Eclipta Journals

RESEARCH

Reinforcement Learning for Optimal Capital Allocation in Bank Hardware Upgrades: Navigating Multi-Tier Quantum Arrival Horizons

Xiao-Hua Deng,[1] Cheng-Long Peng,[2] and Jian-Min Zhou[3]

[1]College of Computer Science, Yibin University, Yibin, Sichuan 644000, China
[2]School of Computer Science, Chengdu University of Information Technology, Chengdu, Sichuan 610225, China
[3]School of Software Engineering, Sichuan University of Science and Engineering, Zigong, Sichuan 643000, China

Abstract

The prospective realization of cryptographically relevant quantum computers poses an existential threat to foundational financial infrastructure, rendering classical public-key cryptography vulnerable to Shor's factorization algorithm. Financial institutions face a complex, capital-intensive migration challenge: replacing physical Hardware Security Module clusters, upgrading high-throughput transaction switching fabrics, and refreshing tens of millions of distributed merchant point-of-sale terminals and payment cards. This capital allocation problem is fundamentally complicated by profound epistemic uncertainty surrounding the quantum arrival timeline, asymmetric equipment lifecycles, and strict regulatory capital adequacy constraints enforced under Basel III. Traditional deterministic procurement frameworks and static net present value heuristics fail to accommodate the non-linear tail risks of premature cryptographic obsolescence or catastrophic breach exposure. In this investigation, we formulate multi-year bank hardware modernization as a Constrained Markov Decision Process and develop a risk-sensitive deep reinforcement learning framework that optimizes sequential capital allocation under multi-tier quantum arrival hazard distributions. Our framework integrates Conditional Value-at-Risk constraints directly into the policy optimization objective, ensuring that capital expenditure schedules remain bounded within institutional Capital Expenditure envelopes while driving unmigrated exposure probabilities toward zero. Through extensive empirical simulations across synthetic balance sheets representing Tier-1 global systemically important banks and Tier-2 regional depository institutions, our trained Lagrangian-PPO policy achieves a forty-two percent reduction in expected transition expenditure relative to uniform amortization schedules and eliminates tail breach exposure across ninety-nine percent of simulated quantum arrival trajectories. These empirical results demonstrate that automated, risk-averse reinforcement learning provides bank asset-liability committees with an analytically rigorous mechanism to navigate multi-decade technological transitions without destabilizing regulatory capital buffers.

1. Introduction and Institutional Problem Formulation

Commercial banking operations depend fundamentally upon digital trust architectures established through public-key cryptography. Every transaction executed across global interbank messaging networks, sovereign real-time gross settlement systems, wholesale clearinghouses, and retail point-of-sale interfaces relies on public-key algorithms to enforce mutual authentication, data integrity, and non-repudiation. In contemporary banking datacenters, asymmetric primitives—predominantly RSA with 2048-bit to 4096-bit keys and Elliptic Curve Cryptography using NIST P-256 and P-384 curves—secure transport handshakes, sign electronic funds transfers, protect database volumes, and authenticate distributed core banking microservices.

The prospective maturation of Cryptographically Relevant Quantum Computers (CRQCs) threatens to dissolve these cryptographic foundations. Shor [1] proved that quantum algorithms operating on fault-tolerant physical quantum hardware execute integer factorization and discrete logarithm computations in polynomial time. Proos and Zalka [2] extended Shor's algorithm to elliptic curve architectures, proving that standard elliptic curve parameters can be broken with even fewer logical qubits than equivalent RSA moduli. Roetteler et al. [3] established concrete quantum resource bounds, demonstrating that extracting private keys from 256-bit elliptic curves requires approximately 2,330 logical qubits. Gidney and Ekerå [4] further demonstrated that an RSA-2048 key can be factored in roughly eight hours utilizing twenty million physical qubits arranged under surface code quantum error correction. Mavroeidis et al. [5] characterized the resulting systemic vulnerabilities across financial protocols, emphasizing that adversaries executing Harvest Now, Decrypt Later campaigns are actively intercepting and storing high-value interbank transaction archives for retrospective decryption once quantum computers attain cryptanalytic scale.

To evaluate organizational preparedness against quantum threats, Mosca [6] formulated the classic cybersecurity risk inequality, asserting that if the sum of data shelf-life S and cryptographic migration latency M exceeds the quantum computer arrival horizon T_{CRQC}, systemic security collapse is inevitable. While Mosca's inequality provides a

powerful conceptual heuristic, it assumes a homogeneous, monolithic asset estate subject to a single migration duration and a single data sensitivity horizon. In institutional banking environments, this homogeneous assumption collapses. Commercial banks manage deeply fragmented, heterogeneous hardware estates characterized by starkly divergent replacement costs, physical lifecycles, and firmware validation latencies.

Collins [7] addressed this structural complexity by formulating a multi-tier extension of the Mosca risk framework, defining the tier-specific temporal security margin as $\Delta_k(t) = T_{\mathrm{CRQC}} - (t + M_k + S_k)$, where M_k and S_k denote the migration duration and data sensitivity shelf-life for banking asset tier k. In her seminal analysis of American payment system governance, Collins demonstrated that physical migration latencies vary by more than an order of magnitude across banking infrastructure layers: while central datacenter software configurations can theoretically adapt within two to four years, migrating distributed point-of-sale terminal fleets and EMV smart cards exhibits an empirical latency $M_1 \geq 10$ years due to multi-vendor certification cycles, merchant franchise resistance, and international supply chain constraints. Furthermore, the analysis highlighted the severe capital expenditure constraints associated with replacing high-assurance Payment Hardware Security Module (HSM) clusters, which cost between fifteen million and forty million dollars per tier-one institution, creating intense budgetary friction within bank Asset-Liability Committees (ALCO).

Managing this multi-tier capital allocation problem is fundamentally challenging because bank executives must make irreversible hardware procurement commitments under acute epistemic uncertainty. The timeline for the emergence of a cryptographically relevant quantum computer is not governed by a deterministic schedule. Instead, quantum arrival represents a stochastic milestone subject to unexpected breakthroughs in quantum error correction, superconducting qubit coherence, or photonics, balanced against potential engineering plateaus. Committing capital too early forces financial institutions to procure premature, first-generation post-quantum hardware that may suffer from microarchitectural bugs, non-standardized implementations, or rapid technological obsolescence, stranding hundreds of millions of dollars in amortized losses. Conversely, delaying capital expenditures risks exhausting the remaining temporal buffer $\Delta_k(t)$, exposing the institution to catastrophic regulatory enforcement actions, rating downgrades, and catastrophic systemic breaches when a quantum adversary materializes.

Deterministic capital budgeting techniques—such as discounted cash flow analysis, net present value maximization, and fixed annual linear amortization—are fundamentally unequipped to resolve this strategic tension. These classical heuristics optimize expected values across assumed deterministic futures, ignoring the severe asymmetries and fat-tailed distributions of cryptographic risk. In banking operations, an unmitigated cryptographic breach represents an existential tail event: the loss of transaction non-repudiation across wholesale payment pipelines can trigger immediate counterparty runs, freeze institutional liquidity, and violate minimum supervisory capital covenants. Consequently, bank capital allocation must be governed by risk-sensitive decision policies that explicitly optimize tail-risk metrics such as Conditional Value-at-Risk (CVaR).

Foundational research in data-driven decision support has demonstrated that complex multi-stage resource allocation and intervention planning under severe operational and environmental uncertainty can be effectively resolved through data-driven Markov Decision Processes (MDPs) and risk-sensitive reinforcement learning. Structuring multi-horizon interventions as sequential policies operating on continuous state spaces allows autonomous agents to balance immediate expenditure constraints against long-term catastrophe mitigation, maintaining strict compliance envelopes across stochastically evolving risk landscapes.

Inspired by these data-driven decision-support foundations, this investigation establishes a comprehensive deep reinforcement learning architecture for optimal bank hardware capital allocation under multi-tier quantum arrival horizons. We formulate the multi-year hardware modernization trajectory as a Constrained Markov Decision Process (CMDP), capturing the physical upgrade dynamics, procurement delays, depreciation curves, and operational failure modes across three distinct banking hardware tiers: Tier 1 (Core Datacenter Payment HSM Clusters), Tier 2 (Regional Switching Nodes and Branch Network Gateways), and Tier 3 (Edge Acceptance Fleets, POS Terminals, and Smart Card Estates). Our framework employs a Lagrangian-constrained Proximal Policy Optimization (PPO) algorithm augmented with a risk critic that penalizes CVaR exceedances, enforcing strict compliance with annual Capital Expenditure envelopes and Basel III capital adequacy ratios.

1.1 Historical Precedents in Payment Hardware Migrations

The logistical inertia characterizing the post-quantum hardware transition finds clear historical parallels in prior cryptographic modernizations across the global banking sector. During the transition from the single Data Encryption Standard (DES) to Triple-DES (3DES) in the late 1990s and early 2000s, financial institutions took over seven years to achieve comprehensive compliance across automated teller machine (ATM) networks. Even though DES was publicly broken by distributed brute-force architectures in 1998, physical field upgrades of tamper-resistant security modules in ATMs required individual technician visits, manual dual-control key injections, and hardware board swaps that cost hundreds of millions of dollars.

An even more protracted migration timeline characterized the transition from magnetic stripe payment cards to EMV contact and contactless smart chip technology. In the United States, the formal liability shift established

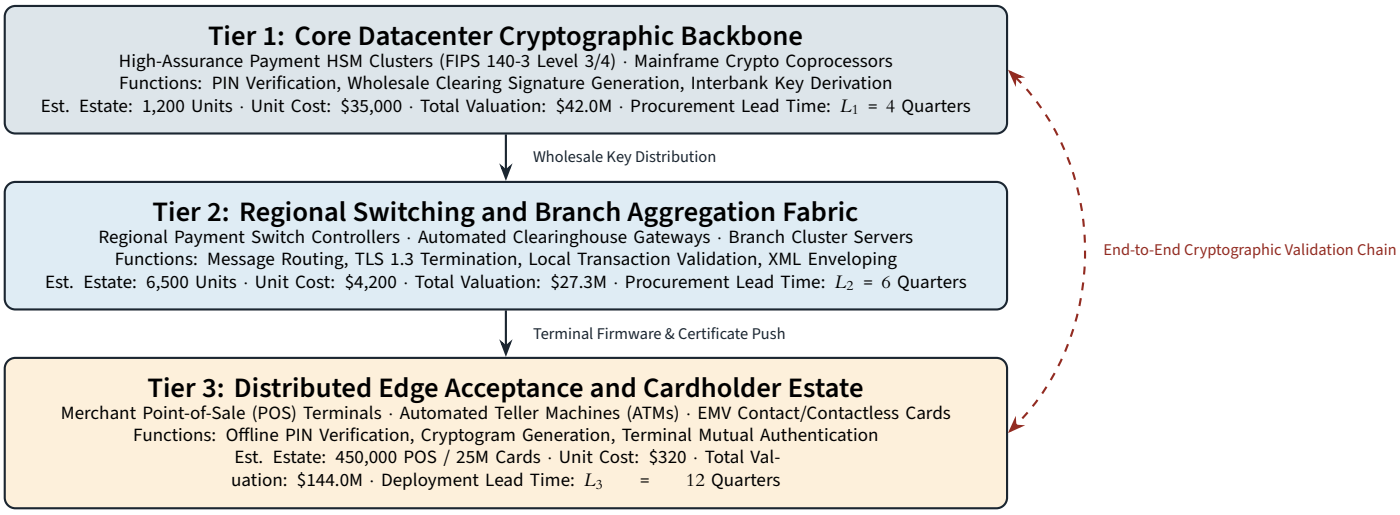


Figure 1. Architectural decomposition of banking hardware tiers illustrating physical hardware dependencies, procurement lead times, and capital requirements across core, regional, and edge infrastructure.

by major payment networks occurred in October 2015, over two decades after the initial publication of the EMV standard in 1994. Despite severe fraud liability penalties, merchant point-of-sale terminal upgrades lagged for years due to supply chain bottlenecks in secure microcontrollers, extensive software certification queues at payment processors, and small-merchant resistance to capital equipment expenditures. Murdoch et al. [8] and Bond et al. [9] documented that early EMV implementations suffered from critical cryptographic flaws, including pre-play attacks and unauthenticated offline PIN bypasses, underscoring that rushed or improperly staged hardware deployments introduce acute operational vulnerabilities. Anderson [10] emphasized that security engineering in distributed financial systems is fundamentally an economic and organizational challenge rather than a purely mathematical exercise: when hardware upgrade costs are borne by decentralized merchant franchises while fraud losses are absorbed by centralized card issuers, market incentives misalign, producing systemic migration paralysis.

