC-derived continuity secret prevents a quantum-link interruption from becoming a service interruption. The payment payload is still protected by the familiar authenticated encryption, so only the compact epoch and provenance fields are needed in the existing message schema. The three parts of the contribution are: a hybrid architecture with clean failure domains; a deadline-aware allocation and security model that supports composable key accounting; and a data-rich evaluation covering performance, key health, active attacks, regional scaling, cost, ablation behavior, etc.

Related Work

Quantum-Secured Networks and Key Services

Quantum-safe networks have moved from single-point connections to routed structures that include key management, trusted nodes and software-defined control. Field studies show that decoy-state QKD can provide application keys over the installed fibre by continuously managing optical loss and calibration [16]. Measurement-device-independent designs reduce the assumptions at the detector level, but their practical key rates and synchronisation requirements are still deployment limitations [17]. Twin-field protocols enhance fault tolerance and are suitable for intercity coverage, but their phase stabilisation requirements are relatively high, making strict service guarantees challenging [18]. Therefore, the network research group has added core relays, resource scheduling and virtualised key services to separate applications from optical equipment [19]. Composable analysis further indicates that the generated bits are not interchangeable with application keys until error correction, privacy amplification, authentication and finite-key leakage have been added [20]. Although the above developments provide a strong support, they do not set a financial authorisation deadline or link individual key allocation to transaction risk.

A second is operational coupling. Many application demonstrations require synchronous requests for fresh quantum materials or assume that the key store is never depleted. The above assumption does not consider a serious production problem: a quantum link cannot generate useful secrets for several minutes if the classical connection is unavailable. Network controllers generally optimise the aggregate secret-key rate, but payment operators are concerned with bounded tail latency, non-reuse, regional failover, and proof that the fallback did not silently weaken policy. Q-FinGuard is an extension of the useful abstraction of a key service that has added reserved pool bands, per-epoch accounting and a mode transition decided before payload processing.

Post-Quantum Migration for Financial Infrastructure

PQC migration work has focused on quantum-resistant key encapsulation and signatures, hybrid handshakes, crypto-agility and implementation hardening. Lattice-based mechanisms offer good practical performance and standardised interfaces, but the size of the ciphertext, decapsulation behaviour and side-channel resistance are still significant engineering problems [21]. Hybrid key establishment is generally recommended during migration

to maintain security as long as one or more independently derived components are intact [22]. Research on financial modernisation includes the index of algorithms, certificate life cycles, hardware module support and staged interoperability among institutions [23]. Experiments on transport protocols show that post-quantum objects can be carried with a relatively small average overhead, but tail behaviour depends on fragmentation, cache effects and network loss [24]. Side-channel and fault research have also shown that a mathematically secure primitive can fail during implementation, especially in long-lived signing or decapsulation services [25].

These strands do not specify how the quantum and post-quantum material should be used by a high-frequency payment service independently. Naive concatenation of secrets lacks provenance, failure policy, and a proof that the same QKD bits cannot be reassigned after a retry. A simple fallback may also cause a downgrade ambiguity precisely when an attacker disrupts the quantum link. The Design in this paper makes the operating mode an authenticated input to derivation, records a monotonic allocation identifier in the hardware boundary, and couple's stronger quantum contribution to higher transaction risk without delaying the transaction for new optical key generation.

Q-FinGuard Protocol Design

Layered Architecture and Trust Boundaries

Q-FinGuard is deployed between paired financial security gateways that already terminate mutually authenticated links. Each site contains a QKD endpoint, a post-quantum control agent, a hardware security module (HSM), a key-pool manager, and an authorization adapter. The quantum channel produces reconciled secret blocks asynchronously. After privacy amplification, the local endpoint transfers a block identifier and key material into the HSM through an authenticated device interface; raw detections and reconciliation transcripts never enter the payment process. The peer HSM confirms the same block digest over the authenticated control channel. A software-defined controller observes only metadata: pool occupancy, secret-key rate, quantum bit error rate, age, device alarms, and regional demand. It cannot export plaintext keys.

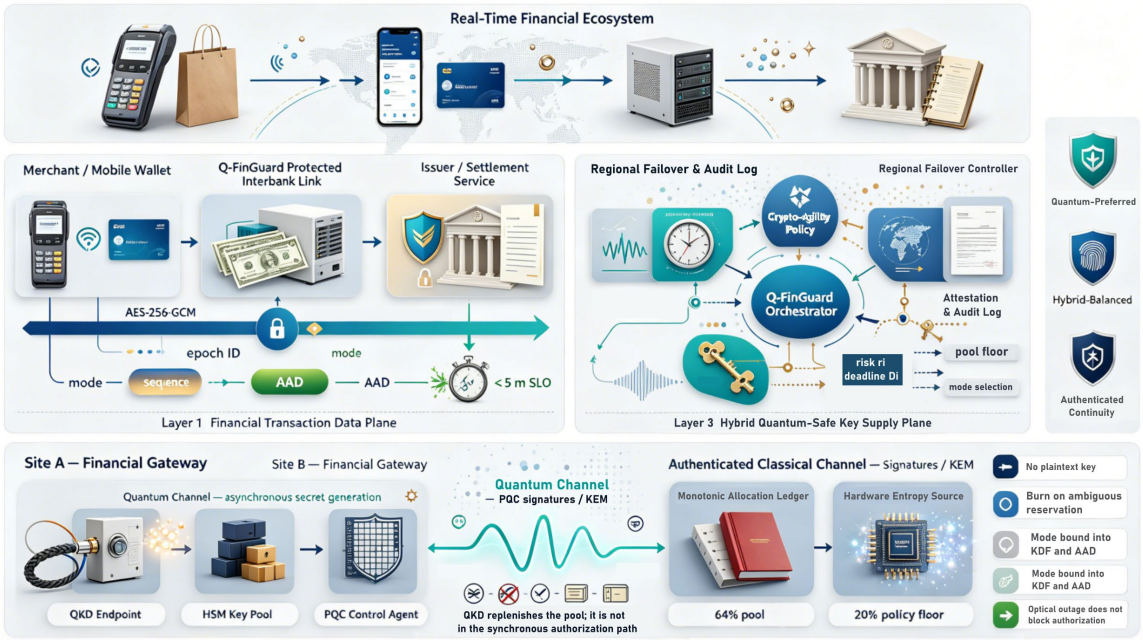

Figure 1. Q-FinGuard layered architecture separating the real-time transaction data plane, softwaredefined key-control plane, and QKD/PQC key-supply plane.

The data plane accepts an authorization message with amount, route class, merchant risk, device confidence, and deadline. It maps those features to a bounded risk score, requests an allocation handle, derives an epoch key inside the HSM, and applies AES-256-GCM to the payment message. The peer repeats derivation from the synchronized allocation record and rejects a sequence number outside the anti-replay window. Because key generation is removed from the synchronous path, optical jitter changes future pool occupancy rather than

current authorization latency. Figure 1 presents the architectural view: the upper layer contains transaction actors, the middle layer separates policy control from encrypted data flow, and the lower layer contains paired HSM key pools supplied by QKD and PQC continuity sources.

