Ramanujan College
University of Delhi
(Accredited Grade 'A++' by NAAC)

DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science (Honours) Computer Science

Post-Quantum TLS in Serverless Systems: Cold-Start, Tail Latency, and Cost Amplification under Autoscaling

Submitted by:
Chirag
Roll No: 20221414

Under the Supervision of:
Dr. Nikhil Kumar Rajput
Assistant Professor, Department of Computer Science

Co-Supervisor:
Mr. Sahil Pathak
Assistant Professor, Department of Computer Science

Department of Computer Science
Ramanujan College, University of Delhi
oct 2025

CERTIFICATE

This is to certify that the dissertation entitled **"Post-Quantum TLS in Serverless Systems: Cold-Start, Tail Latency, and Cost Amplification under Autoscaling"** submitted by **Chirag**, Roll No. **20221414**, in partial fulfillment of the requirements for the award of the degree of **Bachelor of Science (Honours) Computer Science** of University of Delhi, is a record of the candidate's own work carried out by him/her under my supervision and guidance.

The matter embodied in this dissertation is original and has not been submitted for the award of any other degree or diploma.

Dr. Nikhil Kumar Rajput
Assistant Professor
Department of Computer Science
Ramanujan College
University of Delhi

Mr. Sahil Pathak
Professor
Department of Computer Science
Ramanujan College
University of Delhi

Date: _____ **Place:** New Delhi

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all those who have contributed to the completion of this dissertation.

First and foremost, I am deeply grateful to my supervisor, **Dr. Nikhil Kumar Rajput**, for her invaluable guidance, continuous support, and patience throughout my research. Her expertise in machine learning and software engineering has been instrumental in shaping this work. I also extend my heartfelt thanks to my co-supervisor, **Mr. Sahil Pathak**, for his insightful feedback and encouragement.

I would like to thank the faculty members of the Department of Computer Science, Ramanujan College, for providing me with a strong foundation in computer science principles and for creating an environment conducive to learning and research.

My sincere appreciation goes to my fellow students and researchers who provided valuable discussions, feedback, and moral support throughout this journey. Special thanks to the lab members who helped me with data collection and experimental setup.

I am grateful to the developers and maintainers of the open-source projects and datasets that made this research possible. The availability of quality resources significantly contributed to the success of this work.

I would also like to acknowledge the computational resources provided by the college's computer lab, which were essential for conducting the experiments presented in this dissertation.

Finally, I express my deepest gratitude to my family and friends for their unwavering support, understanding, and encouragement throughout my academic journey. Their belief in me has been a constant source of motivation.

Chirag
Roll No: 20221414
oct 2025

ABSTRACT

PRE-EXPERIMENT VERSION;

The advent of post-quantum cryptography (PQC) presents a new performance challenge for cloud-native systems, particularly in ephemeral environments such as Function-as-a-Service (FaaS). While traditional TLS key exchanges (e.g., X25519) are optimized for low-latency connections, PQC schemes such as Kyber introduce significantly larger key sizes and computation costs, potentially amplifying cold-start and handshake delays in serverless workloads. This research aims to empirically evaluate the performance impact of NIST FIPS-approved post-quantum and hybrid key-exchange algorithms across representative serverless platforms — AWS Lambda, Google Cloud Functions, and OpenFaaS. Using a TLS 1.3 microservice instrumented with the Open Quantum Safe (OQS) OpenSSL library, the study will benchmark handshake latency, memory footprint, CPU utilization, and cold-start overhead under varying request concurrency (10–1000 req/s). It is hypothesized that pure PQC handshakes (e.g., Kyber768) will incur higher latency and resource usage than classical TLS, and that hybrid configurations (e.g., X25519 + Kyber512) may mitigate these penalties while maintaining quantum resilience. The findings are expected to provide a quantitative foundation for optimizing PQC-hybrid adoption in serverless deployments, balancing post-quantum security with operational efficiency.

Keywords: Post-Quantum Cryptography, Serverless Computing, TLS 1.3, Hybrid Key Exchange, Kyber, Cold Start Latency, Open Quantum Safe

Chapter 1
INTRODUCTION

1.1 Background

The cryptic principles of cloud-based infrastructure are changing radically. The existence of large scale quantum computers poses a threat to the security of broadly used public-key cryptographic schemes like RSA, Diffie-Hellman, and Elliptic Curve Cryptography (ECC), which are used to provide Transport Layer Security (TLS) and most of the global Internet traffic is already secured **[3,10]**.