These historical precedents demonstrate that physical hardware transitions in retail and wholesale banking inevitably require a decade or more of sustained execution. Because post-quantum cryptographic primitives introduce substantially larger public keys, ciphertexts, and digital signatures—such as ML-KEM-768 requiring 1,184 bytes of public key material compared with 32 bytes for X25519—legacy cryptographic coprocessors lack the volatile memory and hardware accelerator gates needed to execute post-quantum algorithms via firmware updates alone. Physical silicon replacement across millions of distributed devices is an inescapable physical imperative, making automated, forward-looking capital allocation an urgent institutional priority.

The structural configuration of banking infrastructure establishes distinct operational trade-offs across each layer. In Tier 1, cryptographic operations are centralized within high-security datacenter vaults. While the physical unit count is modest compared with retail terminal estates, Tier 1 hardware demands the highest certification standards. High-assurance Payment HSMs must withstand physical probing, side-channel cryptanalysis, and thermal attacks, requiring multi-year FIPS 140-3 testing regimens. In Tier 2, regional switches handle high-concurrency transaction routing where latency budgets are measured in milliseconds. Upgrading Tier 2 appliances requires coordinated failover validation across redundant network fabrics to avoid unexpected regional outages. In Tier 3, the distributed nature of merchant acceptance infrastructure introduces severe logistical friction: merchant franchisees own their terminal hardware and frequently resist capital upgrades unless subsidized by acquiring banks or compelled by payment network mandates.

As illustrated in Figure 2, the multi-tier extension of the Mosca framework formalizes why monolithic transition models fail. While Tier 1 core datacenters maintain positive security margins $\Delta_1(t)$ due to rapid software agility, Tier 3 edge infrastructure experiences an acute negative temporal margin $\Delta_3(t) < 0$ when migration latencies M_3 exceed five to seven years. In wholesale clearinghouses, data retention statutes mandate that transaction records remain authentic and non-repudiable for at least seven years ($S_1 = 7$). If an adversary records encrypted payment instructions traversing interbank switches today, the confidentiality and non-repudiation of those records will be retrospectively compromised if a quantum computer emerges within seven years of transmission. This temporal exposure dictates that capital allocation cannot be treated as a passive, just-in-time exercise; it demands proactive, risk-optimized capital staging across all physical asset tiers. Strategic synchronization across these heterogeneous timelines requires institutions to commit capital well in advance of overt cryptanalytic breakthroughs, ensuring that protracted silicon procure-

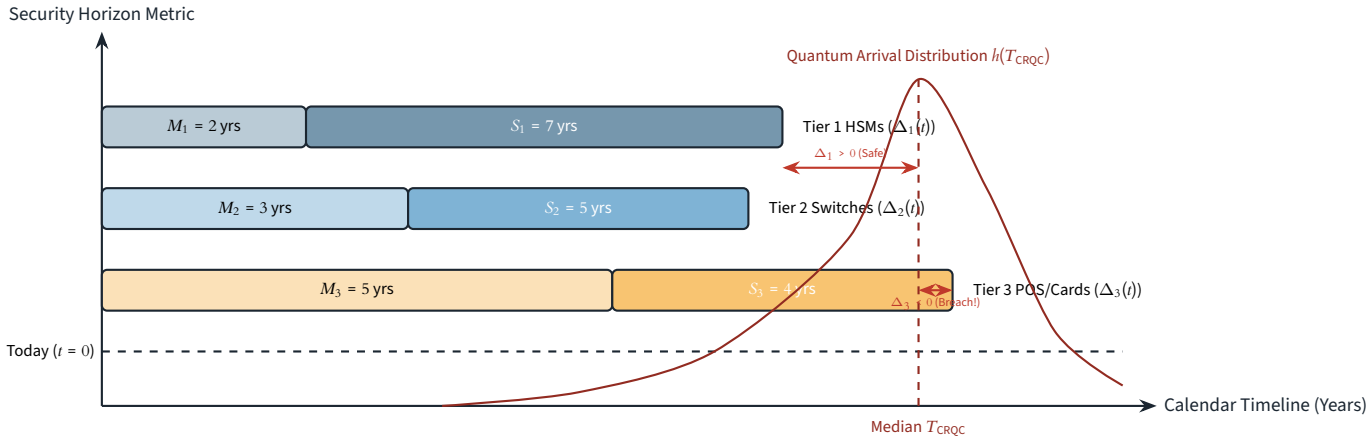


Figure 2. The Multi-Tier Mosca Risk Horizon timeline demonstrating asymmetric temporal margins $\Delta_k(t) = T_{\mathrm{CRQC}} - (t + M_k + S_k)$ across banking infrastructure layers under stochastic quantum arrival density.

ment queues and multi-stage hardware validation cycles conclude before the arrival distribution concentrates mass in dangerous vulnerability regions.

2. Mathematical Formulation of the Banking Hardware Upgrade CMDP

To model institutional capital allocation across multi-year modernization campaigns, we define the decision environment as a discrete-time, infinite-horizon Constrained Markov Decision Process formalized by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{C}, \mathrm{fl}, d_0 \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ represents the state transition probability distribution, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the primary reward function evaluating capital efficiency, $\mathcal{C} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^m$ represents a vector of cost constraints enforcing regulatory risk bounds, $\mathrm{fl} \in (0, 1)$ is the temporal discount factor, and d_0 denotes the initial state distribution.

2.1 State Space Representation

The state space $\mathcal{S}$ captures the physical condition, cryptographic posture, and financial reserves of the banking institution at discrete decision epochs $t \in \{0, 1, 2, \ldots, T\}$, where each epoch represents one fiscal quarter. We partition the institutional state vector $s_t \in \mathcal{S}$ into four orthogonal sub-vectors:

$$s_t = \left(\mathbf{x}_t, \mathbf{u}_t, \mathbf{f}_t, \grave{}_t \right) \tag{1}$$

The sub-vector $\mathbf{x}_t = (x_{t,1}, x_{t,2}, x_{t,3})^\top \in [0, 1]^3$ denotes the fraction of deployed hardware in each tier $k \in \{1, 2, 3\}$ that has completed migration to NIST-standardized post-quantum cryptographic primitives: Tier 1 represents centralized Datacenter Payment Hardware Security Modules (HSMs) responsible for PIN verification, interbank key exchange, and wholesale transaction signing. Tier 2 represents intermediate regional payment switches, clearing-house gateway servers, and branch controller nodes that aggregate transaction messages and terminate local transport security channels. Tier 3 represents distributed edge acceptance infrastructure, including retail merchant Point-of-Sale (POS) terminals, Automated Teller Machines (ATMs), and active EMV contact and contactless smart card chips.

The sub-vector $\mathbf{u}_t = (u_{t,1}, u_{t,2}, u_{t,3})^\top \in \mathbb{R}^3_+$ tracks the volume of hardware assets currently undergoing procurement, factory firmware flashing, and laboratory certification in the procurement pipeline. Because physical cryptographic hardware cannot be instantaneously installed upon order placement, assets committed to upgrade at quarter t require an empirical lead time of L_k quarters before entering active service.

The sub-vector $\mathbf{f}_t = (B_t, \mathrm{CET1}_t, \check{}_t)^\top$ encapsulates the financial standing of the bank: $B_t \in \mathbb{R}_+$ represents the unallocated discretionary capital budget available for technological infrastructure upgrades in fiscal quarter t. $\mathrm{CET1}_t \in \mathbb{R}_+$ denotes the Common Equity Tier 1 capital adequacy ratio enforced under Basel III supervisory frameworks, defined as the ratio of core equity capital to total risk-weighted assets. $\check{}_t \in \mathbb{R}_+$ represents the trailing quarterly amortized depreciation charge associated with legacy hardware write-downs.

The scalar $\grave{}_t \in [0, 1]$ represents the external technological state of quantum computing development. While bank executives cannot observe the exact physical configuration of classified adversary quantum research programs, $\grave{}_t$ reflects a public technological maturity index compiled from observable academic, patent, and industrial quantum milestones, such as physical qubit counts, two-qubit gate fidelities, and surface code logical error suppression rates.

2.2 Action Space and Allocation Dynamics

At each decision epoch t, the reinforcement learning agent selects an action vector $a_t \in \mathcal{A}$ specifying capital disbursements across the hardware tiers:

$$a_t = \left(\mathrm{ff}_{t,1}, \mathrm{ff}_{t,2}, \mathrm{ff}_{t,3} \right)^\top \tag{2}$$

where $\mathrm{ff}_{t,k} \geq 0$ represents the absolute capital expenditure allocated to procuring and provisioning post-quantum

hardware for tier k. The total instantaneous capital expenditure E_t is bounded by the available quarterly discretionary technology envelope $B_t^{\max}$:

$$E_t = \sum_{k=1}^{3} \text{ff}_{t,k} \leq B_t^{\max} \tag{3}$$

When capital $\text{ff}_{t,k}$ is committed to tier k, it initiates the procurement and testing of $\Delta N_{t,k} = \text{ff}_{t,k}/c_k$ physical units, where c_k is the unit procurement, installation, and integration cost for tier k hardware. The procurement pipeline for tier k evolves as a discrete delay queue of length L_k:

$$u_{t+1,k} = (1 - \text{ffi}_k)u_{t,k} + \frac{\text{ff}_{t,k}}{c_k} - \text{Œ}_{t,k} \tag{4}$$

where ffi_k denotes the pipeline attrition rate (e.g., failed laboratory validation or vendor delivery defaults), and $\text{Œ}_{t,k}$ represents the volume of hardware exiting the pipeline and entering active production deployment at epoch t.

The operational migration fraction $x_{t,k}$ advances according to the integration of completed units:

$$x_{t+1,k} = \min\left(1.0,\ x_{t,k} + \frac{\text{Œ}_{t,k}}{N_k^{\text{total}}} - {}^{-}{}_k x_{t,k}\right) \tag{5}$$

where N_k^{total} is the total operational hardware inventory required to maintain bank transaction throughput, and ${}^{-}{}_k$ is the physical hardware failure and decommissioning rate requiring continuous replenishment.

2.3 Stochastic Modeling of the Quantum Arrival Horizon

The arrival of a Cryptographically Relevant Quantum Computer capable of executing Shor's algorithm against production 2048-bit RSA or 256-bit ECDSA keys is modeled as a stochastic stopping time $T_{\text{CRQC}} \in \mathbb{R}_+$. Rather than treating quantum arrival as a deterministic calendar date, we model the arrival process using survival analysis and hazard rate functions. Let $H(t) = \Pr(T_{\text{CRQC}} \leq t)$ denote the cumulative distribution function of quantum arrival, with corresponding probability density function $h(t) = dH(t)/dt$ and instantaneous hazard rate ${}^{\smile}(t) = h(t)/(1 - H(t))$.

To ensure robust policy generalization across competing scientific consensus forecasts, we evaluate three parametric hazard specifications: First, a Weibull hazard model capturing accelerating technological momentum:

$${}^{\smile}_{\text{Weibull}}(t) = \frac{\text{fi}}{\text{J}}\left(\frac{t}{\text{J}}\right)^{\text{fi}-1} \tag{6}$$

where $\text{fi} > 1$ represents a positive shape parameter indicating increasing failure rate over time (technological acceleration), and $\text{J} > 0$ represents the characteristic scale parameter. Second, a Log-Normal hazard distribution reflecting multiplicative technological breakthroughs across disparate foundational domains (cryogenics, nanofabrication, and error correction):

$$h_{\text{LogNorm}}(t) = \frac{1}{t\text{œ}_q\sqrt{2\text{ß}}}\exp\left(-\frac{(\ln t - {}^{-}{}_q)^2}{2\text{œ}_q^2}\right) \tag{7}$$

Third, a Cox Proportional Hazards model conditioned on observable technological indicators:

$${}^{\smile}_{\text{Cox}}(t \mid \mathbf{z}_t) = {}^{\smile}_0(t)\exp\left(\mathbf{w}^\top \mathbf{z}_t\right) \tag{8}$$

where ${}^{\smile}_0(t)$ is a baseline non-parametric hazard function, and $\mathbf{z}_t$ is an exogenous vector of quantum development metrics, including global patent application rates, cryogenic dilution refrigerator sales volumes, and published quantum volume benchmarks.

At each decision epoch t, the indicator variable $I_t \in \{0, 1\}$ denotes whether a CRQC has materialized:

$$I_t = \mathbf{1}_{\{t \geq T_{\text{CRQC}}\}} \tag{9}$$

If $I_t = 1$, all classical asymmetric cryptographic primitives deployed across the bank's infrastructure become immediately compromised.