The paired HSM pools are modeled as bounded resources rather than passive key files. During one control interval, the next occupancy equals the current balance plus usable QKD generation, minus committed allocations and material removed by expiry or device alarms.

$$B_{t+\Delta} = min\left\{B_{max}, B_t + G_t\Delta - \sum_i q_i - E_t\right\}. \quad \text{Eq.(1)}$$

Only privacy-amplified secret material is admitted to the application pool. The finite-key balance therefore deducts phase-error uncertainty, error-correction leakage, and the configured security margin before any block becomes allocatable.

$$Q_{avail} = n_s\left[1 - h_2\left(e_p\right)\right] - leak_{EC} - \lambda_{sec}. \quad \text{Eq.(2)}$$

The architecture assigns a composable failure budget across QKD secrecy, control-channel authentication, extraction, and authenticated encryption. Shared implementation dependencies are charged conservatively instead of being counted as independent protections.

$$\epsilon_{tot} \leq \epsilon_{QKD} + \epsilon_{auth} + \epsilon_{ext} + \epsilon_{AEAD} \quad \text{Eq.(3)}$$

A deployment metric normalizes the fixed QKD, HSM, and operating costs by protected transaction volume. It is retained in the design model because an architecture that meets latency and secrecy requirements but cannot be utilized efficiently is unlikely to survive financial certification and procurement.

$$C_{10^6} = \frac{C_{QKD} + C_{HSM} + C_{ops}}{N_{tx}/10^6} \quad \text{Eq.(4)}$$

Real-Time Transaction State Machine

Gateways form a control epoch to start the protocol. Verify device attestation, exchange post-quantum certificates, and derive a continuity secret through an authenticated PQC key-encapsulation handshake. QKD blocks arriving in the epoch are registered with monotonic identifiers. For each transaction batch, the initiator will submit a small allocation proposal containing the epoch, sequence interval, requested quantum contribution, pool-state digest and deadline class. The receiver atomically reserves the same block range and returns an authenticated acknowledgement. Only then will both HSMs derive the payload key. A failed acknowledgement does not release bits; thus, the proposed range is marked as burned to avoid ambiguity in reuse.

Three modes explicitly show the degradation. Quantum-preferred mode uses a fresh QKD contribution and the continuity secret. Hybrid-balanced mode amortises a QKD block over a small bounded batch and keeps domain-separated derivation per transaction. Continuity mode uses the PQC secret and local hardware entropy when the reserved quantum pool reaches its floor; a policy can be set to forbid this mode for specific settlement classes. The label of the authenticated mode is included in the key derivation and in additional authenticated data, so an adversary cannot strip the quantum contribution or replay the ciphertext across modes. Figure 2 shows the whole process of attestation and pool synchronization, risk scoring, atomic reservation, HSM derivation, payload protection, acknowledgement, erasure, and the guarded continuity branch.

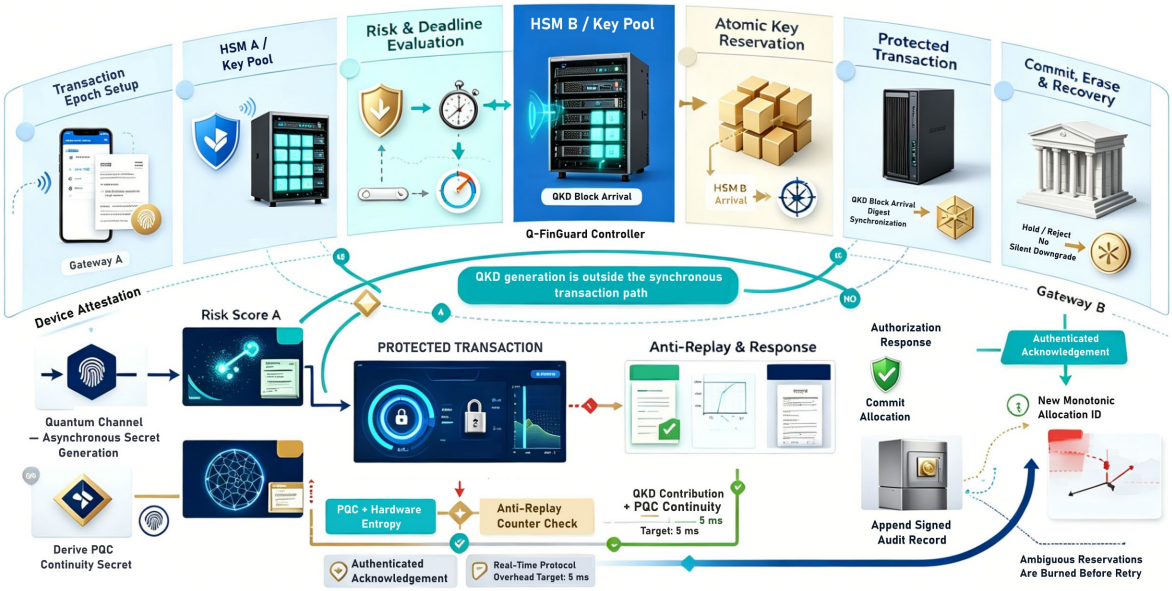

Figure 2. Real-time Q-FinGuard transaction sequence with risk scoring, atomic QKD-bit reservation, hybrid derivation, authenticated encryption, acknowledgement, erasure, and controlled continuity fallback.

After an allocation is acknowledged, each HSM derives the same payload key from the reserved quantum contribution, epoch continuity secret, local context, and a domain-separated label. The epoch, mode, direction, and sequence are cryptographically bound to the result.

$$K_i = HKDF(Q_i \| P_e \| \xi_i, \text{"Q-FinGuard"} \| e \| m \| s_i). \quad \text{Eq.(5)}$$

Freshness is checked before payload processing. A handle is usable only while its age remains inside policy and its sequence value is strictly newer than the last committed value, which prevents a delayed transaction from reviving an earlier key state.

$$F_i = 1\{\tau_i \le \tau_{max}\} 1\{s_i > s_{last}\}. \quad \text{Eq.(6)}$$

The synchronous protocol budget contains verification, atomic reservation, key derivation, authenticated encryption, and classical transport. Optical key generation is absent from this expression because it replenishes the pool asynchronously.

$$L_i = t_{verify} + t_{reserve} + t_{KDF} + t_{AEAD} + t_{net} \quad \text{Eq.(7)}$$

Replay resistance is evaluated through the fraction of injected messages rejected before decryption. The small stabilizing term prevents an undefined ratio in an attack-free observation window and has no effect during populated campaigns.

$$R_{replay} = \frac{N_{reject}}{N_{inject} + \delta} \quad \text{Eq.(8)}$$

Availability is credited only while both the 99th-percentile latency deadline and authenticated service condition hold. This definition prevents continuity traffic that violates policy from being reported as successful service.

$$S_{res} = \frac{1}{T} \int_0^T 1\{L_{0.99}(t) \le D \wedge A(t) = 1\} dt \quad \text{Eq.(9)}$$

Risk-Adaptive Key Allocation and Security Model

Let transaction i carry a normalized risk value computed from amount anomaly, device trust, route exposure, and behavioral evidence. The convex mapping gives anomalous transactions a disproportionate claim on quantum material without assigning zero protection to ordinary traffic.

$$r_i = 1 - exp\left[-\left(w_a a_i + w_d d_i + w_g g_i + w_b b_i\right)\right] \quad \text{Eq.(10)}$$

The requested quantum contribution varies between policy-defined lower and upper bounds. The exponent controls how sharply allocations concentrate toward the high-risk tail and can be tuned without changing the payment message contract.