To counter this, the National Institute of Standards and Technology (NIST) has made a new generation of post-quantum cryptographic (PQC) algorithms, including lattice-based key encapsulation schemes such as ML-KEM (Kyber) to be resistant to quantum adversaries **[10]**.

At the same time, cloud computing has continued to develop into serverless and microservice-driven execution patterns that focus on the elastic autoscaling and ephemeral execution environments as well as strict tail-latency service level objectives **[8]**.

Function-as-a-Service (FaaS) platforms like AWS Lambda or Google Cloud Functions are dynamically scaled production instances, responding to demand, which is better at utilizing resources, but also creates cold-start latency, connection churn, and limited potential connection reuse **[6]**.

The two trends intercepted on TLS layer. Cryptographic handshakes are amortized in long-lived connections in traditional persistent server settings **[5]**.

By comparison, serverless architectures re-executes the TLS handshakes several times in cold and warm execution modes. The increased key sizes, ciphertexts, and computational cost of PQC and hybrid TLS handshakes are thus multiplied by autoscaled churn and fan-out of microservice to make cryptographic overhead a leading factor in tail latency and operational cost **[1,11]**.

Although strategies to implement PQC TLS have already been proved feasible in the context of more traditional server and network models **[3,5,10]**, its influence at the system level in elastic serverless systems has not been studied extensively.

1.2 Problem Statement

Current testing of post-quantum and hybrid TLS key exchange systems are centered largely around persistent server settings and network-oriented metrics of key exchange performance, including handshake latency with deterministic packet loss and bandwidth conditions. These analyses make the assumption of long lived server processes, connection reuse, and amortised cost of cryptography.

These assumptions are however fundamentally broken by serverless computing platforms. Short lived execution, churn of instances in autoscaling and fan-out of connections in microservices results in multiplied TLS handshakes, which generate a feedback loop where cryptographic overhead directly amplifies cold-start latency, warm-path tail latency, and usage-based billing cost. Although the introduction of FaaS has led to the widespread usage of the platforms to implement latency-sensitive microservices, there is no existing literature that analyses the impact of post-quantum and hybrid TLS handshakes on performance in terms of latency distributions, autoscaling behavior, and operational cost in elastic serverless environments in a systematic, empirical manner.

The absence of quantitative models and empirical evidence to describe the system-level and economic contribution of post-quantum and hybrid TLS key exchange of autoscaling serverless execution is thus the main issue of this dissertation.

1.3 Research Objectives

The main aim of the study is to estimate and model the amplification of cryptographic overhead by post-quantum and hybrid TLS 1.3 key exchanges in elastic serverless microservice systems.

The specific objectives are:

• Measure cold-start latency, warm-path tail latency, and connection reuse collapse in classical, hybrid and PQC-only TLS.

Scale D. Measure autoscaling churn and hand shake amplification with scale.

Model under usage-based billing of servers is:

• Determine the cryptographic dominance levels where the actual cryptographic overhead becomes the most important factor of tail latency.

• Suggest deployment-level optimization measures that do not degrade quantum security and reduce performance reduction.

1.4 Research Questions

RQ1: What are the impact of post-quantum and hybrid TLS key exchange mechanisms on cold-start latency, warm-path tail latency, and connection reuse in serverless microservice systems?

RQ2: How does autoscaling churn increase cryptographic overhead with concurrency?

RQ3: What are the best PQC and hybrid key exchange configurations in terms of balancing quantum resistance and operation efficiency in elastic FaaS environments?

RQ4: How hybrid key exchange systems alleviate cryptographic dominance in comparison to PQC-only systems?

RQ5: How can cryptographic amplification be minimized and still ensure quantum security?

1.5 Research Methodology Overview

The research study is based on an experimental approach, in which the controlled system deployment, empirical measurement, and statistical analysis are merged.

The PQC-enabled TLS stacks (Envoy with OQS-OpenSSL) are used to implement a Netflix-style fan-out microservice architecture on serverless platforms. It produces controlled workload bursts to trigger autoscaling churn behavior and cold-start behavior. The metrics are gathered using the distributed tracing and the Prometheus instrumentation to measure the latency, handshake dominance, arriving connections reuse collapse, and the inflation of the billing cost. The statistical analysis is conducted through non-parametric procedures in order to determine significant differences between performance of different cryptographic settings.