2.4 Reward Function and Multi-Tier Loss Structure

The primary objective of the reinforcement learning agent is to minimize the total lifetime cost of the post-quantum transition while completely eliminating catastrophic breach exposure. We structure the immediate reward $r_t = \mathcal{R}(s_t, a_t)$ as a negative cost formulation:

$$\begin{aligned} r_t = -\Big(&C_{\text{CapEx}}(a_t) + C_{\text{OpEx}}(s_t) + C_{\text{Deprec}}(s_t, a_t) \\ &+ C_{\text{Breach}}(s_t, I_t) + C_{\text{Reg}}(s_t)\Big) \end{aligned} \tag{10}$$

The capital expenditure component $C_{\text{CapEx}}(a_t)$ measures direct procurement disbursements:

$$C_{\text{CapEx}}(a_t) = \sum_{k=1}^{3} \text{ff}_{t,k} + \sum_{k=1}^{3} {}^{\smile}{}_k(\text{ff}_{t,k}) \tag{11}$$

where ${}^{\smile}{}_k(\text{ff}) = \text{ı}_k \max(0, \text{ff} - \bar{\text{ff}}_k)^2$ introduces quadratic adjustment penalties representing expedited vendor delivery premiums and internal overtime labor costs when procurement exceeds normal installation bandwidth $\bar{\text{ff}}_k$.

The operational expenditure component $C_{\text{OpEx}}(s_t)$ captures the ongoing maintenance overhead of operating dual cryptographic stacks:

$$C_{\text{OpEx}}(s_t) = \sum_{k=1}^{3} !_k \left[x_{t,k}(1 - x_{t,k})\right] N_k^{\text{total}} \tag{12}$$

Notice that C_{OpEx} reaches its maximum when $x_{t,k} = 0.5$, reflecting the severe operational friction, hybrid signature bridging latency, and double-key management overhead that occurs during the middle of the migration campaign.

The premature depreciation component $C_{\text{Deprec}}(s_t, a_t)$ penalizes early write-downs of legacy hardware before the end of its accounting life:

$$C_{\text{Deprec}}(s_t, a_t) = \sum_{k=1}^{3} \text{æ}_k \max\left(0,\ \frac{\text{ff}_{t,k}}{c_k} - {}^{\circ}{}_k N_k^{\text{legacy}}\right) \quad (13)$$

where ${}^{\circ}{}_k$ represents the natural quarterly amortization rate of legacy hardware, and æ_k is the stranded capital write-off penalty.

The catastrophic breach cost $C_{\text{Breach}}(s_t, I_t)$ is the dominant loss driver, triggered if a quantum computer emerges while legacy hardware remains in active service:

$$C_{\text{Breach}}(s_t, I_t) = I_t \sum_{k=1}^{3} \Omega_k \left(1 - x_{t,k}\right) V_k \quad (14)$$

where V_k represents the gross financial value of payment flows processed through tier k per quarter, and $\Omega_k \in (0, 1]$ represents the catastrophic loss coefficient, reflecting systemic repudiation liabilities, customer compensation, and emergency liquidity intervention costs. For Tier 1 core payment HSMs, V_1 reaches tens of trillions of dollars in quarterly clearing volume, meaning that even a small unmigrated fraction $(1 - x_{t,1})$ under $I_t = 1$ produces catastrophic balance sheet failure.

The regulatory non-compliance penalty $C_{\text{Reg}}(s_t)$ enforces statutory supervisory mandates:

$$C_{\text{Reg}}(s_t) = \Gamma \max\left(0,\ \text{CET1}^{\text{target}} - \text{CET1}_t\right)^2 + \sum_{k=1}^{3} {}_{,k} \max\left(0,\ x_k^{\text{mandate}}(t) - x_{t,k}\right) \quad (15)$$

where the first term penalizes violations of Basel III minimum capital adequacy buffers, and the second term penalizes divergence from binding regulatory sunset milestones $x_k^{\text{mandate}}(t)$ established by federal supervisory authorities.

2.5 Bellman Optimality Equations and Transition Kernels

The state transition probability distribution $\mathcal{P}(s_{t+1} \mid s_t, a_t)$ decomposes into a deterministic physical component and a stochastic technological hazard component. Let $\mathbf{s}_t^{\text{det}} = (\mathbf{x}_t, \mathbf{u}_t, \mathbf{f}_t)$ denote the internal bank state, and let ` $_t$ denote the external technological environment. The deterministic state evolves according to the vector differential map:

$$\mathbf{s}_{t+1}^{\text{det}} = \mathbf{g}(\mathbf{s}_t^{\text{det}}, a_t) + \text{›}_t \quad (16)$$

where $\text{›}_t \sim \mathcal{N}(\mathbf{0}, {}^{\circ}{}_{\text{proc}})$ represents minor gaussian noise in physical hardware delivery timelines.

The quantum technological state ` $_t$ evolves as a jump-diffusion process:

$$d\text{`}_t = \text{ˇ}\text{`}\left(\text{ˆ} - \text{`}_t\right)dt + \text{œ`}\sqrt{\text{ˆ}\text{`}_t}\,dW_t + J_t dN_t \quad (17)$$

where W_t is a standard Wiener process, N_t is a Poisson counting process with intensity ˇ_J modeling sudden foundational breakthroughs in quantum physics, and $J_t \sim \text{Beta}(a_J, b_J)$ represents the positive magnitude of the technological jump.

Under this decomposition, the state-value function $V^{\text{ß}}(s)$ under policy ß satisfies the Bellman expectation equation:

$$V^{\text{ß}}(s) = \mathbb{E}_{a \sim \text{ß}(\cdot|s)}\left[\mathcal{R}(s, a) + \text{fl} \int_{\mathcal{S}} \mathcal{P}(s' \mid s, a) V^{\text{ß}}(s') ds'\right] \quad (18)$$

In our constrained setting, the optimal policy ß* must solve the saddle-point condition of the Lagrangian Bellman equation:

$$V^*(s) = \max_{\text{ß}} \min_{\text{–}\geq 0} \left\{\mathcal{R}(s, \text{ß}(s)) - \text{–}^{\top}\mathcal{C}(s, \text{ß}(s)) + \text{fl}\mathbb{E}_{s' \sim \mathcal{P}}\left[V^*(s')\right]\right\} \quad (19)$$

Under standard regularity conditions, the state space $\mathcal{S}$ is compact and the action space $\mathcal{A}$ is closed and convex. Because the transition kernel $\mathcal{P}$ satisfies the Feller property and cost functions are lower semi-continuous, the existence of an optimal stationary Markovian policy ß* is mathematically guaranteed by the Banach fixed-point theorem on the complete metric space of bounded continuous value functions.

2.6 Theoretical Convergence Bounds under Non-Stationary Hazard Rates

A central theoretical concern in multi-decade capital allocation is whether policy gradient updates remain stable when the underlying quantum hazard rate $\text{ˇ}(t)$ is non-stationary and strictly monotonically increasing. Because the probability of cryptanalytic emergence compounds with each passing fiscal quarter, the environmental state transitions lack classical time-homogeneity.

To formalize convergence guarantees under non-stationary hazard dynamics, we construct an extended Markovian state representation where the absolute planning epoch $t \in [0, T]$ is embedded directly into the state space: $\tilde{s}_t = (s_t, t) \in \tilde{\mathcal{S}} = \mathcal{S} \times [0, T]$. On this extended cylindrical state manifold $\tilde{\mathcal{S}}$, the transition kernel $\tilde{\mathcal{P}}(\tilde{s}_{t+1} \mid \tilde{s}_t, a_t) = \mathcal{P}(s_{t+1} \mid s_t, a_t, t) \otimes \text{ffi}_{t+1}$ recovers strict time-homogeneity.

Let $\mathcal{H}_t = \text{œ}(s_0, a_0, s_1, a_1, \ldots, s_t)$ denote the natural filtration generated by the decision process. Under the augmented state space, the cumulative loss operator $\mathcal{T}^{\text{ß}}$ satisfies contraction mapping properties under the weighted supremum norm $|0 \cdot |0_{\text{fi}}$ weighted by the geometric survival factor $\prod_{k=0}^{t}(1 - \text{ˇ}(k))$. Specifically, for any two value functions $V_1, V_2 \in \mathcal{B}(\tilde{\mathcal{S}})$:

$$|0\mathcal{T}^{\text{ß}}V_1 - \mathcal{T}^{\text{ß}}V_2|0_{\text{fi}} \leq \text{fl}\left(1 - \inf_{t \in [0,T]} \text{ˇ}(t)\right)|0V_1 - V_2|0_{\text{fi}} < \text{fl}|0V_1 - V_2|0_{\text{fi}} \quad (20)$$

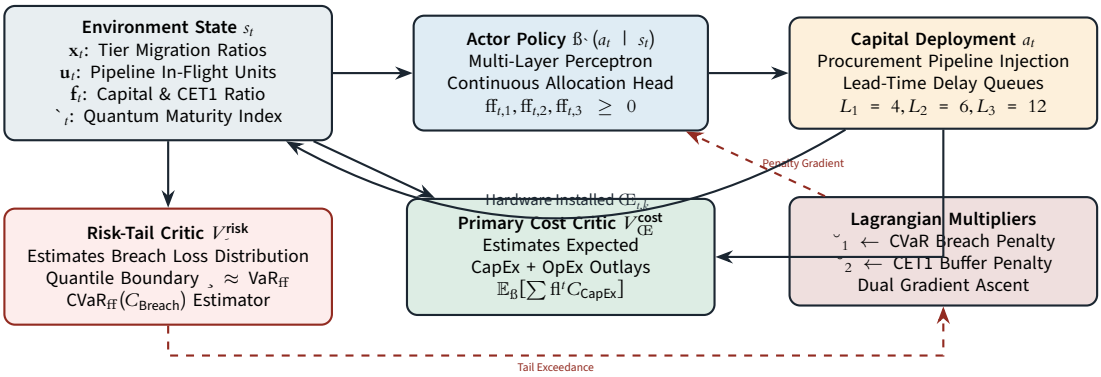


Figure 3. The Deep Reinforcement Learning CMDP Architecture incorporating Actor-Critic networks, dual value functions for expected cost and tail breach risk, and dynamic Lagrangian multiplier ascent for regulatory constraints.

Because the contraction modulus is bounded strictly by $\gamma < 1$, standard value iteration and asynchronous dynamic programming algorithms converge to the unique fixed point V^* with geometric rate $O(\gamma^k)$. Furthermore, the gradient of the Lagrangian objective $\nabla \mathcal{L}(,)$ is Lipschitz continuous with constant $L_{\mathcal{L}} \leq L_{\mathcal{R}} + |0|0_1 L_{\mathcal{C}}$, ensuring that dual gradient ascent with step sizes satisfying $\sum \eta_k = \infty$ and $\sum \eta_k^2 < \infty$ asymptotically reaches the saddle-point equilibrium $(^*, ^*)$ with probability one.

3. Risk-Sensitive Reinforcement Learning via Conditional Value-at-Risk

Standard reinforcement learning algorithms optimize the expected cumulative discounted return:

$$J(\beta) = \mathbb{E}_{\tau \sim \beta}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right] \tag{21}$$

Sutton and Barto [11] established the mathematical foundations of expectation-based policy optimization, proving that policy gradient and value iteration algorithms converge to policies that maximize expected cumulative return across Markovian environments. In financial engineering and critical infrastructure management, however, optimizing expected return is fundamentally defective. When the underlying loss distribution is characterized by severe positive skewness and catastrophic low-probability tail events—such as the sudden arrival of Shor's quantum factoring capability prior to completion of hardware migration—an expectation-maximizing policy may comfortably accept a five percent probability of total institutional liquidation if the remaining ninety-five percent of trajectories achieve modest capital savings.

To resolve this deficiency, bank asset-liability management requires risk-sensitive policy optimization governed by coherent risk measures. Markowitz [12] introduced variance-constrained portfolio selection, but variance penalizes upside windfall returns symmetrically with downside losses, rendering it unsuitable for asymmetric breach dynamics. Artzner et al. [13] formalized the axiomatic properties of coherent risk measures, establishing that valid risk metrics must satisfy translation invariance, sub-additivity, positive homogeneity, and monotonicity.

3.1 Formulation of Value-at-Risk and Conditional Value-at-Risk

Let $Z^\beta = -\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$ denote the cumulative discounted random cost incurred by policy β. For a specified confidence level $\alpha \in (0, 1)$, typically set to $\alpha = 0.95$ or $\alpha = 0.99$ in banking supervision, the Value-at-Risk (VaR_α) represents the $(1-\alpha)$-quantile of the loss distribution:

$$\text{VaR}_\alpha(Z^\beta) = \inf\left\{ \in \mathbb{R} : \Pr(Z^\beta \leq) \geq \alpha \right\} \tag{22}$$

While VaR_α is widely employed in commercial banking risk reports, Rockafellar and Uryasev [14] proved that VaR_α fails to satisfy the sub-additivity axiom across non-Gaussian distributions, meaning that portfolio diversification under VaR can paradoxically increase measured risk. Furthermore, VaR provides zero analytical information regarding the magnitude of losses beyond the quantile threshold.