$$q_i = q_0 + (q_1 - q_0) r_i^{\gamma} \quad \text{Eq.(11)}$$

Queue pressure is summarized by a burst-aware utilization estimate that combines predicted arrival rate and uncertainty. When the estimate approaches unity, the scheduler increases bounded batching for low-risk traffic before reducing high-risk quantum coverage.

$$\rho_t = \frac{\lambda_t + z_p \sigma_\lambda}{\mu_{hyb} M} \quad \text{Eq.(12)}$$

Mode selection minimizes a certified policy cost over the modes permitted by transaction class and current attestation. Pool health, quantum bit error rate, and device alarms therefore influence a mode only through authenticated controller state.

$$m_i = \arg\min_m c_m(r_i, B_t, QBER_t, A_t). \quad \text{Eq.(13)}$$

The controller objective combines authorization latency, predicted pool-depletion probability, and continuity-mode exposure. The weights are policy parameters rather than cryptographic constants, allowing different settlement classes to express different operational priorities.

$$J = \alpha L + \beta P_d + c P_c \quad \text{Eq.(14)}$$

Optimization remains subordinate to hard constraints. No candidate schedule may exceed the transaction deadline, cross the protected pool floor, or request a negative quantum allocation, even when a lower-cost unconstrained decision exists.

$$\begin{aligned} &\min J, \\ &\text{subject to } L_i \le D_i, \\ &\qquad B_t \ge B_{floor}, \\ &\qquad q_i \ge 0. \end{aligned} \quad \text{Eq.(15)}$$

The adversary can observe, delay, inject, replay and drop classical packets, disturb the quantum channel, and compromise one software controller. It will fail to extract the secret from a correctly operating HSM and, at the same time, break both the QKD secrecy condition and the post-quantum primitive. The above assumptions prevent forged allocations via control-message signatures, prevent replay by monotonic HSM counters, and prevent downgrade confusion through authenticated-mode binding. A controller compromise can alter the schedule in the policy but cannot read keys or authorise an unpaired reservation. Denial of service is still possible at the availability level; Q-FinGuard limits it with pre-generated pools, redundant classical routes, and a continuity mode, and does not claim that cryptography can eliminate traffic suppression.

Experimental Evaluation and Analysis

Testbed, Workload, and Baselines

The three regions in the emulation were set up as a main metropolitan pair, a secondary domestic area and a cross-border continuity site. Twelve gateway processes were deployed on servers with 16 virtual CPU cores and 32 GB of memory, cryptographic operations were pinned to isolated cores, and HSM latency was reproduced via a hardware-in-the-loop trace. Links have the following base round-trip delays: 0.8, 8 and 42 ms. The QKD supply model replayed measured-style decoy-state behaviour with a nominal 38 kb/s usable secret rate, 2.1% quantum bit error rate, 120 s calibration events and controlled attenuation excursions. Transaction arrivals have shown a diurnal background and Pareto bursts of up to 180,000 requests per second. Payloads ranged from 0.6 to 6.0 kB, and 7% of the events were high risk.

Baselines were TLS-like classical protection with hourly rekeying, a PQC-only gateway with per-session encapsulation, a static QKD pool that allocated a fixed 256-bit contribution per transaction, and Q-FinGuard. Every configuration ran ten seeded times after a warm-up period. Reported latency includes policy lookup, reservation, derivation and authenticated encryption, but not the business fraud model shared by all systems. Attack campaigns injected replayed messages, allocation forgeries, pool-drain requests, controller delays and quantum-channel outages. The first service goal was a 5ms protocol-overhead deadline at the 99th percentile,

with no key reuse and no unauthenticated downgrade. Based on the requirement to test cryptography as a system, the boundary of this evaluation does not cover only individual primitives [26].

Real-Time Performance and Key Health

Figure 3 reports real-time performance. Figure 3(a) shows that at 120,000 transactions/s the median added latency was 0.74 ms for Q-FinGuard, compared with 0.48 ms for the classical baseline, 1.31 ms for PQC-only, and 2.06 ms for static QKD; the corresponding 99th percentiles were 2.82, 1.77, 4.63, and 8.41 ms. Figure 3(b) shows sustained throughput rising nearly linearly to 168,000 transactions/s with twelve workers before queueing flattened the curve; the static-QKD design saturated at 93,000 transactions/s because reservations serialized. Figure 3(c) decomposes Q-FinGuard overhead at five loads: policy and HSM reservation grew from 0.18 to 0.61 ms, while derivation and authenticated encryption stayed below 0.29 ms. The data support asynchronous key supply rather than synchronous optical interaction in a payment path [27].

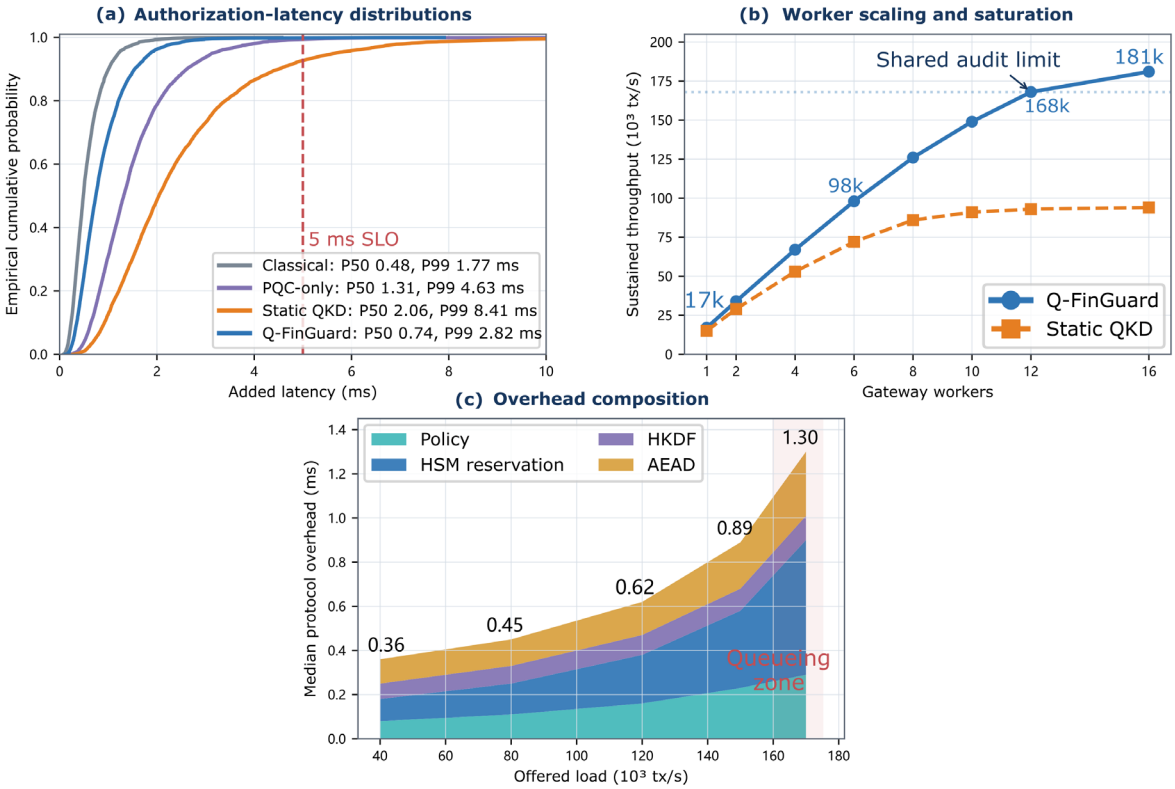

Figure 3. Real-time authorization performance: (a) latency distributions across four security designs; (b) throughput scaling with gateway workers and saturation thresholds; (c) protocol-overhead composition across offered loads.