1.6 Scope and Limitations

1.6.1 Scope

This research focuses on:

• Quantum-resistant and hybrid 1.3 key exchange.

Elastic serverless platforms such as AWS Lambda, Google Cloud Functions and OpenFaaS.

Cold-start latency, warm-path tail latency, autoscaling churn and operational cost metrics.

Lattice based NIST-standardized algorithms (primarily ML-KEM (Kyber)).

1.6.2 Limitations

• VM-based and bare-metal environments are not considered persistent.

• TLS handshake behavior is only tested; long-term session throughput is not tested.

Network variability is regulated instead of being entirely representative of the common Internet.

Energy efficiency and PQC hardware acceleration are not taken into consideration.

1.7 Contributions

Previous system-level empirical analysis of PQC TLS with elastic serverless execution.

The authors were able to identify cryptographic phase-transition behavior with autoscaling churn.

PQC TLS Cost inflation models under usage-based billing.

• A repeatable benchmarking model of PQC analysis in FaaS systems.

• Implementation advice of quantum-safe cloud services.

1.8 Organization of the Dissertation

The rest of this dissertation is arranged in the following way:

Chapter 2 provides a review on related work on PQC, hybrid TLS, and serverless performance.

Chapter 3 covers the statistical analysis, instrumentation and the methodology of the experiment.

Chapter 4 demonstrates the experimental findings and relative assessment.

The implications, limitations, and future research directions are discussed in Chapter 5.

Chapter 2
LITERATURE REVIEW

2.1 Introduction

The upcoming creation of resilient quantum computers is a paradigm shift to the cryptography underpinning of contemporary digital infrastructure. RSA, DiffieHellman and Elliptic Curve Cryptography (ECC), mathematically vulnerable to the Shor quantum factoring and discrete logarithm algorithms, are used to secure the vast majority of Internet traffic via public-key cryptographic schemes. Post-Quantum Cryptography (PQC) has in turn become a focus of research worldwide, leading to the standardization of lattice-based, code-based and hash-based quantum-resistant primitives by NIST.

At the same time the paradigm shift in cloud computing has been to serverless and cloud-native microservice architectures defined by transient execution, extreme auto-scaling, and hard-tail service level guarantees. These systems are fundamentally different to traditional persistent server environments where the cost of TLS handshakes and cryptography is amortized over long-lived connections. Rather, Function-as-a-Service (FaaS) systems recycle and demolish execution environments time and time again, leading to sequence of TLS handshakes and multiplying cryptographic overhead.

A significant amount of recent research has stringently tested PQC in TLS 1.3 in controlled network scenarios, assessing the latency of a handshake, bandwidth inflation, and throughput effect. Nevertheless, in its synthesis as presented in this chapter, the literature on the subject is strictly limited to the view of the persistent client-server model and network analysis. PQC has not been previously systematically studied and how it varies cold-start latency, warm-path tail latency, autoscaling churn, and cost of operation of elastic serverless microservice fabrics. This is a critical gap that is covered in this dissertation.

2.2 Post-Quantum Cryptography and the Quantum Threat Landscape

One of the common topics in the literature reviewed is the acute awareness of the threat model harvest-now, decrypt-later (HNDL) [3]. Several research works point to the fact that even the rivalry can be gathering encrypted traffic to decrypt it in the future when quantum computers appear, and an active transition to PQC is not only a theoretical necessity but also an operational emergency.

Matta and Bolli showed that although organizational awareness of quantum risk is high the maturity of actual PQC deployment is moderate, making the adoption of PQC both an architectural and governance issue as much as a cryptographic one. According to their results, it is generally accepted that hybrid migration strategies comprising classical and post-quantum primitives are the most effective transitional strategy since they allow finding a balance between security and performance, and they are also backward-compatible [2, 10].

These results position PQC as an algorithmic substitution but rather as a structural change that modulates network behavior, resource usage, and cost dynamics through cloud infrastructures [1, 11].

2.3 Performance of Post-Quantum TLS 1.3

2.3.1 Network-Level TLS Benchmarks

Another line of research is predominant; it is concerned with the performance aspects of PQC being integrated into TLS 1.3. One of the most detailed early studies was given by Sosnowski et al., who showed that lattice-based implementation schemes like Kyber and Dilithium tend to be as efficient as or more efficient than classical RSA and ECC [5], with the bandwidth overhead being the key performance bottleneck. Their effort also confirmed that deployment of hybrid TLS does not have serious computational costs and thus can be deployed in the near future as long as the TCP congestion window parameters are adjusted correctly.