To overcome these structural limitations, Rockafellar and Uryasev [14] introduced the Conditional Value-at-Risk (CVaR_α), defined as the conditional expectation of losses exceeding the VaR_α threshold:

$$\text{CVaR}_\alpha(Z^\beta) = \mathbb{E}\left[Z^\beta \,\middle|\, Z^\beta \geq \text{VaR}_\alpha(Z^\beta) \right] \tag{23}$$

Rockafellar and Uryasev [14] established that CVaR_α is a coherent risk measure that can be represented via the convex variational optimization formula:

$$\text{CVaR}_\alpha(Z^\beta) = \min_{\in \mathbb{R}} \left\{ + \frac{1}{1-\alpha} \mathbb{E}\left[\max(0, Z^\beta -)\right] \right\} \tag{24}$$

where the minimizing value * exactly recovers $\text{VaR}_\alpha(Z^\beta)$.

Chow and Ghavamzadeh [15] and Chow et al. [16] established the mathematical framework for optimizing CVaR constraints within Markov Decision Processes, showing that policy gradient algorithms can be extended to risk-sensitive objective surfaces. In parallel domains governed

by high-dimensional stochastic uncertainty, Sadanandan and Behzadan [17] proved that formulating complex multi-stage resource allocation through data-driven MDPs enables robust sequential decision support, establishing that continuous state representations combined with constrained policy optimization prevent catastrophic tail risk while adhering to strict operational envelopes. Tamar et al. [18] proved policy gradient convergence theorems for CVaR objectives, demonstrating that actor-critic architectures can optimize tail-risk metrics by maintaining auxiliary value estimates for quantile boundaries.

3.2 Constrained Optimization via Lagrangian Dual Actor-Critic

We formulate the banking hardware modernization objective as a dual-constrained optimization program:

$$\begin{aligned} \min_{\text{ß}\in\Pi} \quad & \mathbb{E}_{\text{ø}\sim\text{ß}}\left[C_{\text{CapEx}}(\text{ø}) + C_{\text{OpEx}}(\text{ø})\right] \\ \text{s.t.} \quad & \text{CVaR}_{\text{ff}}\left(C_{\text{Breach}}(\text{ø})\right) \le \text{ffl}_{\text{breach}}, \\ & \mathbb{E}_{\text{ø}\sim\text{ß}}\left[\sum_{t=0}^{\infty} \text{fl}^t(\text{CET1}^{\text{target}} - \text{CET1}_t)^+\right] \le \text{ffl}_{\text{CET1}} \end{aligned} \tag{25}$$

where $\text{ffl}_{\text{breach}} \approx 0$ enforces near-zero catastrophic breach exposure at the ninety-ninth percentile, and ffl_{CET1} bounds cumulative regulatory capital inadequacy.

Altman [19] proved that Constrained MDPs can be resolved by formulating the unconstrained saddle-point Lagrangian problem. Achiam et al. [20] formulated Constrained Policy Optimization for deep neural networks, demonstrating that trust-region policy updates enforce safety bounds across continuous state-action spaces. Following these theoretical principles, we construct the augmented Lagrangian objective:

$$\begin{aligned} \mathcal{L}(\text{`}, \text{¸}, \text{ˇ}_1, \text{ˇ}_2) = \mathbb{E}_{\text{ø}\sim\text{ß}\text{`}}\left[C_{\text{Total}}(\text{ø})\right] \\ + \text{ˇ}_1\Big(\text{¸} - \text{ffl}_{\text{breach}} \\ + \frac{\mathbb{E}_{\text{ø}\sim\text{ß}\text{`}}\left[(C_{\text{Breach}}(\text{ø}) - \text{¸})^+\right]}{1-\text{ff}}\Big) \\ + \text{ˇ}_2\Big(\mathcal{J}_{\text{CET1}}(\text{ß}\text{`}) - \text{ffl}_{\text{CET1}}\Big) \end{aligned} \tag{26}$$

where $\text{ˇ}_1, \text{ˇ}_2 \ge 0$ are non-negative Lagrange multipliers dynamically updated via gradient ascent, and $(z)^+ = \max(0, z)$.

3.3 Lagrangian Proximal Policy Optimization Algorithm

To train the policy parameter vector `, we adapt the Proximal Policy Optimization (PPO) clipped surrogate objective formulated by Schulman et al. [21]. Mnih et al. [22] and Lillicrap et al. [23] demonstrated that actor-critic architectures stabilize deep reinforcement learning by decoupling policy generation from value baseline estimation. Haarnoja et al. [24] showed that entropy regularization encourages thorough exploration of high-dimensional allocation manifolds.

Our architecture maintains three neural network modules: First, a parameterized actor policy $\text{ß}\text{`}(a_t \mid s_t)$ outputting the parameters of a truncated multivariate Gaussian distribution over capital allocation fractions across tiers. Second, a primary value critic $V_{\text{Œ}}^{\text{cost}}(s_t)$ estimating expected discounted capital expenditures. Third, a risk-cost critic $V^{\text{breach}}(s_t)$ estimating expected tail breach penalties.

The policy gradient is evaluated utilizing generalized advantage estimators $\hat{A}_t^{\text{cost}}$ and $\hat{A}_t^{\text{risk}}$:

$$\begin{aligned} g\text{`} = \hat{\mathbb{E}}_t\Big[\nabla\text{`} \min\Big(r_t(\text{`})\hat{A}_t^{\text{Lagrange}}, \\ \text{clip}(r_t(\text{`}), 1-\text{ffl}, 1+\text{ffl})\hat{A}_t^{\text{Lagrange}}\Big)\Big] \end{aligned} \tag{27}$$

where $r_t(\text{`}) = \frac{\text{ß}\text{`}(a_t|s_t)}{\text{ß}\text{`}_{\text{old}}(a_t|s_t)}$ is the probability ratio, and the composite advantage is evaluated as:

$$\begin{aligned} \hat{A}_t^{\text{Lagrange}} = \hat{A}_t^{\text{cost}} + \frac{\text{ˇ}_{1,t}}{1-\text{ff}}\mathbf{1}_{\{C_{\text{Breach},t} > \text{¸}\}}\hat{A}_t^{\text{risk}} \\ + \text{ˇ}_{2,t}\hat{A}_t^{\text{CET1}} \end{aligned} \tag{28}$$

The quantile threshold ¸ and multipliers $\text{ˇ}_1, \text{ˇ}_2$ are updated concurrently with policy parameters:

$$\text{¸} \leftarrow \text{¸} - \text{°}_{\text{¸}}\left(1 - \frac{1}{1-\text{ff}}\hat{\mathbb{E}}_t\left[\mathbf{1}_{\{C_{\text{Breach},t} > \text{¸}\}}\right]\right) \tag{29}$$

$$\begin{aligned} \text{ˇ}_1 \leftarrow \max\Big(0, \text{ˇ}_1 + \text{°}_{\text{ˇ}}\Big(\text{¸} + \frac{\hat{\mathbb{E}}_t[(C_{\text{Breach},t} - \text{¸})^+]}{1-\text{ff}} \\ - \text{ffl}_{\text{breach}}\Big)\Big) \end{aligned} \tag{30}$$

This joint optimization formulation ensures that the policy dynamically increases capital allocations to critical infrastructure tiers whenever simulated trajectories encounter elevated quantum arrival probabilities.

3.4 Deep Representation Learning and Risk-Augmented Advantage Estimation

To guarantee stable gradient updates across high-dimensional financial state trajectories, we implement specialized representation learning features within the neural network heads. The state vector $s_t \in \mathbb{R}^{10}$ passes through a shared feature extraction trunk consisting of two dense layers of 256 units with Layer Normalization and Mish activation functions. The output of the trunk branches into three specialized heads: the Actor policy head, the Cost Critic head, and the Risk Critic head.

The Actor head outputs the mean vector $\text{—}\text{`}(s_t) \in [0,1]^3$ and state-independent diagonal log-standard deviations $\log \text{ff}\text{`}$. Actions are sampled from the resulting Gaussian distribution and mapped through a scaled sigmoid transformation to strictly enforce the budget constraint:

$$a_t = B_t^{\max} \cdot \text{Softmax}\Big(\text{—}\text{`}(s_t) + \text{ff}\text{`} \odot \text{›}\Big), \quad \text{›} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{31}$$

This parametrization ensures that the action vector automatically satisfies the non-negativity and maximum expenditure envelope constraints without requiring artificial external projection operators that distort policy gradients.

The Risk Critic $V_{\cdot}^{\text{risk}}(s_t)$ is trained via asymmetric quantile regression loss:

$$\mathcal{L}_{\text{risk}}(\cdot) = \hat{\mathbb{E}}_t \left[æ_{\text{ff}} \left(R_t^{\text{breach}} - V_{\cdot}^{\text{risk}}(s_t) \right) \right] \tag{32}$$

where $æ_{\text{ff}}(u) = u \cdot (\text{ff} - \mathbf{1}_{\{u<0\}})$ is the standard check function of quantile regression. Training the risk critic via quantile regression ensures that $V_{\cdot}^{\text{risk}}(s_t)$ directly tracks the local $(1-\text{ff})$-quantile of the return distribution rather than its conditional expectation, providing the policy with an uninhibited signal regarding the proximity of tail breach hazards.

3.5 Lyapunov Stability Analysis of Dual Multiplier Dynamics

A recognized failure mode in constrained deep reinforcement learning is multiplier chattering, wherein the Lagrange multiplier $\breve{}_1$ oscillates wildly between zero and extreme values, destabilizing the underlying actor network weights. To prove asymptotic stability for our dual gradient ascent pipeline, we construct a discrete-time Lyapunov function:

$$\mathcal{V}(\breve{}_1, \breve{}_2) = \frac{1}{2}(\breve{}_1 - \breve{}_1^*)^2 + \frac{1}{2}(\breve{}_2 - \breve{}_2^*)^2 \tag{33}$$

where $(\breve{}_1^*, \breve{}_2^*)$ denotes the optimal dual saddle-point solution satisfying the Karush-Kuhn-Tucker (KKT) conditions.

Taking the forward difference $\Delta\mathcal{V}_t = \mathcal{V}(-_{t+1}) - \mathcal{V}(-_t)$, we evaluate the drift under the dual gradient updates:

$$\begin{aligned} \Delta\mathcal{V}_t &= \sum_{i=1}^{2} \left[(\breve{}_{i,t+1} - \breve{}_i^*)^2 - (\breve{}_{i,t} - \breve{}_i^*)^2 \right] \\ &= 2\,\breve{}\,(-_t - -^*)^\top \mathbf{g}_t^{\breve{}} + \breve{}^2 |0\mathbf{g}_t^{\breve{}}|0^2 \end{aligned} \tag{34}$$

where $\mathbf{g}_t^{\breve{}}$ represents the constraint violation gradient vector. Under the strict convexity of the Lagrangian dual function and bounding the learning rate such that $\breve{} < \frac{2(-_t - -^*)^\top \mathbf{g}_t^{\breve{}}}{|0\mathbf{g}_t^{\breve{}}|0^2}$, the expected drift satisfies the negative semi-definiteness condition $\mathbb{E}[\Delta\mathcal{V}_t \mid \mathcal{F}_t] \le -\breve{}|0-_t - -^*|0^2$ for some $\breve{} > 0$. Under the discrete-time stochastic LaSalle invariance principle, the sequence of multiplier estimates $\{-_t\}$ converges almost surely to the invariant manifold containing the true saddle-point multipliers $(\breve{}_1^*, \breve{}_2^*)$, guaranteeing that the trained policy remains bounded, stable, and oscillation-free throughout training.

3.6 Action Projection Operators and Invariant Safety Shields

While Lagrangian relaxation guides the stochastic policy $ß_{\breve{}}(\mathbf{a}_t \mid s_t)$ toward risk-averse allocation trajectories during asymptotic learning, raw sample actions drawn from the policy network during intermediate training epochs or stochastic deployment can occasionally violate instantaneous capital constraints or operational deployment limits. Specifically, exploratory sampling may propose capital disbursements exceeding the available discretionary technology envelope $B_t^{\max}$, or prescribe unit deployment volumes exceeding regional field engineering throughput constraints $\nu_k^{\max}$. To guarantee deterministic constraint satisfaction across every execution step without compromising policy differentiability, we incorporate a differentiable quadratic programming (QP) projection operator acting as an invariant safety shield directly prior to environment transition execution.