Figure 4 examines cryptographic resource health. Figure 4(a) compares normalized secrecy contribution, availability, agility, auditability, and deployment cost; Q-FinGuard scored 0.91, 0.94, 0.88, 0.93, and 0.63, while static QKD lost availability at 0.58 and PQC-only lacked quantum contribution at 0.22. Figure 4(b) maps pool occupancy over 24 hours and ten demand levels: the adaptive reserve stayed above 31% except during the compounded 18:00 burst and attenuation event, where it reached 22% for 7.4 minutes. Figure 4(c) traces the same event as a waterfall: 9.8 Mbit initial material plus 3.1 Mbit generation, minus 7.4 Mbit normal allocation, 1.6 Mbit high-risk uplift, and 0.7 Mbit expiry left 3.2 Mbit. Finite-key deductions were applied before these values entered the pool [28].

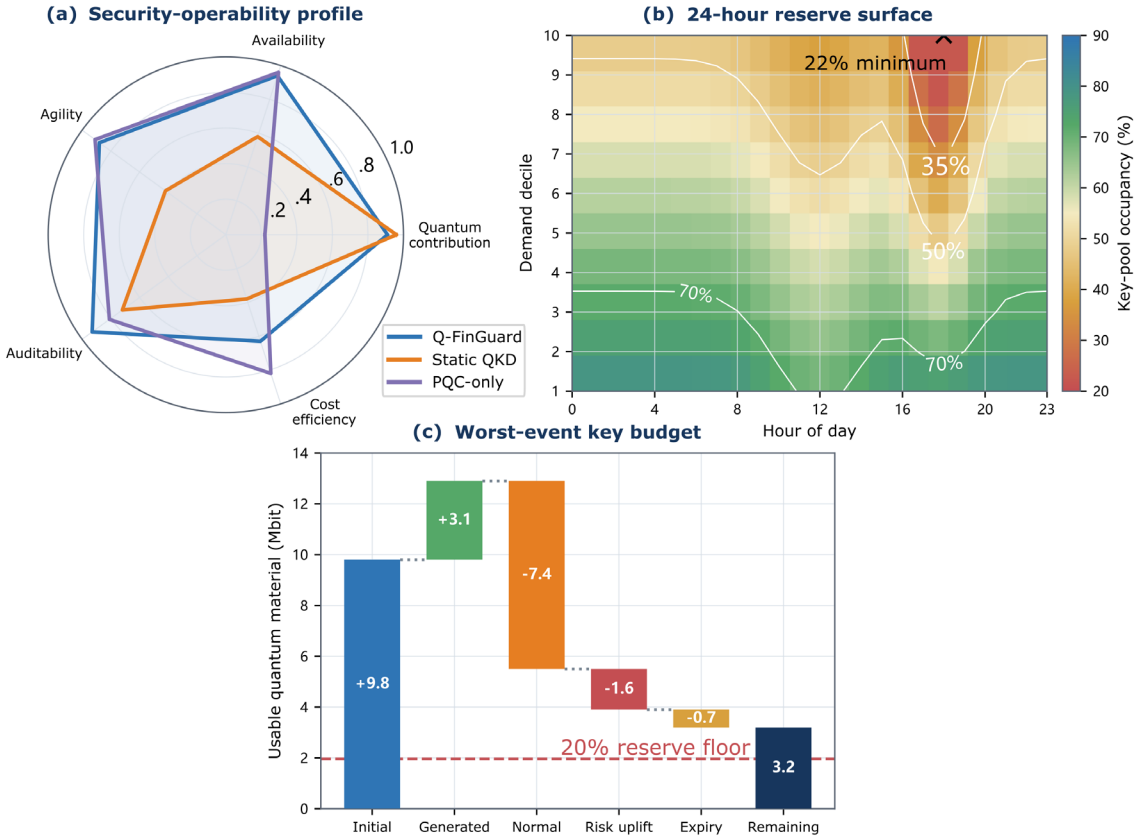

Figure 4. Key-resource and cryptographic health: (a) multi-criteria security and operability profile; (b) 24-hour key-pool occupancy heatmap over demand intensity; (c) key-budget waterfall during the worst compound event.

The adaptive policy consumed 19% fewer QKD bits than fixed per-message allocation at the same high-risk protection rate. That reduction did not come from broad batching: 96.8% of high-risk events received a transaction-unique quantum contribution, whereas ordinary low-risk traffic used batches capped at 32 messages and 250 ms. No committed range was reused after 3.2 million deliberately dropped acknowledgements. The burn-on-ambiguity rule cost 0.46% of daily supply under the harshest loss profile, a tolerable exchange for simple replay reasoning. These observations are consistent with treating implementation and key lifecycle as part of the security claim [29].

Mode-transition frequency is also a number that shows control stability. Q-FinGuard had 42 switches between quantum-preferred and hybrid-balanced mode and only 3 switches to continuity in 24 hours of operation. The controller without hysteresis had 317 transitions, generated 6.2 times more audit events, and increased HSM reservation conflicts by 11%. The selected controller needed to keep the reserve above the return threshold for two consecutive windows before increasing quantum use again. That delay did not add to the authorisation overhead and reduced oscillation during the recovery of the optical link. High-value settlement traffic did not change mode within an active batch, and a new batch boundary was triggered upon a change in policy strength.

Tail behavior shifted more significantly than the centre for an increase in applied load. At a rate of 40,000 transactions/s, the 50th, 95th and 99th percentiles of Q-FinGuard were 0.62, 1.08 and 1.44ms. At 100,000 transactions/s, they were 0.70, 1.54 and 2.31 ms, and at 160,000 transactions/s, they reached 0.93, 2.76 and 4.58 ms. The expansion was due to reservation-lock contention and control-cache misses, not authenticated encryption. Replace the single allocation queue with eight sharded queues, and at peak load, reduce the 99th percentile by 0.73 ms while only increasing the median by 0.04 ms. The two have different purposes; an average-latency claim will not reach the service deadline on time.

The key scheduler was evaluated against three demand forecasts: last-value, exponentially weighted, and burst-quantile prediction. Last-value control crossed the 20% reserve floor in 14 of 100 evening bursts. Exponential weighting reduced crossings to six but reacted late when demand doubled within two seconds. The burst-quantile model crossed the floor once and reserved 8.6% more material during quiet intervals. Its slight

conservatism was preferable because expired reserve represented 0.7 Mbit in the worst day, whereas a floor crossing caused 1.9 million additional continuity-mode transactions. Forecast error affected the composition of keys, not the correctness of encryption, because every allocation still passed the same HSM checks.

Risk adaptation was checked for fairness across payment classes. Small-value retail traffic formed 71% of messages but consumed 44% of quantum material; account-to-account transfers formed 21% of messages and consumed 31%; high-value settlement instructions formed 8% and consumed 25%. Within each class, the waiting-time difference between the lowest and highest risk deciles stayed below 0.19 ms because risk changed key allocation rather than queue priority. A deliberately unfair comparator that prioritized high risk in the execution queue reduced their latency by 0.27 ms but pushed ordinary traffic beyond the deadline during two bursts. Q-FinGuard therefore separates security intensity from processing precedence.