This analysis was extended by Heinrich et al. who measured TLS handshake latency with simulated packet loss, jitter and bandwidth [5,10]. Their findings showed that structured lattice-based KEMs (Kyber, Saber, NTRU) are robust even in the face of most network conditions and are often faster than classical ECDH, but conservative schemes like FrodoKEM, HQC, and BIKE are highly performance sensitive to adverse network conditions because of the large-sized public keys and ciphertexts [10]. This paper also demonstrated that PQC handshake latency depends not only on raw computation, but also on TCP fragmentation, initial congestion windows, and MTU constraints.

Complementary experiments that assessed Time-To-Last-Byte (TTLB) behavior have shown that although PQC handshakes can increase the initial connection establishment time, the overhead is rapidly decreasing with the volume of application data. Kampanakis demonstrated that TLS connections using PQC incur a TTLB degradation of less than 5-15 percent where hundreds of kilobytes are transferred over the network, which supports the finding that PQC can be deployed with the traditional web workloads[4].

All these works point to the conclusion that post-quantum TLS is not only computationally feasible, but the performance is also controlled by bandwidth and TCP dynamics instead of the cryptographic computation itself.

2.4 PQC in Cloud-Native and Microservice Environments

Stepping further, however, and beyond the single client-server model, more recent efforts have been put into researching the PQC in cloud-native microservice ecosystems. In a 2023 study of scalable end-to-end encryption of microservices, sidecar-based, PQC-enabled service mesh was proposed, showing that hybrid Kyber-based key exchange increased the latency of handshakes by nearly 69 percent but had little effect on steady-state throughput. There was a moderate (~6.8) CPU overhead, which was considered operationally acceptable and it justified the viability of PQC in horizontally scalable microservice fabrics.

On the same note, cloud-security transition surveys highlighted the fact that a PQC migration presents an initial complexity of architectural, governance, and compliance challenges that extend beyond cryptographic replacement. Bigger keys and ciphertexts augment memory strain, network fragmentation, and resource drain- problems, which are exacerbated in scalable cloud settings.

Nevertheless, these works still presuppose sustained service cases and long-lasting relations and have a critical gap in terms of ephemeral serverless execution.

2.5 Serverless Cold Starts and Elasticity

Literature has extensively shown that the cold-start latency and churn of autoscaling dominate the serverless platforms particularly during bursty workloads. Cold yields also mean a larger initialisation delay, dependency loading and initialisation network overhead which cumulatively impacts tail latency in fan-out microservice graphs. These phenomena represent a fundamental contrast to persistent server settings where TLS handshake overhead can be amortized thousands of requests.

However, all the reviewed PQC TLS studies do not measure the interactions of quantum-safe cryptography with cold starts, instance churn, connection reuse collapse, and billing physics in FaaS systems. Current literature assesses the handshake latency separately without accounting for the feedback mechanism that arises with autoscaling and transient execution.

2.6 Identified Research Gap

Summarizing all the reviewed literature, it is possible to draw three conclusions:

Computationally feasible, Post-quantum TLS 1.3 and hybrid TLS 1.3 are frequently competitive with classical cryptography on the network and server.

It is mainly bandwidth and TCP behavior that determines performance degradation and not the mere cryptographic computation.

PQC overheads can be tolerated by cloud-native microservices in persistent deployments.

Nevertheless, there are no previous studies that assess the compounded impact of PQC on cold-start latency, warm-path tail latency, autoscaling churn, connection reuse, and the cost of operation in elastic serverless systems. This vulnerability is essential, since these are the very places where the TLS handshakes are most numerous, amortization is the least strong, and the billing models directly convert the cryptographic overhead into the financial cost [6,8].

2.7 Economic Cost Models and PQC in Usage-Based Cloud Billing

Although network-centric performance models determine the level of computational feasibility of PQC, they do not deal with the issue of the translation of cryptographic overhead into cost in the form of usage-based billing models. Cryptographic operations are monetized directly using serverless platforms which charge per millisecond of CPU time and allocated memory.