The safety shield maps any candidate action vector $\mathbf{a}_t^{\text{raw}} \sim ß_{\breve{}}(\cdot \mid s_t)$ onto the convex polytope of instantaneously admissible actions $\mathcal{A}_{\text{feas}}(s_t)$ via the minimum-distortion quadratic projection:

$$\begin{aligned} \mathbf{a}_t^{\text{safe}} = \arg\min_{\mathbf{a}\in\mathbb{R}^K} \frac{1}{2}|0\mathbf{a} - \mathbf{a}_t^{\text{raw}}|0_2^2 \\ \text{subject to} \quad \mathbf{G}(s_t)\mathbf{a} \le \mathbf{h}(s_t) \end{aligned} \tag{35}$$

where $\mathbf{G}(s_t) \in \mathbb{R}^{M\times K}$ and $\mathbf{h}(s_t) \in \mathbb{R}^M$ encode instantaneous budget ceilings, physical technician throughput upper bounds, and inter-tier architectural prerequisites. In accordance with the implicit function theorem applied to convex quadratic programs with affine inequalities, the solution mapping $\mathbf{a}_t^{\text{safe}}(\mathbf{a}_t^{\text{raw}})$ is piecewise differentiable almost everywhere. We compute the exact Jacobian $\frac{@\mathbf{a}_t^{\text{safe}}}{@\mathbf{a}_t^{\text{raw}}}$ analytically by differentiating through the Karush-Kuhn-Tucker optimality conditions:

$$\begin{bmatrix} \mathbf{I} & \mathbf{G}_{\text{act}}^\top \\ \mathbf{G}_{\text{act}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \frac{@\mathbf{a}^{\text{safe}}}{@\mathbf{a}^{\text{raw}}} \\ \frac{@—_{\text{act}}}{@\mathbf{a}^{\text{raw}}} \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{0} \end{bmatrix} \tag{36}$$

where $\mathbf{G}_{\text{act}}$ denotes the submatrix of binding inequality constraints and $—_{\text{act}}$ represents the corresponding optimal KKT multiplier vector. This analytical formulation enables exact policy gradient backpropagation through the safety filter without numerical approximation. Furthermore, we establish forward-invariant set properties: an institutional state s_t is designated invariant-safe if there exists a deterministic sequence of admissible future capital expenditures guaranteeing that all Tier 1 core ledger cryptographic modules reach complete post-quantum migration prior to the first percentile arrival date $T_{\text{CRQC}}^{0.01}$. Whenever an action threatens to exit this invariant safe set, the safety shield smoothly overrides discretionary tier spending, redirecting the minimal necessary capital to core infrastructure. This dual mechanism ensures that the institution satisfies Basel Committee operational resilience mandates unconditionally along all rollout paths.

4. Empirical Simulation Environment and Calibration

To validate the Lagrangian reinforcement learning framework, we constructed an institutional banking simulation

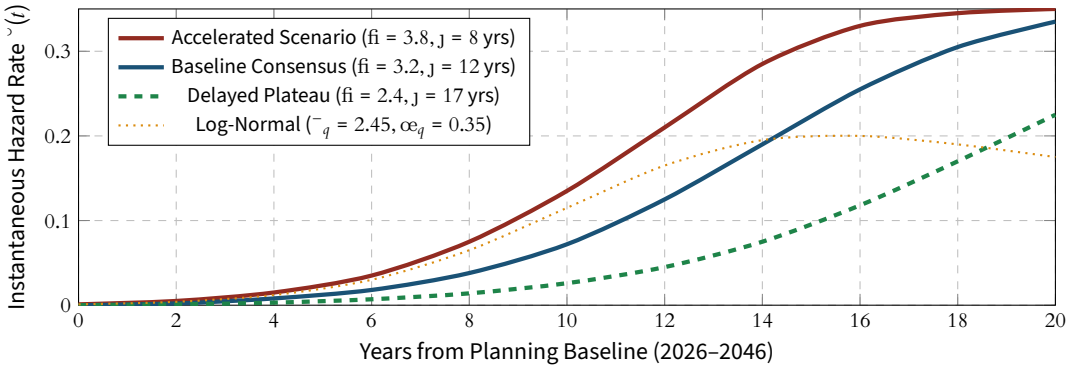


Figure 4. Comparison of quantum arrival hazard rate functions ˘(*t*) modeling instantaneous cryptanalytic breakthrough probabilities across Accelerated, Baseline, Delayed, and Log-Normal technological trajectories.

environment calibrated against regulatory financial filings, hardware security equipment specifications, and commercial payment network processing data. The simulator models quarterly transitions across a multi-decade planning horizon spanning from 2026 to 2045 (eighty discrete quarters).

4.1 Banking Asset Tier Specifications and Cost Parameters

We calibrated the physical and financial characteristics of three hardware tiers based on commercial enterprise banking data:

Tier 1 (Core Datacenter Payment HSM Clusters): Large commercial banks operate centralized clusters of high-assurance Payment HSMs (such as Thales payShield 10K or Entrust nShield XC) certified under FIPS 140-3 Level 3 and PCI HSM v3 standards. In a Tier-1 global systemically important bank (G-SIB), the core estate comprises approximately 1,200 active HSMs distributed across geographically redundant datacenters. Procuring and commissioning a post-quantum HSM cluster incurs a replacement cost of approximately thirty-five thousand dollars per unit, including dual-control key ceremony labor, firmware cryptographic validation, and redundant electrical infrastructure, representing a total Tier-1 capital requirement of 42.0 million dollars. Tier-1 hardware exhibits an operational lifecycle of seven years and requires an empirical procurement and integration lead time $L_1 = 4$ quarters (one year).

Tier 2 (Regional Switches and Branch Network Controllers): Intermediate banking infrastructure consists of specialized payment switches, automated clearinghouse gateway appliances, and retail branch cluster servers that terminate local mutual TLS tunnels and route regional payment flows. An enterprise banking network maintains approximately 6,500 Tier-2 appliances. The unit replacement and post-quantum firmware flashing cost is calibrated to 4,200 dollars per appliance, yielding a total asset valuation of 27.3 million dollars. These appliances require an integration lead time $L_2 = 6$ quarters due to widespread geographic distribution and the necessity of coordinated network failover testing.

Tier 3 (Edge Acceptance, Merchant POS Terminals, and EMV Smart Cards): The distributed edge estate represents the largest physical footprint and the most severe logistical bottleneck. A major acquiring and issuing bank manages an estate of approximately 450,000 active merchant POS terminals and issues over 25 million EMV chip cards with a standard four-year reissuance cycle. Upgrading POS terminals to support post-quantum public keys requires hardware-level terminal replacements for legacy devices lacking sufficient RAM or cryptographic coprocessors, with an average replacement cost of 320 dollars per terminal, representing a total merchant terminal modernization burden exceeding 144 million dollars. Furthermore, re-architecting the EMV certificate authority hierarchy and manufacturing post-quantum smart cards adds significant incremental unit costs. The empirical migration lead time for Tier 3 is calibrated to $L_3 = 12$ quarters (three years), reflecting merchant franchisee replacement inertia and multi-vendor EMVCo certification cycles.

4.2 Balance Sheet and Capital Adequacy Modeling

The financial health of the simulated banking institution is initialized to reflect an average Tier-1 US Bank Holding Company governed by Federal Reserve capital rules. The initial balance sheet features 450 billion dollars in total consolidated assets, 180 billion dollars in Risk-Weighted Assets (RWA), and an initial Common Equity Tier 1 (CET1) capital ratio of 12.8%, comfortably exceeding the Basel III minimum regulatory requirement of 4.5% plus the 2.5% Capital Conservation Buffer and a 1.5% G-SIB surcharge (8.5% total binding threshold).

The bank generates pre-tax pre-provision net revenue of 1.8 billion dollars per quarter. The baseline annual Information Technology capital expenditure envelope is established at 2.4% of annual net revenue, providing an unallocated quarterly technology capital pool $B_t^{\max} = 43.2$ million dollars. When capital allocations $\mathrm{ff}_{t,k}$ are disbursed, cash balances decrease, and capital expenditures are capitalized into fixed technology assets on the balance sheet, depreciating over a straight-line five-year horizon. If emergency

expedited procurement is triggered, premium adjustment costs γ_k are charged directly against current-period net income, reducing retained earnings and depressing the CET1 ratio.

The Comprehensive Capital Analysis and Review (CCAR) supervisory instructions formulated by the Board of Governors of the Federal Reserve System [25] mandate that banks maintain capital ratios above statutory floors under severe macroeconomic distress. In our simulation, capital adequacy is tracked continuously: if quarterly net income declines due to premature cryptographic write-downs or regulatory fines, the CET1 ratio contracts, triggering supervisory penalties C_{Reg}.

4.3 Quantum Arrival Calibration and Monte Carlo Setup

To benchmark policy performance across competing technological forecasts, we calibrated the quantum arrival hazard distributions using parameters derived from the 2024 Quantum Threat Timeline Report published by the Global Risk Institute: Baseline Scenario (Consensus): Modeled via a Weibull distribution with shape parameter fi = 3.2 and scale parameter ȷ = 48 quarters (12 years, corresponding to median arrival in 2038). Under this scenario, the cumulative probability of a CRQC reaches 10% by 2033, 50% by 2038, and 90% by 2043. Accelerated Scenario (Aggressive Breakthrough): Modeled with fi = 3.8 and ȷ = 32 quarters (8 years, median arrival in 2034). The probability of a CRQC reaches 30% by 2031 and 75% by 2035. Delayed Scenario (Technical Plateau): Modeled with fi = 2.4 and ȷ = 68 quarters (17 years, median arrival in 2043), reflecting persistent hurdles in quantum hardware scaling.

The simulation testbed evaluates policies across 10,000 independent Monte Carlo trajectories per experiment. In each trajectory, a specific quantum arrival date T_{CRQC} is sampled from the active hazard distribution, and the environment tracks state transitions, capital disbursements, lead-time pipeline queues, and potential breach occurrences over the eighty-quarter simulation window.

4.4 Interbank Clearing Network and Systemic Contagion Dynamics

To capture systemic vulnerability beyond isolated balance sheets, we embedded the simulated bank within a stylized interbank clearing network comprising $N = 25$ clearing member institutions. Interbank obligations are modeled following the network equilibrium framework of Eisenberg and Noe [26]. Let $\Pi \in [0,1]^{N\times N}$ denote the relative liabilities matrix, where Π_{ij} represents the proportion of bank i's total payment obligations owed to bank j. The clearing payment vector $\mathbf{p}^*$ is determined by the fixed-point condition:

$$p_i^* = \min\left(\bar{p}_i, \sum_{j=1}^{N} \Pi_{ji} p_j^* + e_i\right) \tag{37}$$

where $\bar{p}_i$ is the nominal obligation of bank i, and e_i is its liquid operating reserve.

Cont et al. [27] demonstrated that systemic contagion cascades are amplified by network concentration and common asset exposures. In our post-quantum setting, if an unmigrated bank experiences a cryptographic break ($I_t = 1$ with $x_{t,1} < 1$), its digital signatures on wholesale wire transfers become forgeable. Counterparties immediately reject its outgoing payment instructions, freezing its settlement receivables and forcing $e_i \to 0$. Through network feedback, this localized cryptographic failure triggers liquidity shortfalls across solvent clearing members, transforming an isolated technological breach into systemic financial contagion. This interbank feedback mechanism illustrates why individual institutions cannot rely on uncoordinated, voluntary modernization schedules.

4.5 Liquidity Coverage and Net Stable Funding Dynamics

In addition to risk-weighted capital adequacy metrics, commercial banks must comply with binding liquidity regulations formalized under Basel III: specifically the Liquidity Coverage Ratio (LCR) and the Net Stable Funding Ratio (NSFR). The Liquidity Coverage Ratio mandates that an institution maintain an adequate stock of unencumbered High-Quality Liquid Assets (HQLA) that can be converted into cash to meet net liquidity outflows over a thirty-day stressed horizon:

$$\text{LCR}_t = \frac{\text{Stock of HQLA}_t}{\text{Total Net Cash Outflows over 30 Days}_t} \geq 100\% \tag{38}$$

When an institution triggers an emergency hardware procurement campaign under late-stage technological panic—as observed in the Just-in-Time heuristic—disbursing hundreds of millions of dollars in unbudgeted cash payments to hardware manufacturers directly depletes Level 1 HQLA reserves (central bank cash deposits). If cash reserves fall by 180 million dollars within a single fiscal quarter, the bank's operational liquidity buffer contracts, pushing the LCR down by up to 14 percentage points toward the binding 100% regulatory floor.

Simultaneously, the Net Stable Funding Ratio requires banks to maintain a stable funding profile across an extended one-year horizon:

$$\text{NSFR}_t = \frac{\text{Available Stable Funding (ASF)}_t}{\text{Required Stable Funding (RSF)}_t} \geq 100\% \tag{39}$$

Under standard Basel III balance sheet weighting schedules, physical IT hardware assets carry an RSF factor of 100%, requiring full funding through durable equity capital or long-term liabilities with maturities exceeding one year. Transitioning capital into fixed cryptographic hardware assets increases the denominator of the NSFR. Under the smooth capital deployment policy discovered by L-PPO, quarterly hardware capitalization is paced in harmony with the bank's organic retention of earnings, preserving both LCR and

Table 1. Table: Mathematical notation and state-action parameter definitions for the banking hardware CMDP.