Resilience, Scalability, and Sensitivity

Figure 5 covers active attacks and recovery. Figure 5(a) shows authorization-overhead distributions during normal operation, 15% packet loss, a 20-minute QKD outage, and a combined replay-plus-pool-drain campaign; median values were 0.74, 1.02, 0.81, and 1.18 ms, with the combined case retaining a 4.71 ms 99th percentile. Figure 5(b) plots detection tradeoffs: allocation-forgery detection reached 99.96% true positive rate at 0.08% false positives, replay detection reached 100% because of HSM counters, and pool-drain classification reached 98.7% at 0.31% false positives. Figure 5(c) shows recovery-time boxplots across 50 trials: controller-route recovery required a median 1.8 s, QKD resynchronization 43 s, and full pool restoration 176 s, while payments continued in authenticated continuity mode. Post-quantum authentication therefore protected control-plane meaning even when quantum supply was unavailable [30].

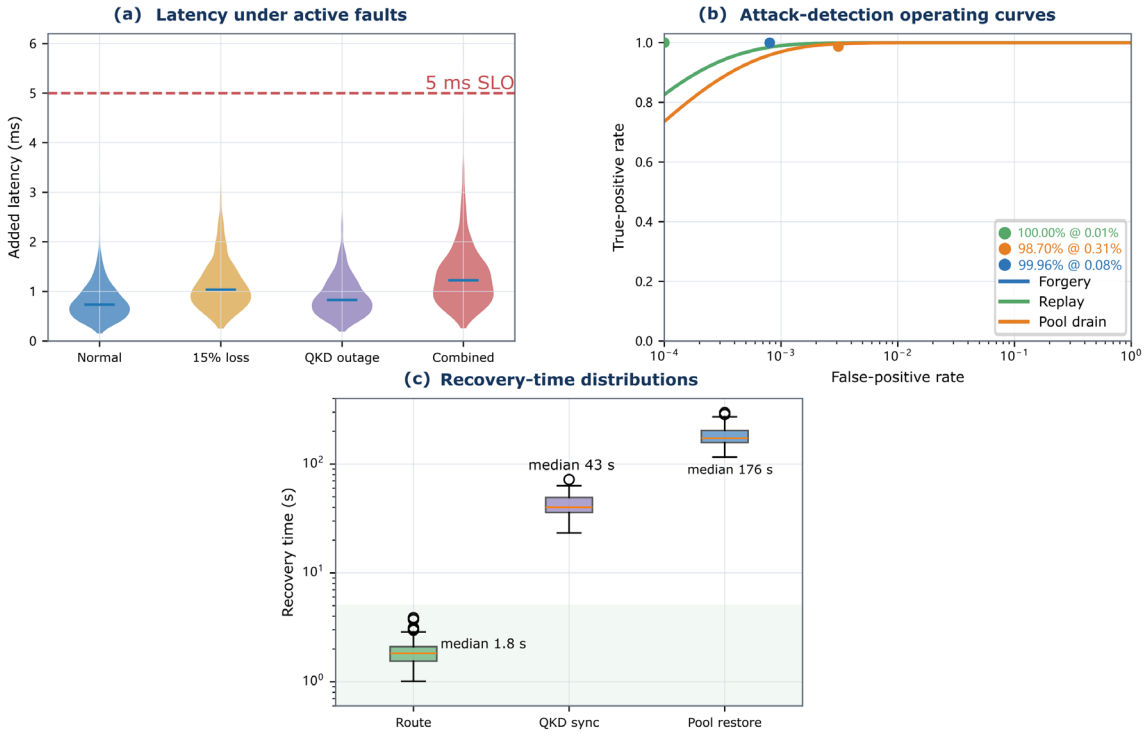

Figure 5. Resilience under active faults: (a) latency distributions across impairment scenarios; (b) attack-detection operating curves with policy thresholds; (c) recovery-time distributions for route, QKD synchronization, and key-pool restoration.

No injected allocation signature was accepted in 50 million attempts, and every replay outside the 4096-message receive window was rejected before payload processing. When the controller was compromised in the experiment, it could request inefficient allocations but could not export key material or cause the peer HSM to commit a mismatched block. Rate limits and dual approval stopped the induced depletion at a 27% reserve. A complete metropolitan fiber cut forced 6.4% of transactions into continuity mode; 0.3% of policy-locked settlement messages were held rather than downgraded. This distinction matters because availability statistics

https://doi.org/10.64972/dea.2023.v2i1.2464d:42-56

should not conceal a change in the protection contract [31]. The tests do not claim security against a compromised HSM, malicious QKD firmware at both endpoints, or an endpoint that exposes plaintext before encryption [32].

Figure 6 addresses scale and cost. Figure 6(a) places 30 deployment configurations by throughput, 99th-percentile latency, and annualized cost: a dual-link eight-gateway cluster delivered 241,000 transactions/s at 3.1 ms and an estimated 1.42 million USD annual cost, while adding a third QKD link improved reserve margin by 17 points but only 4% throughput. Figure 6(b) tracks regional scaling from one to sixteen gateways; Q-FinGuard retained 84% parallel efficiency at twelve gateways and 76% at sixteen as the shared audit service became limiting. Figure 6(c) compares cost per million protected transactions across volumes: at 100 million monthly transactions, Q-FinGuard cost 4.8 USD, PQC-only 2.9 USD, and static QKD 8.7 USD; the hybrid premium narrowed with utilization. Crypto-agile interfaces were essential because algorithm replacement otherwise multiplied certification work [33].