According to surveys on PQC migration readiness, cost uncertainty rather than algorithmic security is the most prevalent impediment to adoption. The larger the size of the PQC keys and ciphertexts, the more memory pressure and CPU channels per connection, which in the elastic billing models lead to high cost per request [1, 10]. These costs are amortized in long-lived connections in persistent server models; in serverless environments, instance churn is so high that there is no amortization. In such a way, PQC overhead will be a first-order cost determinant.

The state of the art in PQC TLS does not measure this economic amplification, and cloud operators have no actionable cost models of deploying PQC in an elastic environment [1, 11].

2.8 Fan-Out Amplification and Tail-Latency Physics

Dean and Barroso generalized tail latency into a performance bottleneck dominant in distributed systems [8]. With microservice architectures, a request can fan out with dozens of downstream services, increasing per-hop latency in tailing end-to-end.

In fan-out intersects autoscaling in serverless fabrics forming a feedback loop:

The incidence of bursty demand causes churning.

• Instance churn fails TLS connection reuse.

• Reuse failure doubles TLS handshakes.

The effects of handshakes: Handshake amplification enlarges cold and warm latencies.

• Autoscaling is further instigated by inflated latency.

Persistent server environments do not have this loop, and current PQC TLS benchmarks do not even pay attention to this. Consequently, the previous literature grossly underestimates the amplification of PQC overhead in the system level [1, 10, 11].

2.9 Why Existing TLS Benchmarks Are Incomplete

In all literature reviewed on PQC TLS, three de facto assumptions prevail:

Long-lived server processes

Stable connection reuse

Minute churn in autoscaling.

These are not true in the FaaS environment where instances can serve just a few requests before they are destroyed. Therefore, current literature is unable to capture:

• Cold-start amplification

• Warm-path hand shake dominance.

• Reuse collapse

• Cost feedback loops

Tail-latency explosion under fan-out.

Thus, although current literature is right regarding the fact that PQC is cryptographically feasible, it does not describe the characteristics of whether PQC is economically and architecturally viable in the modern elastic serverless computing.

2.10 Positioning of the Present Work

This dissertation introduces the first empirical evaluation of PQC and hybrid TLS 1.3 under elastic serverless microservice execution. It extends existing PQC TLS benchmarks into a new operational regime by incorporating:

- Cold-start latency
- Warm-path tail latency (p95/p99)
- Autoscaling churn
- Fan-out amplification
- Connection reuse collapse
- Billing cost inflation

By isolating cryptography as the sole independent variable, the study quantifies the **quantum tax** imposed on serverless architectures.

2.11 Summary

This chapter synthesized literature spanning quantum-threat modeling, NIST PQC standardization, PQC TLS 1.3 benchmarking, cloud-native microservices, and serverless cold-start dynamics. While extensive research demonstrates the cryptographic feasibility of PQC and hybrid TLS under persistent deployments, no prior work evaluates their compounded effects on cold-start latency, warm-path tail latency, autoscaling churn, and operational cost in elastic serverless environments.

This dissertation fills this gap by delivering the first system-level characterization of PQC's performance and economic impact in serverless microservice fabrics.

3. Methodology

3.1 Research Objective

The objective of this research is to quantitatively characterize how post-quantum and hybrid TLS 1.3 key exchange mechanisms alter latency, tail-latency behavior, autoscaling dynamics, and operational cost in elastic serverless microservice architectures. The study focuses on the system-level amplification of cryptographic overhead arising from ephemeral execution environments, autoscaling churn, and fan-out communication graphs — conditions absent in prior PQC TLS benchmarks.

3.2 Experimental Architecture

A Netflix-style fan-out microservice topology was implemented using serverless functions. The architecture consists of three logical tiers:

- **Service A (Frontend):** Receives client requests and fans out parallel calls.
- **Services B_1–B_3 (Workers):** Perform independent processing in parallel.
- **Service C (Aggregator):** Aggregates worker responses and returns the final result.

All inter-service communication occurs over TLS-secured HTTP/2 connections. TLS termination is implemented inside each function using Envoy sidecar proxies compiled against OQS-OpenSSL to allow runtime selection of cryptographic algorithms.

3.3 Cryptographic Configurations

Four mutually exclusive cryptographic universes are evaluated:

Configuration	TLS 1.3 Key Exchange
Classical	X25519

Hybrid-512	X25519 + Kyber512
Hybrid-768	X25519 + Kyber768
PQC-Only	Kyber768

All cryptographic policy is applied uniformly across every service instance. All non-cryptographic software components are held constant.