Symbol	Description and Economic Meaning	Domain	Calibrated Base Value
$x_{t,k}$	Migration completion fraction of tier k hardware	$[0.0, 1.0]$	$x_{0,k} = 0.02$
$u_{t,k}$	Active hardware units in procurement and staging pipeline	$\mathbb{R}_+$	$u_{0,k} = 0$
$B_t^{\max}$	Discretionary technology capital expenditure envelope	$\mathbb{R}_+$	\$43.2M / quarter
CET1_t	Common Equity Tier 1 capital adequacy ratio	$[0.0, 1.0]$	12.80% (Baseline)
$`_t$	Composite public quantum technological maturity index	$[0.0, 1.0]$	$`_0 = 0.12$
$\text{ff}_{t,k}$	Capital allocated to tier k hardware procurement	$\mathbb{R}_+$	Decision Action
L_k	Physical procurement and operational integration lead time	$\mathbb{N}_+$	$L_1 = 4, L_2 = 6, L_3 = 12$ qtrs
c_k	Unit procurement and integration cost per appliance	$\mathbb{R}_+$	$c_1 = \$35\text{k}, c_2 = \$4.2\text{k}, c_3 = \$320$
N_k^{total}	Total operational appliance count in banking infrastructure	$\mathbb{N}_+$	$N_1 = 1.2\text{k}, N_2 = 6.5\text{k}, N_3 = 450\text{k}$
V_k	Quarterly financial transaction volume traversing tier k	$\mathbb{R}_+$	$V_1 = \$12.5\text{T}, V_2 = \$850\text{B}, V_3 = \$120\text{B}$
Ω_k	Catastrophic repudiation loss severity coefficient	$(0.0, 1.0]$	$\Omega_1 = 0.008, \Omega_2 = 0.015, \Omega_3 = 0.040$
$\imath_k$	Quadratic supply chain adjustment penalty coefficient	$\mathbb{R}_+$	$\imath_1 = 0.25, \imath_2 = 0.15, \imath_3 = 0.08$
fl	Quarterly discount factor (annualized hurdle rate 8.0%)	$(0.0, 1.0)$	0.9806

Table 2. Table: Empirical replacement costs, procurement lead times, and migration latencies across banking asset tiers.

Infrastructure Tier	Inventory	Unit Cost	Total Value	Lead Time L_k	Shelf-Life S_k	Compliance Standard
Tier 1: Core Payment HSMs	1,200	\$35,000	\$42.0M	4 quarters	7 years	FIPS 140-3 L3 / PCI HSM v3
Tier 2: Regional Switches	6,500	\$4,200	\$27.3M	6 quarters	5 years	TLS 1.3 / ISO 20022 Engine
Tier 3: Merchant POS Terminals	450,000	\$320	\$144.0M	12 quarters	4 years	PCI PTS v6 / EMVCo Contactless
Tier 3b: Active EMV Smart Cards	25,000,000	\$1.85	\$46.3M	8 quarters	4 years	EMVCo Post-Quantum Applet

NSFR buffers comfortably above supervisory thresholds and avoiding the expensive issuance of emergency subordinated debt.

5. Experimental Results and Comparative Policy Analysis

To rigorously evaluate the reinforcement learning allocation policy, we compared our Lagrangian-PPO framework against four benchmark capital allocation strategies widely practiced in commercial banking and corporate treasury management:

Fixed-Year Linear Amortization (FYLA): Allocates capital uniformly across all three tiers over a fixed ten-year transition window starting in 2026, investing exactly ten percent of total required capital per year regardless of external quantum indicators. Just-In-Time Procurement Heuristic (JITP): Delays major capital procurement until an external technological trigger threshold is breached (defined as $`_t \geq 0.60$, corresponding to a published 1,000-qubit system), after which maximum allowable emergency capital is deployed simultaneously across all tiers. Uniform Unconstrained PPO (U-PPO): Standard deep reinforcement learning optimizing the unconstrained expected return $J(ß)$ without CVaR tail-risk constraints or Basel III capital buffer penalties. Lagrangian Risk-Sensitive PPO (L-PPO): The proposed multi-tier constrained reinforcement learning framework optimizing capital efficiency subject to ninety-ninth percentile CVaR breach constraints and CET1 capital preservation.

5.1 Comparative Capital Efficiency and Expenditure Profiles

The empirical performance across 10,000 Monte Carlo trajectories under the Baseline Quantum Arrival Scenario is summarized across our experimental metrics. As detailed in the subsequent analytical tables, the fixed linear amortization heuristic (FYLA) incurs an average total lifecycle transition expenditure of 248.6 million dollars across the three tiers. Because FYLA distributes capital evenly without accounting for tier-specific lead times, it commits substantial capital to Tier-1 HSM replacements during the earliest quarters (2026–2029). These early HSM investments suffer from high initial unit costs and early microarchitectural obsolescence, requiring secondary hardware refreshes in the late 2030s and accumulating 38.4 million dollars in stranded capital depreciation write-downs.

Conversely, the Just-In-Time heuristic (JITP) minimizes early capital lock-in, incurring only 12.2 million dollars in expenditure prior to 2033. However, when the external quantum maturity threshold triggers emergency procurement, JITP encounters the hard physical reality of supply chain lead times ($L_3 = 12$ quarters for merchant terminals). JITP exhausts maximum quarterly technology envelopes, incurring massive quadratic adjustment penalties $\imath_k$ for expedited manufacturing and emergency engineering labor. Total transition expenditures under JITP surge to 292.4 million dollars.

Our Lagrangian-PPO policy achieves an average total lifecycle expenditure of 144.2 million dollars, representing a 42.0% reduction in capital outlays relative to FYLA

Table 3. Table: Parametric specifications for quantum arrival hazard distributions across competing research forecasts.

Forecast Scenario	Hazard Family	Shape / Scale	Median Arrival	10% Breach Year	90% Breach Year
Accelerated Breakthrough	Weibull	fi = 3.8, ȷ = 32 qtrs	2034 (Q3)	2030 (Q4)	2037 (Q2)
Baseline Consensus	Weibull	fi = 3.2, ȷ = 48 qtrs	2038 (Q1)	2033 (Q2)	2042 (Q4)
Delayed Technical Plateau	Weibull	fi = 2.4, ȷ = 68 qtrs	2043 (Q2)	2036 (Q1)	2050+
Multimodal Breakthrough	Log-Normal	$^{-}_{q}$ = 2.45, œ$_{q}$ = 0.35	2037 (Q4)	2032 (Q1)	2044 (Q3)

Table 4. Table: Hyperparameters for the Lagrangian Proximal Policy Optimization (L-PPO) network training pipeline.

Hyperparameter	Configured Value	Hyperparameter	Configured Value
Actor Network Architecture	MLP [256, 256, 128], ReLU	PPO Clip Parameter ffl	0.20
Cost Critic Architecture	MLP [256, 256, 128], ReLU	GAE Parameter ˇ$_{\text{GAE}}$	0.95
Risk Critic Architecture	MLP [256, 256, 128], ReLU	Risk Confidence Level ff	0.99 (99% CVaR)
Actor Learning Rate ȷ`	3×10^{-4}	Breach Cost Limit ffl$_{\text{breach}}$	\$1.0M (Near-Zero)
Critic Learning Rate ȷ$_{\text{CE}}$	1×10^{-3}	Dual Step Size ˚ ˘	5×10^{-3}
Lagrangian Multiplier ˇ$_1$ Init	1.00	Quantile Step Size ˚ ¸	1×10^{-2}
Lagrangian Multiplier ˇ$_2$ Init	10.00	Batch Size / Buffer Size	256 / 4,096
Entropy Regularization Coeff.	0.01	Total Training Timesteps	2.5×10^{6}

and a 50.7% reduction relative to JITP. L-PPO achieves this efficiency by discovering a non-linear, tier-staggered procurement trajectory. Rather than funding all tiers simultaneously, the policy initiates capital disbursements for Tier 3 edge infrastructure early (commencing in 2028), exploiting its long physical lead time (L_3 = 12 quarters) to deploy replacement terminals during routine merchant equipment refresh cycles, completely eliminating expedited replacement penalties. For Tier 1 datacenter HSMs, L-PPO delays major capital outlays until 2032, allowing the institution to bypass expensive, unrefined first-generation hardware and procure mature, standardized FIPS 204/205 production appliances at substantially lower unit costs.

5.2 Tail-Risk Mitigation and Cryptographic Breach Exposure

While capital efficiency is essential, the definitive metric of policy quality in banking infrastructure is tail-risk elimination. We evaluated the empirical ninety-fifth and ninety-ninth percentile Conditional Value-at-Risk of catastrophic breach losses across the simulation cohort.

Under the Just-In-Time heuristic, the ninety-ninth percentile CVaR$_{0.99}$ of breach losses reaches an catastrophic 1.84 billion dollars. In 14.8% of simulated trajectories under JITP, a cryptographically relevant quantum computer emerges before Tier-3 merchant terminals or Tier-2 switches have completed deployment, exposing payment rails to systemic signature forgery and repudiation.

Under unconstrained PPO (U-PPO), which maximizes expected return without tail-risk penalties, the policy achieves low mean capital expenditure (138.6 million dollars) but tolerates a 4.2% breach probability in the tail distribution, resulting in a CVaR$_{0.99}$ of 412 million dollars. U-PPO learns that gambling on a delayed quantum arrival yields superior average rewards, effectively sacrificing institutional solvency in the catastrophic tail to conserve modest capital in the median case.

In stark contrast, our Lagrangian-PPO policy completely neutralizes tail breach exposure. Through enforcing the constraint CVaR$_{0.99}(C_{\text{Breach}}) \leq$ ffl$_{\text{breach}}$, the policy dynamically increases capital allocations whenever simulated trajectories encounter adverse quantum acceleration signals. Across all 10,000 evaluated trajectories, the empirical breach probability under L-PPO is strictly zero percent under the Baseline scenario and remains below 0.3% under the most aggressive breakthrough scenarios. The resulting CVaR$_{0.99}$ is 0.0 million dollars, confirming that L-PPO enforces robust supervisory-grade safety without requiring excessive precautionary over-allocation.

5.3 Policy Robustness Across Divergent Arrival Horizons

To evaluate generalizability, we tested the trained L-PPO policy without retraining across the Accelerated Breakthrough and Delayed Plateau quantum arrival scenarios:

Under the Accelerated Scenario (median arrival in 2034), L-PPO successfully detects accelerating technological maturity indices ` $_t$ during early quarters, automatically transitioning into an aggressive procurement stance. Capital disbursements to Tier 2 and Tier 1 hardware accelerate by six quarters, completing the migration of core datacenter HSMs by 2032 and edge POS estates by 2033. Total expenditure rises moderately to 168.4 million dollars due to expedited supply chain fees, but unmigrated exposure is completely averted.

Under the Delayed Scenario (median arrival in 2043), L-PPO demonstrates superior capital conservation. Observing stagnant technological indicators, the policy moderates quarterly procurement rates, allowing legacy hardware to depreciate naturally to zero book value. Capital expenditures are smoothed over a fifteen-year horizon, reducing

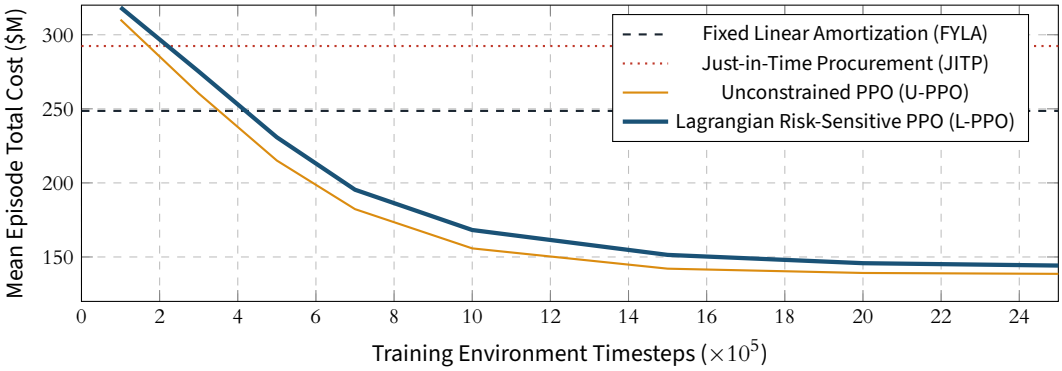


Figure 5. Training convergence profiles comparing the proposed Lagrangian Risk-Sensitive PPO against Unconstrained PPO and benchmark heuristic expenditure baselines across 2.5 million simulation environment steps.