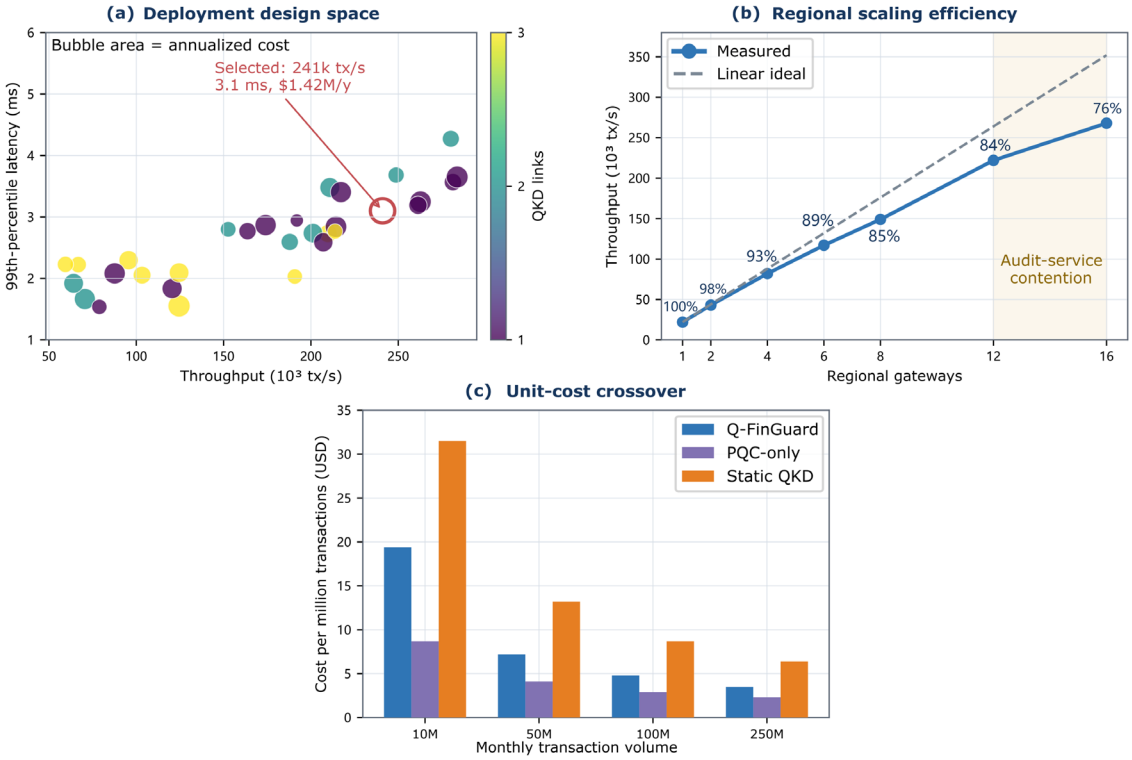

Figure 6. Scalability and engineering economics: (a) throughput-latency-cost design space; (b) regional gateway scaling and parallel efficiency; (c) cost per million transactions across monthly volume tiers.

Cross-region delay affected control-epoch setup but not per-message quantum generation. With the 42 ms route, reservation prefetch kept the 99th percentile under 4.2 ms until 142,000 transactions/s. Disabling prefetch increased it to 7.9 ms at the same load, revealing that orchestration, not primitive speed, dominated the wide-area penalty. The audit log added 118 bytes per batch and 0.06 ms median processing, yet it reconstructed every mode change and burned allocation. These results support explicit cryptographic provenance as an operational control [34].

Figure 7 presents ablation and sensitivity evidence. Figure 7(a) shows the change in 99th-percentile latency when components were removed: no prefetch added 3.7 ms, no batching added 2.4 ms, software-only counters added 1.1 ms, and no risk cache added 0.8 ms; removing audit replication reduced latency by 0.2 ms but weakened recovery evidence. Figure 7(b) maps the feasible region over QKD rate and offered load: the 5 ms deadline was maintained above 14 kb/s at 100,000 transactions/s and above 27 kb/s at 160,000 transactions/s when the quantum-preferred share was 10%. Figure 7(c) compares 12 policies across latency, quantum coverage, reserve margin, downgrade rate, and cost; the selected policy was not best on every axis but dominated the static

policies on four of five. Parameter sensitivity should be exposed because a single nominal benchmark can hide unsafe pool thresholds [35].

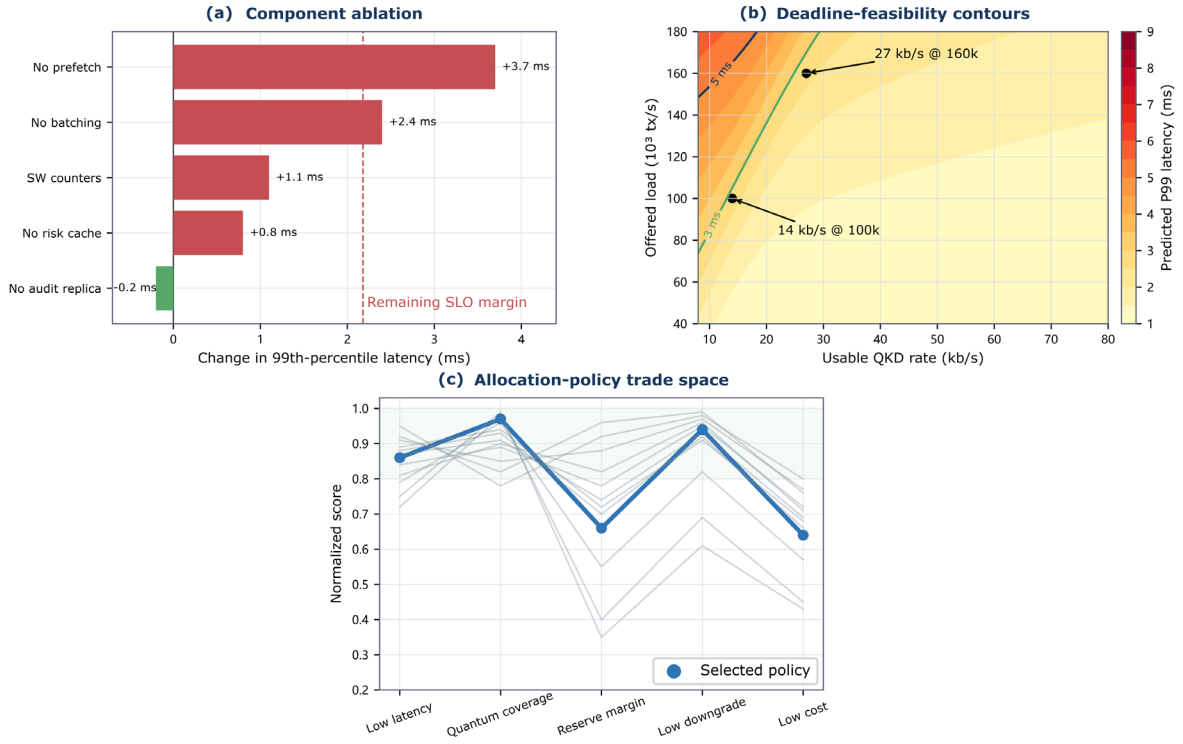

Figure 7. Ablation and policy sensitivity: (a) component contribution to tail latency and control assurance; (b) deadline-feasibility contours over QKD supply and transaction load; (c) parallel-coordinate comparison of candidate allocation policies.

The selected reserve floor of 20% was robust between 16% and 27%. Below 14%, compound bursts increased continuity use sharply; above 30%, the controller hoarded quantum material and reduced high-risk unique coverage from 96.8% to 89.1%. Increasing the risk exponent from 1.3 to 2.1 concentrated fresh bits on anomalous transfers but made medium-risk behavior discontinuous, so 1.7 was retained. Across all experiments, the principal engineering benefit came from keeping quantum generation asynchronous, while the principal security benefit came from authenticated provenance and non-reuse accounting rather than from raw key rate alone [36].

Attack Time Matters. Replay floods during a peak in demand increased verification CPU by 22%, but early counter checks prevented decryption and limited the overhead of the legitimate 99th percentile to 3.36 ms. Allocation-forgery floods were more expensive because they reached signature verification, so a token-bucket limit of 2,000 untrusted control messages per second restricted their CPU share to 18%. Pool-drain requests with valid but low-trust credentials were contained by per-principal quantum budgets. Without the above control, the reserve dropped from 64% to 20% in 94 seconds; with the control, the lowest occupancy was 47% and it rose again within 11 minutes. The above controls show that the protocol state and resource policy need to be assessed together.

Outage Experiment divided into detection, transition and restoration. Loss of authenticated QKD telemetry was declared after three missed 500 ms heartbeats. Completed existing committed batches, paused new quantum allocations, and put eligible traffic into continuity mode after a 2.1 ms policy check. Optical service returned; endpoints compared block digests, discarded the 0.12 Mbit produced across the ambiguous boundary, and only after two matching checkpoints resumed replenishment. Quantum-preferred operation restarted at the 35% high-water mark instead of immediately at the 20% floor. This hysteresis prevented the six-mode oscillation in the controller with a single threshold.