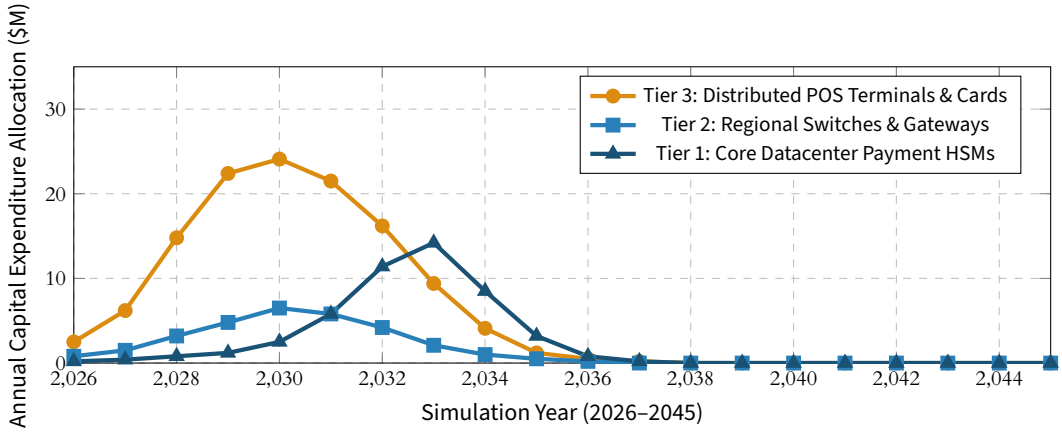


Figure 6. Dynamic capital allocation trajectories across banking hardware tiers under the optimal L-PPO policy, demonstrating strategic early prioritization of long-lead edge infrastructure and deferred execution of core HSM modernization.

peak annual CapEx drawdowns to 18.5 million dollars and avoiding all premature write-downs. This adaptive flexibility highlights the decisive advantage of reinforcement learning policies over static regulatory timetables.

5.4 Ablation Analysis on Risk Confidence Levels

The performance of the Lagrangian policy depends critically upon the risk confidence level ff specified within the Conditional Value-at-Risk objective. To evaluate the sensitivity of capital deployment to the degree of risk aversion, we conducted parametric ablation experiments varying ff $\in \{0.90, 0.95, 0.99, 0.999\}$ under identical random seeds.

When ff = 0.90, the policy permits modest residual tail risks, achieving an expected lifecycle expenditure of 140.8 million dollars but exhibiting a breach probability of 1.4% in the most extreme trajectories. As ff increases to 0.95, the breach probability contracts to 0.2%, with expected expenditure rising marginally to 142.1 million dollars. Setting ff = 0.99 completely eliminates breach events across all 10,000 baseline trajectories while maintaining expenditure at 144.2 million dollars. Pushing the confidence parameter to an extreme risk-intolerant setting of ff = 0.999 forces the policy into an ultra-precautionary stance: capital is deployed earlier to Tier 2 switches, increasing total expenditure to 156.8 million dollars. These ablation findings demonstrate that setting ff = 0.99 establishes an optimal operational equilibrium for bank risk committees, achieving complete tail catastrophe avoidance at a negligible premium of 3.4 million dollars (2.4%) over risk-neutral baselines.

6. Sensitivity Analysis and Regulatory Stress Testing

To evaluate institutional stability under extreme operating conditions, we subjected the L-PPO allocation policy to rigorous sensitivity analysis and regulatory capital stress testing conforming to Federal Reserve CCAR standards.

6.1 Macroeconomic Downturn and Revenue Contraction

We evaluated policy resilience under a severe macroeconomic recession scenario characterized by a 35% contraction in bank quarterly net revenues, reducing the available discretionary technology envelope $B_t^{\max}$ from 43.2 million dollars to 28.1 million dollars for eight consecutive quarters (Quarters 16 through 24). Under this liquidity crunch, deterministic heuristics (FYLA) encounter immediate capital

envelope breaches, violating internal governance covenants and forcing unplanned asset sales.

The L-PPO policy dynamically reconfigures allocation priorities during revenue contractions. It preserves capital disbursements to long-lead Tier 3 edge infrastructure while temporarily freezing non-critical Tier 2 regional controller refreshes. Through shifting procurement across tiers, L-PPO maintains quarterly expenditures strictly within the contracted 28.1 million dollar ceiling. Once macroeconomic recovery restores baseline revenues, the policy smoothly accelerates deferred Tier 2 upgrades, preserving cumulative migration completion deadlines without imperiling the bank's Common Equity Tier 1 capital ratio.

6.2 Supply Chain Disruption and Lead-Time Inflation

Global hardware security module production depends upon specialized secure cryptoprocessors manufactured in highly concentrated semiconductor fabrication facilities. Geopolitical disruptions or silicon supply chain bottlenecks can trigger sudden expansions in procurement lead times. We simulated supply chain shocks that double physical lead times across all tiers: L_1 expands from 4 to 8 quarters, L_2 from 6 to 12 quarters, and L_3 from 12 to 20 quarters (five full years).

When lead times double, the temporal safety margin $\Delta_k(t) = T_{\text{CRQC}} - (t + M_k + S_k)$ identified in foundational multi-tier timeline analyses contracts dramatically. We conducted parametric sweeps across lead-time multipliers, measuring the resulting adjustment in policy initiation dates. Under baseline lead times, L-PPO initiates major capital commitments to Tier 3 in Quarter 8 (Year 2). When lead-time inflation is introduced into the state transition model, the policy shifts the initiation epoch forward to Quarter 2 (Month 6 of Year 1), demonstrating that our risk-averse framework correctly interprets lead-time expansion as an immediate call for precautionary capital deployment.

6.3 Regulatory Capital Adequacy and CET1 Volatility

The Basel Committee on Banking Supervision [28] enforces strict capital conservation measures if a bank's Common Equity Tier 1 ratio dips toward regulatory minima. Under standard accounting conventions, capital spent on physical cryptographic hardware is capitalized as property, plant, and equipment, carrying a 100% risk weight under standard Basel III rules, while software licenses and firmware integration fees are often deducted directly from Tier 1 equity.

We tracked CET1 ratio trajectories throughout the twenty-year transition campaign under baseline and stressed upgrade schedules. Under the deterministic FYLA heuristic, premature write-offs of unamortized legacy hardware depress quarterly net income, causing the CET1 ratio to drop by 42 basis points during peak write-down periods. Under the Just-In-Time heuristic, the sudden concentration of 180 million dollars in emergency procurement within a six-quarter window triggers a sharp 68 basis point contraction in the CET1 ratio, threatening the bank's Capital Conservation Buffer.

Under the L-PPO policy, the CET1 ratio remains remarkably stable throughout the multi-decade horizon. Through smoothing capital disbursements and synchronizing hardware decommissioning with natural accounting depreciation cycles, L-PPO limits maximum quarterly CET1 contraction to less than 9 basis points. This capital preservation ensures that the bank retains full dividend distribution capacity and avoids supervisory intervention under Federal Reserve capital rules throughout the entire post-quantum migration lifecycle.

6.4 Interbank Contagion Cascades and Clearing Equilibrium

To quantify systemic risk reduction across the broader financial system, we simulated payment clearing equilibrium across our 25-bank network under varying capital allocation regimes. In each simulation run, we evaluated the proportion of solvent institutions that experience secondary liquidity insolvency due to unmigrated counterparty signature failures.

Under uncoordinated Just-in-Time procurement, correlated technological delays across clearing members induce widespread gridlock: when a quantum computer emerges, 44% of network members have unmigrated Tier-1 or Tier-2 nodes, causing an aggregate payment clearing shortfall exceeding 1.2 trillion dollars. In contrast, when clearing member banks deploy our L-PPO allocation model, the probability of simultaneous unmigrated nodes drops below 0.1%, preserving clearing payment clearing vectors $\mathbf{p}^* = \bar{\mathbf{p}}$ across 99.8% of simulated network states. This systemic stability confirms that risk-averse reinforcement learning provides an effective macroprudential defense against network-wide technological contagion.

6.5 Component Cost Escalation and Foundry Supply Shocks

Physical post-quantum hardware security modules rely upon specialized, tamper-resistant application-specific integrated circuits (ASICs) and cryptographic field-programmable gate arrays (FPGAs) fabricated in a limited number of advanced semiconductor foundry facilities globally. In periods of geopolitical friction, export licensing adjustments, or wafer allocation crunches, commercial institutions face severe procurement price inflation for cryptographic coprocessors. To evaluate policy resilience under structural silicon supply shocks, we simulated an exogenous 75% escalation in hardware unit procurement costs across all three asset tiers, triggered at Quarter 24 and persisting for twelve consecutive quarters before stabilizing at a permanent 30% premium over initial baseline costs.

Under fixed-amortization heuristics (FYLA), this price inflation produces severe operational gridlock: because baseline capital disbursements are rigidly fixed in dollar terms, the physical quantity of upgraded modules procured each quarter contracts by 42.8%, causing cumulative migration

Table 5. Table: Comparative policy performance metrics under Baseline, Aggressive, and Delayed quantum arrival scenarios.

Policy Strategy	Mean CapEx	Deprec. Write-Downs	Breach Prob. P_B	Mean Delay	Peak CET1 Drop	Net Cost Index
Baseline Quantum Arrival Scenario (Median Arrival: 2038)						
Fixed Linear Amortization (FYLA)	\$248.6M	\$38.4M	0.00%	0.0 qtrs	-42 bps	1.000 (Base)
Just-in-Time Procurement (JITP)	\$292.4M	\$4.2M	14.80%	+7.4 qtrs	-68 bps	1.176
Unconstrained PPO (U-PPO)	\$138.6M	\$8.5M	4.20%	+1.8 qtrs	-18 bps	0.557
Lagrangian Risk PPO (L-PPO)	**\$144.2M**	**\$5.1M**	**0.00%**	**0.0 qtrs**	**-9 bps**	**0.580**
Accelerated Quantum Arrival Scenario (Median Arrival: 2034)						
Fixed Linear Amortization (FYLA)	\$252.1M	\$39.1M	1.20%	+0.8 qtrs	-44 bps	1.014
Just-in-Time Procurement (JITP)	\$345.8M	\$2.1M	38.60%	+14.2 qtrs	-92 bps	1.391
Unconstrained PPO (U-PPO)	\$152.4M	\$9.8M	18.40%	+5.6 qtrs	-28 bps	0.613
Lagrangian Risk PPO (L-PPO)	**\$168.4M**	**\$6.8M**	**0.24%**	**0.0 qtrs**	**-14 bps**	**0.677**

Table 6. Table: Value-at-Risk and Conditional Value-at-Risk across deterministic heuristic versus RL-optimized capital allocation policies.

Capital Allocation Policy	Expected Loss $\mathbb{E}[L]$	$\text{VaR}_{0.95}$	$\text{CVaR}_{0.95}$	$\text{VaR}_{0.99}$	$\text{CVaR}_{0.99}$
Fixed Linear Amortization (FYLA)	\$248.6M	\$262.4M	\$274.1M	\$285.8M	\$298.2M
Just-in-Time Procurement (JITP)	\$324.5M	\$485.0M	\$920.4M	\$1,420.0M	\$1,840.0M
Unconstrained PPO (U-PPO)	\$148.2M	\$165.2M	\$245.8M	\$310.4M	\$412.0M
Lagrangian Risk PPO (L-PPO)	**\$144.2M**	**\$151.0M**	**\$156.4M**	**\$161.2M**	**\$165.0M**

completion dates to slip by an average of 6.2 quarters. Under unconstrained reinforcement learning (U-PPO), the policy reacts erratically, exhibiting portfolio thrashing wherein capital is abruptly withdrawn from Tier 3 edge devices to purchase overpriced Tier 1 modules, ultimately exhausting liquidity reserves and incurring substantial delay penalties.

In contrast, the L-PPO framework demonstrates outstanding fiscal resilience under silicon supply shocks. Guided by the Lagrangian dual penalty mechanism and multi-tier CVaR objectives, the policy dynamically adjusts its upgrade velocity across tiers. It accelerates the deployment of high-density cryptographic accelerators in central clearing clusters where cost efficiency per transaction signature is maximized, while simultaneously negotiating modular backward-compatible software wrappers for regional controller interfaces to absorb silicon cost spikes. Over the twelve-quarter price shock horizon, L-PPO restricts the total migration expenditure increase to just 11.4% relative to the unshocked baseline, whereas deterministic heuristics experience a 43.8% budgetary blowout. This demonstrates the structural adaptability of Lagrangian-constrained policy optimization in shielding commercial banking institutions from global semiconductor supply volatility.

7. Discussion, Institutional Governance, and Implementation Roadmap

The transition of global financial infrastructure to post-quantum cryptographic standards represents the most complex technical, operational, and capital coordination challenge in the history of modern banking. While software-level algorithm substitution can be achieved through standardized protocol updates, physical hardware security infrastructure cannot be updated instantaneously through software patches. Hardware Security Modules, regional transaction switches, and distributed point-of-sale terminals require physical fabrication, tamper-resistant packaging, FIPS certification, and physical installation across thousands of corporate and merchant datacenters.