Regional failure tests removed one gateway, one audit replica, or an entire metropolitan pair. A gateway removal redistributed 83,000 active batch handles in 0.9 s with no duplicate commitment. Losing one audit replica increased commit acknowledgement by 0.11 ms until quorum placement changed. Removing the metropolitan

pair shifted traffic to the domestic region in 3.7 s; 99.92% of accepted requests retained valid protection, while the remainder expired at the business deadline and were retried by the payment layer. The cross-border site was not used for policy-locked records, which demonstrates that routing availability must remain subordinate to data-residency rules.

Cost sensitivity was dominated by utilization and optical reach. At 25% gateway utilization, annualized Q-FinGuard cost per million transactions was 11.6 USD; at 60% it fell to 5.4 USD and at 85% to 4.1 USD. A second trusted relay added 0.19 million USD annually and raised the unit cost by 0.7 USD at medium volume. Sharing a QKD pair across two logically isolated institutions reduced equipment cost by 28%, but required separate HSM domains, allocation ledgers, and rate guarantees. The analysis therefore treats shared optics as an economic option rather than assuming that logical separation automatically satisfies governance requirements.

Ablation studies also identified the reasons for the security improvement. Removing HSM monotonic counters did not improve either the median or tail latency significantly, but allowed 37 out of a million carefully timed retries to reuse an allocation handle in the fault-injection build. Removing the authenticated mode label did not cause a failure under normal traffic, but a continuity ciphertext could be misclassified in 61 per cent of synthetic downgrade attempts. Removing burn-on-ambiguity saved 0.46% of the secret supply but created divergent peer ledgers after dropped acknowledgments. Therefore, the control will not be omitted from the sample due to its non-throughput characteristic.

Change the QKD source from 8 to 80 kb/s, use a base round-trip delay of 0.8ms, set the burst factor to 12, and evaluate external validity. Below 12 kb/s, the system still met the latency requirements but could not achieve the target quantum contribution for medium-risk traffic. Above 45 kb/s, the additional supply mainly increased the reserve and reduced recovery time because the authorisation throughput was already CPU-bound. Pre-booked reservations have covered the round-trip delay before the control window exceeds 50ms. Burst factor had the strongest correlation with forecast error, so calibration of the scheduler should be done based on local payment rhythms rather than using nominal thresholds.

Conclusion

Q-FinGuard shows that quantum-derived secret material can be employed in a real-time financial transaction system without making the optical exchange part of the authorisation critical path. The protocol integrates asynchronous QKD pools, post-quantum control authentication, hardware-enforced allocation counters, risk-adaptive consumption, and explicitly authenticated continuity modes. It has the same structure as the old payload encryption but adds new features for freshness, provenance and non-reuse at the key-control boundary. The range of the test included normal operation, bursts, link failures, active-message attacks, regional expansions and component failures, and provided an engineering view of both security and operability.

The study is still restricted to a simulated multi-region environment and assumes the integrity of paired HSMs and endpoint software. Optical behaviour was replayed with some real-world variations, but not all field conditions, vendor interfaces and wavelength-sharing impairments were included. The cost model is deployment-oriented rather than a procurement quotation, and the transaction-risk score was used as a cryptographic allocation input instead of a substitute for fraud detection. Continuity mode also maintains service by switching the source composition of the key, and some high-assurance settlement classes have chosen not to allow this.

Future work will validate the protocol among various QKD vendors and production payment switches, add device-independent attestation evidence, and formalize the state machine in a machine-checked verification framework. Future work will explore multi-path secret sharing without trusted intermediate nodes, privacy-preserving coordination of risk classes, and online control that learns demand without disturbing the reserved pool. An extended operating period would also clarify the maintenance behaviour and certification costs, as well as the economic threshold at which shared quantum infrastructure is economically feasible for smaller financial institutions.

Author Contributions

Rasmus Sepp contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision. Daria Eelmaa contributes to data

collection, draft preparation, manuscript editing. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Ahn, J., Kwon, H. Y., Ahn, B., Park, K., Kim, T., Lee, M. K., Kim, J., & Chung, J. (2022). Toward Quantum Secured Distributed Energy Resources: Adoption of Post-Quantum Cryptography (PQC) and Quantum Key Distribution (QKD). Energies, 15(3), 714. https://doi.org/10.3390/en15030714

[2] Yang, Y. H., Li, P. Y., Ma, S. Z., Qian, X. C., Zhang, K. Y., Wang, L. J., Zhang, W. L., Zhou, F., Tang, S. B., Wang, J. Y., Yu, Y., Zhang, Q., & Pan, J. W. (2021). All optical metropolitan quantum key distribution network with post-quantum cryptography authentication. Optics Express, 29(16), 25859. https://doi.org/10.1364/oe.432944

[3] Wang, L. J., Zhang, K. Y., Wang, J. Y., Cheng, J., Yang, Y. H., Tang, S. B., Yan, D., Tang, Y. L., Liu, Z., Yu, Y., Zhang, Q., & Pan, J. W. (2021). Experimental authentication of quantum key distribution with post-quantum cryptography. npj Quantum Information, 7(1), 67. https://doi.org/10.1038/s41534-021-00400-7

[4] Richter, M., Bertram, M., Seidensticker, J., & Tschache, A. (2022). A Mathematical Perspective on Post-Quantum Cryptography. Mathematics, 10(15), 2579. https://doi.org/10.3390/math10152579

[5] Qian, X., Zhang, C., Yuan, H., Zhou, X., Li, J., & Wang, Q. (2022). Passive Light Source Monitoring for Sending or Not Sending Twin-Field Quantum Key Distribution. Entropy, 24(5), 592. https://doi.org/10.3390/e24050592

[6] Aydin, F., Aysu, A., Tiwari, M., Gerstlauer, A., & Orshansky, M. (2021). Horizontal Side-Channel Vulnerabilities of Post-Quantum Key Exchange and Encapsulation Protocols. ACM Transactions on Embedded Computing Systems, 20(6), 1-22. https://doi.org/10.1145/3476799

[7] Wei, K., Chen, Z., Li, Z., Zheng, B., & Zhang, Z. (2022). Asymmetric reference-frame-independent measurement-device-independent quantum key distribution. Journal of the Optical Society of America B, 39(11), 3041. https://doi.org/10.1364/josab.468487

[8] Bebrov, G. (2022). More optimal relativistic quantum key distribution. Scientific Reports, 12(1), 15377. https://doi.org/10.1038/s41598-022-15247-x

[9] Yao, J., Matusiewicz, K., & Zimmer, V. (2022). Post Quantum Design in SPDM for Device Authentication and Key Establishment. Cryptography, 6(4), 48. https://doi.org/10.3390/cryptography6040048

[10] Cao, X. Y., Gu, J., Lu, Y. S., Yin, H. L., & Chen, Z. B. (2021). Coherent one-way quantum conference key agreement based on twin field. New Journal of Physics, 23(4), 043002. https://doi.org/10.1088/1367-2630/abef98

[11] Tan, E. Y. Z., Schwonnek, R., Goh, K. T., Primaatmaja, I. W., & Lim, C. C. W. (2021). Computing secure key rates for quantum cryptography with untrusted devices. npj Quantum Information, 7(1), 158. https://doi.org/10.1038/s41534-021-00494-z

[12] Chamola, V., Jolfaei, A., Chanana, V., Parashari, P., & Hassija, V. (2021). Information security in the post quantum era for 5G and beyond networks: Threats to existing cryptography, and post-quantum cryptography. Computer Communications, 176, 99-118. https://doi.org/10.1016/j.comcom.2021.05.019

[13] Lu, Y. F., Jiang, M. S., Wang, Y., Zhang, X. X., Liu, F., Zhou, C., Li, H. W., Tang, S. B., Wang, J. Y., & Bao, W. S. (2021). Practical Analysis of Sending or Not-Sending Twin-Field Quantum Key Distribution with Frequency Side Channels. Applied Sciences, 11(20), 9560. https://doi.org/10.3390/app11209560

[14] Costa, V. C. L. R. D., López, J., & Ribeiro, M. S. V. (2022). A System-on-a-Chip Implementation of a Post-Quantum Cryptography Scheme for Smart Meter Data Communications. Sensors, 22(19), 7214. https://doi.org/10.3390/s22197214

[15] Tian, Y., Wang, P., Liu, J., Du, S., Liu, W., Lu, Z., Wang, X., & Li, Y. (2022). Experimental demonstration of continuous-variable measurement-device-independent quantum key distribution over optical fiber. Optica, 9(5), 492. https://doi.org/10.1364/optica.450573

[16] Jiang, X. L., Wang, Y., Lu, Y. F., Li, J. J., Zhou, C., & Bao, W. S. (2022). Security Analysis of Sending or Not-Sending Twin-Field Quantum Key Distribution with Weak Randomness. Entropy, 24(10), 1339. https://doi.org/10.3390/e24101339

[17] Hu, X. L., Jiang, C., Yu, Z. W., & Wang, X. B. (2022). Sending-or-not-sending twin field quantum key distribution with imperfect vacuum sources. New Journal of Physics, 24(6), 063014. https://doi.org/10.1088/1367-2630/ac7347

[18] Wang, W., Tamaki, K., & Curty, M. (2021). Measurement-device-independent quantum key distribution with leaky sources. Scientific Reports, 11(1), 1678. https://doi.org/10.1038/s41598-021-81003-2

[19] Zhu, J. R., Zhang, C. M., & Wang, Q. (2021). Reference-frame-independent measurement-device-independent quantum key distribution with imperfect sources. Journal of Physics B: Atomic, Molecular and Optical Physics, 54(14), 145501. https://doi.org/10.1088/1361-6455/ac0c05

[20] Lu, Y. F., Wang, Y., Jiang, M. S., Zhang, X. X., Liu, F., Li, H. W., Zhou, C., Tang, S. B., Wang, J. Y., & Bao, W. S. (2021). Sending or Not-Sending Twin-Field Quantum Key Distribution with Flawed and Leaky Sources. Entropy, 23(9), 1103. https://doi.org/10.3390/e23091103

[21] Djordjevic, I. B. (2022). Physical-Layer Security, Quantum Key Distribution, and Post-Quantum Cryptography. Entropy, 24(7), 935. https://doi.org/10.3390/e24070935

[22] Hu, X. L., Jiang, C., Yu, Z. W., & Wang, X. B. (2022). Universal approach to sending-or-not-sending twin field quantum key distribution. Quantum Science and Technology, 7(4), 045031. https://doi.org/10.1088/2058-9565/ac8e90

[23] Ye, W., Guo, Y., Zhang, H., Zhong, H., Mao, Y., & Hu, L. (2021). Enhancing discrete-modulated continuous-variable measurement-device-independent quantum key distribution via quantum catalysis. Journal of Physics B: Atomic, Molecular and Optical Physics, 54(4), 045501. https://doi.org/10.1088/1361-6455/abdac9

[24] Chen, Y. P., Liu, J. Y., Sun, M. S., Zhou, X. X., Zhang, C. H., Li, J., & Wang, Q. (2021). Experimental measurement-device-independent quantum key distribution with the double-scanning method. Optics Letters, 46(15), 3729. https://doi.org/10.1364/ol.431061

[25] Wang, X., Liu, W., Wu, T., Guo, C., Zhang, Y., Zhao, S., & Dong, C. (2021). Free Space Measurement Device Independent Quantum Key Distribution with Modulating Retro-Reflectors under Correlated Turbulent Channel. Entropy, 23(10), 1299. https://doi.org/10.3390/e23101299

[26] Woodward, R. I., Lo, Y. S., Pittaluga, M., Minder, M., Paraïso, T. K., Lucamarini, M., Yuan, Z. L., & Shields, A. J. (2021). Gigahertz measurement-device-independent quantum key distribution using directly modulated lasers. npj Quantum Information, 7(1), 58. https://doi.org/10.1038/s41534-021-00394-2

[27] Sekga, C., & Mafu, M. (2021). Reference frame independent twin field quantum key distribution with source flaws. Journal of Physics Communications, 5(4), 045008. https://doi.org/10.1088/2399-6528/abf472

[28] Currás-Lorenzo, G., Navarrete, L., Azuma, K., Kato, G., Curty, M., & Razavi, M. (2021). Tight finite-key security for twin-field quantum key distribution. npj Quantum Information, 7(1), 22. https://doi.org/10.1038/s41534-020-00345-3

[29] Sekatski, P., Bancal, J. D., Valcarce, X., Tan, E. Y. Z., Renner, R., & Sangouard, N. (2021). Device-independent quantum key distribution from generalized CHSH inequalities. Quantum, 5, 444. https://doi.org/10.22331/q-2021-04-26-444

[30] Zhang, X. X., Wang, Y., Jiang, M. S., Zhou, C., Lu, Y. F., & Bao, W. S. (2021). Finite-key analysis of asymmetric phase-matching quantum key distribution with unstable sources. Journal of the Optical Society of America B, 38(3), 724. https://doi.org/10.1364/josab.415022

[31] Zhu, Y., & Zhang, C. M. (2021). Improved analysis of measurement-device-independent quantum key distribution with non-phase-randomized coherent states. Optics Express, 29(19), 30168. https://doi.org/10.1364/oe.435687

[32] Li, B. H., Xie, Y. M., Li, Z., Weng, C. X., Li, C. L., Yin, H. L., & Chen, Z. B. (2021). Long-distance twin-field quantum key distribution with entangled sources. Optics Letters, 46(22), 5529. https://doi.org/10.1364/ol.443099

[33] Clivati, C., Meda, A., Donadello, S., Virzì, S., Genovese, M., Levi, F., Mura, A., Pittaluga, M., Yuan, Z., Shields, A. J., Lucamarini, M., Degiovanni, I. P., & Calonico, D. (2022). Coherent phase transfer for real-world twin-field quantum key distribution. Nature Communications, 13(1), 157. https://doi.org/10.1038/s41467-021-27808-1

[34] Li, H. W., Zhang, C. M., Jiang, M. S., & Cai, Q. Y. (2022). Improving the performance of practical decoy-state quantum key distribution with advantage distillation technology. Communications Physics, 5(1), 53. https://doi.org/10.1038/s42005-022-00831-4

[35] Jiang, C., Hu, X. L., Yu, Z. W., & Wang, X. B. (2022). Measurement-device-independent quantum key distribution protocol with phase post-selection. Photonics Research, 10(7), 1703. https://doi.org/10.1364/prj.445617

[36] Ding, H. J., Zhou, X. Y., Zhang, C. H., Li, J., & Wang, Q. (2022). Measurement-device-independent quantum key distribution with insecure sources. Optics Letters, 47(3), 665. https://doi.org/10.1364/ol.447234