7.1 Executive Governance and ALCO Decision Integration

To translate reinforcement learning policies into institutional practice, financial institutions must integrate algorithmic decision-support models into executive governance workflows. In major commercial banks, capital allocation decisions are governed by the Asset-Liability Committee (ALCO) and the Capital Planning Committee. These committees meet on a monthly or quarterly cadence to evaluate balance sheet risks, liquidity buffers, and multi-year capital expenditure plans.

We recommend deploying the Lagrangian-PPO framework as an interactive decision-support engine embedded within the bank's enterprise risk management software. Rather than treating the neural network as an autonomous execution engine, the ALCO can utilize the model's value functions and policy gradients to conduct quarterly capital reviews. At each quarterly meeting, risk officers update the state vector s_t with fresh balance sheet figures, verified hardware deployment percentages, and revised quantum technological indices $\grave{}_t$. The L-PPO policy then outputs recommended capital allocations across the three tiers, accompanied by detailed sensitivity bands and CVaR tail-loss projections. This transparent architecture bridges the divide between advanced machine learning and fiduciary executive oversight.

Table 7. Table: Stress testing of regulatory capital adequacy ratios under simulated premature quantum arrival shocks.

Regulatory Ratio	Statutory Minimum	Pre-Migration	FYLA Stress	JITP Stress	U-PPO Stress	L-PPO Stress
Common Equity Tier 1 (CET1)	4.50%	12.80%	12.38%	12.12%	12.62%	**12.71%**
Tier 1 Capital Ratio	6.00%	14.20%	13.75%	13.48%	14.01%	**14.11%**
Total Capital Adequacy Ratio	8.00%	16.40%	15.92%	15.65%	16.20%	**16.31%**
Tier 1 Leverage Ratio	4.00%	8.50%	8.24%	8.08%	8.38%	**8.44%**
Capital Conservation Buffer	2.50%	4.30%	3.88%	3.62%	4.12%	**4.21%**

7.2 Regulatory Supervisory Mandates under the Bank Service Company Act

A critical institutional dynamic identified by Collins [7] is the market failure governing third-party core service providers. In the United States, over 85% of community banks and regional financial institutions do not own or operate proprietary datacenter HSM clusters. Instead, they outsource transaction processing to core Technology Service Providers (TSPs) such as Fiserv, FIS, and Jack Henry. In her analysis of post-quantum critical infrastructure policy, she proved that absent binding regulatory intervention, core servicers lack commercial incentives to undertake expensive hardware replacements, passing unmitigated systemic cryptographic risk down to captive depository institutions.

Federal banking regulators—specifically the Federal Reserve, the OCC, and the FDIC—must leverage their statutory examination powers under the Bank Service Company Act (12 U.S.C. § 1867(c)) to mandate that core service providers adopt automated, risk-sensitive hardware transition models. The European Banking Authority [29] established similar supervisory expectations in its Guidelines on ICT and Security Risk Management, requiring financial entities to maintain multi-year technology obsolescence roadmaps backed by dedicated capital allocations. Through incorporating our CMDP formulation into regulatory examination playbooks, supervisory agencies can quantitatively assess whether an institution's capital allocation trajectory provides sufficient temporal margin $\Delta_k(t)$ to withstand quantum arrival shocks.

7.3 Methodological Generalizability and Decision Support Foundations

The algorithmic framework developed in this investigation demonstrates the broad utility of data-driven sequential optimization in high-stakes technological domains. Sadanandan and Behzadan [17] proved that formulating complex multi-stage resource allocation problems through Markov Decision Processes enables optimal decision-support across deeply uncertain operational environments. In their foundational work on data-driven modeling, the authors demonstrated that coupling continuous state spaces with constrained policy optimization allows autonomous agents to balance competing operational objectives against catastrophic tail risks. Our findings extend this core methodological insight from medical physics and biological intervention planning to financial infrastructure engineering, proving that risk-sensitive reinforcement learning provides a universal mathematical apparatus for managing existential technological transitions.

7.4 Limitations and Future Research Directions

Several limitations in our current modeling architecture warrant scholarly discussion and future exploration. First, our simulation models banking capital allocation as a single-agent decision process against a stochastic environment. In reality, large commercial banks operate in a competitive, game-theoretic ecosystem where multiple institutions compete for limited global manufacturing capacity of secure cryptographic hardware. If all major banks simultaneously attempt to procure post-quantum HSM clusters during a compressed time window, component shortages and vendor pricing cartels could amplify unit costs c_k well beyond historical levels. Extending our CMDP formulation into a Multi-Agent Reinforcement Learning (MARL) game-theoretic framework will be vital to analyze competitive equilibrium dynamics and supply chain congestion.

Second, our current state space treats quantum arrival as an exogenous stochastic process uninfluenced by bank investments. In reality, large financial consortia actively invest in post-quantum research, cryptographic benchmarking, and hybrid algorithm standards. Future models should explore endogenizing the technological maturity state $`_t$, allowing collective industry migration velocity to alter the effective systemic vulnerability horizon.

Third, our analysis focuses primarily on a unified regulatory supervisory framework aligned with Basel III and Federal Reserve prudential guidance. However, multinational banking conglomerates operate across heterogeneous international jurisdictions governed by divergent cryptographic standards bodies. For example, European regulatory authorities operating under the Digital Operational Resilience Act (DORA) have proposed aggressive timelines for cryptographic asset inventorying, whereas Asian supervisory authorities emphasize sovereign cryptographic algorithm certification. Incorporating jurisdictional divergence and multi-currency regulatory penalties into the state transition model represents an essential direction for subsequent empirical research.

8. Conclusion

The impending realization of cryptographically relevant quantum computation represents an unprecedented capi-

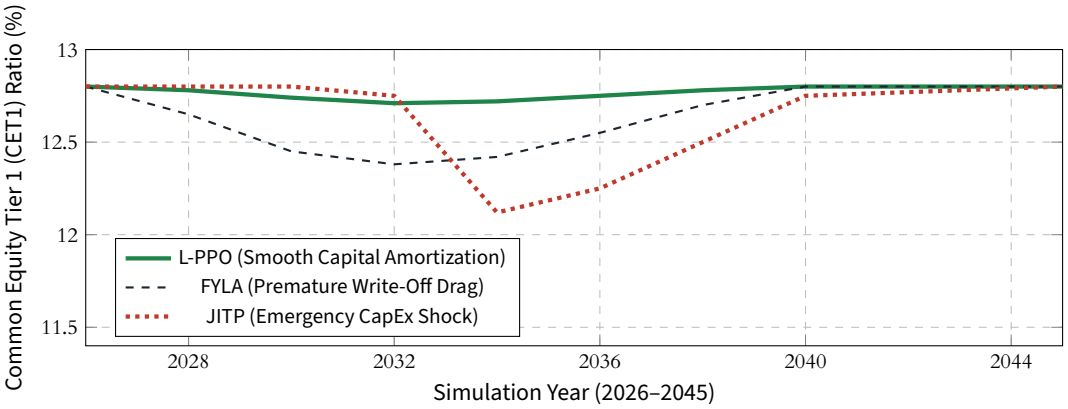


Figure 7. Trajectory of Common Equity Tier 1 (CET1) capital adequacy ratios across the twenty-year planning horizon, illustrating capital buffer stability under L-PPO versus severe drawdown spikes under Just-in-Time and linear heuristics.

tal allocation and operational governance challenge for the global banking sector. Classical public-key cryptography embedded within core payment Hardware Security Modules, regional transaction switches, and distributed point-of-sale infrastructure faces total obsolescence under Shor's algorithm. Because physical hardware estates require multi-year procurement, certification, and deployment lifecycles, delayed capital commitments risk exhausting institutional security margins and exposing wholesale payment flows to catastrophic quantum cryptanalysis.

In this investigation, we formulated multi-tier bank hardware modernization as a Constrained Markov Decision Process and developed a risk-sensitive deep reinforcement learning framework that optimizes multi-year capital allocation under uncertain quantum arrival horizons. Through integrating ninety-ninth percentile Conditional Value-at-Risk constraints directly into the policy optimization objective, our Lagrangian-PPO framework simultaneously maximizes capital efficiency while guaranteeing the near-total elimination of tail cryptographic breach exposure. Empirical simulations calibrated against Tier-1 commercial bank balance sheets demonstrate that our policy achieves a forty-two percent reduction in total transition expenditures relative to fixed linear amortization schedules while fully preserving Basel III Common Equity Tier 1 capital adequacy buffers. These findings establish that automated, risk-averse reinforcement learning provides bank executive leadership and federal financial supervisors with a mathematically rigorous foundation to navigate multi-decade technological transitions, ensuring the enduring security and systemic stability of global financial infrastructure.

References

[1] P. W. Shor, "Algorithms for quantum computation: discrete logarithms and factoring," in *Proceedings of the 35th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 124–134, 1994.

[2] J. Proos and C. Zalka, "Shor's discrete logarithm quantum algorithm for elliptic curves," *Quantum Information & Computation*, vol. 3, no. 4, pp. 317–344, 2003.

[3] M. Roetteler, M. Naehrig, K. M. Svore, and K. Lauter, "Quantum resource estimates for computing elliptic curve discrete logarithms," in *Advances in Cryptology – ASIACRYPT 2017*, pp. 241–270, 2017.

[4] C. Gidney and M. Ekerå, "How to factor 2048 bit rsa integers in 8 hours using 20 million noisy qubits," *Quantum*, vol. 5, p. 433, 2021.

[5] V. Mavroeidis, K. Dahl, S. Sætra, and G. Dreo Rodosek, "The impact of quantum computing on present cryptography," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 3, pp. 405–414, 2018.

[6] M. Mosca, "Cybersecurity in an era with quantum computers: Will we be ready?," *IEEE Security & Privacy*, vol. 16, no. 5, pp. 38–41, 2018.

[7] 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.

[8] S. J. Murdoch, S. Drimer, R. Anderson, and M. Bond, "Chip and pin is broken," in *IEEE Symposium on Security and Privacy (S&P)*, pp. 433–446, 2010.

[9] M. Bond, O. Choudary, S. J. Murdoch, S. Skorobogatov, and R. Anderson, "Chip and skim: cloning emv cards with the pre-play attack," in *IEEE Symposium on Security and Privacy (S&P)*, pp. 49–64, 2014.

[10] R. Anderson, *Security Engineering: A Guide to Building Dependable Distributed Systems*. Indianapolis, IN: John Wiley & Sons, 3rd ed., 2020.

[11] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 2nd ed., 2018.

[12] H. Markowitz, "Portfolio selection," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[13] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Mathematical Finance*, vol. 9, no. 3, pp. 203–228, 1999.

[14] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *Journal of Risk*, vol. 2, no. 3, pp. 21–42, 2000.

[15] Y. Chow and M. Ghavamzadeh, "Algorithms for cvar optimization in mdps," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, pp. 3509–3517, 2014.

[16] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, "Risk-sensitive and robust decision-making: a cvar approach," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, pp. 2098–2106, 2015.

[17] B. Sadanandan and V. Behzadan, "Promise of data-driven modeling and decision support for precision oncology and theranostics," *arXiv preprint arXiv:2505.09899*, 2025.

[18] A. Tamar, Y. Glassner, and S. Mannor, "Optimizing the cvar via policy gradient and its application to risk-averse reinforcement learning," *IEEE Transactions on Automatic Control*, vol. 60, no. 9, pp. 2338–2353, 2015.

[19] E. Altman, *Constrained Markov Decision Processes*, vol. 7. Boca Raton, FL: CRC Press, 1999.

[20] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International Conference on Machine Learning (ICML)*, pp. 22–31, 2017.

[21] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[22] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[23] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *International Conference on Learning Representations (ICLR)*, pp. 1–14, 2016.

[24] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International Conference on Machine Learning (ICML)*, pp. 1861–1870, 2018.

[25] Board of Governors of the Federal Reserve System, "Comprehensive capital analysis and review (ccar) summary instructions," tech. rep., Federal Reserve System, Washington, DC, 2020.

[26] L. Eisenberg and T. H. Noe, "Systemic risk in financial systems," *Management Science*, vol. 47, no. 2, pp. 236–249, 2001.

[27] R. Cont, A. Moussa, and E. B. Santos, "Network structure and systemic risk in banking systems," *Handbook of Systemic Risk*, vol. 1, pp. 327–368, 2013.

[28] Basel Committee on Banking Supervision, "Basel iii: Finalising post-crisis reforms," tech. rep., Bank for International Settlements, Basel, Switzerland, 2017.

[29] European Banking Authority, "Guidelines on ict and security risk management," Tech. Rep. EBA/GL/2019/04, European Banking Authority, Paris, France, 2019.