3.4 Deployment Platforms

Experiments are conducted on three serverless platforms representing distinct autoscaling models:

Platform	Model
AWS Lambda	Proprietary managed FaaS
Google Cloud Functions	Managed container FaaS
OpenFaaS	Kubernetes-native autoscaling

3.5 Workload Model

Traffic is generated using **k6** to induce autoscaling churn and cold-start events. Each experiment consists of cyclic traffic phases:

Phase	Load	Duration
Baseline	50 RPS	300 s
Burst	1000 RPS	60 s
Cooldown	50 RPS	300 s

Each full cycle is repeated **30 times per cryptographic universe**.

3.6 Measurement Instrumentation

Each function instance exports structured metrics:

Metric	Definition
tls_handshake_seconds	Time to complete TLS handshake
request_latency_seconds	End-to-end request latency
cpu_seconds	CPU consumed per request
memory_rss_bytes	Peak resident memory
instance_lifetime_seconds	Duration instance remains active

handshakes_per_instance	Number of TLS handshakes
reuse_ratio	Requests per instance
cold_start	Binary cold/warm indicator

Metrics are collected via Prometheus and correlated using distributed tracing.

3.7 Latency Decomposition

Total request latency is decomposed as:

ttotal=thandshake+tapplicationt_{total} = t_{handshake} + t_{application}ttotal=thandshake+tapplication

This separation allows causal attribution of cryptographic overhead.

3.8 Derived Metrics

Handshake Dominance Ratio (HDR)

HDR=thandshakettotalHDR = \frac{t_{handshake}}{t_{total}}HDR=ttotalthandshake

Effective Hot-Path Latency Inflation (EHLI)

EHLI=p99PQC−p99ClassicalEHLI = p99_{PQC} - p99_{Classical}EHLI=p99PQC−p99Classical

Operational Cost Inflation

Cost=∑(CPUsec×Pcpu+GBsec×Pmem)Cost = \sum (CPU_{sec} \times P_{cpu} + GB_{sec} \times P_{mem})Cost=∑(CPUsec×Pcpu+GBsec×Pmem)

3.9 Phase-Transition Detection

Serverless platforms are nonlinear to latencies and concurrently scaled under the autoscaling limits, connection reuse failure, and cold start amplification. A phase-transition detection methodology is used to determine the point at which cryptographic overhead is the most important contributor to the system latency.

During the burst of one workload cycle controlled increments of concurrency are gradually increased. To compute the Handshake Dominance Ratio (HDR) for every level of concurrency, the ratio is calculated:

HDR=thandshakettotalHDR = \frac{t_{handshake}}{t_{total}}HDR=ttotalthandshake

where thandshaket_{handshake}thandshake is the TLS handshake latency and ttotalt_{total}ttotal is the total request latency.

A system is considered to have entered the **cryptographic dominance regime** when:

HDR>0.3HDR > 0.3HDR>0.3

This value is used to represent that cryptographic processes are adding over 30 percent of the overall request latency and a qualitative shift of behavior of the system. Beyond this, additional concurrency increases result in super-linear increase in tail latency as handshake amplification and reuse collapse. The phase-transition boundary is to compare the resilience of classical and hybrid and PQC settings to the effect of autoscaling on cryptographic amplification.

3.10 Statistical Analysis

In order to represent the average and worst case performance behavior, the latency distributions are summarized with the p50 (median), p95 and p99 percentiles. These percentiles become particularly applicable to serverless systems in which Service Level Objectives (SLOs) are usually formulated in terms of tail latency.

Every experimental setup is repeated 30 times yielding statistically significant latency distributions. The MannWhitney U test is used to test the significance of observed differences between cryptographic configurations with a significance level of alpha=0.05875. The use of this non-parametric test is based on the fact that serverless latency distributions are non-Gaussian and heavy-tailed.

The effect sizes are also calculated to measure the extent of the cryptographic effects beyond statistical significance, and therefore, observed differences are statistically and practically significant.

3.11 Reproducibility Controls

To ensure experimental validity and reproducibility, strict control variables are enforced across all experimental runs:

• **Infrastructure-as-Code (IaC)** is used to provision all cloud resources, ensuring identical deployment configurations across experiments.
• **Cryptographic configuration is the sole independent variable.** All other software components, runtime dependencies, and container images are fixed.
• **Cold-start rate, concurrency ceilings, memory allocation, CPU limits, and execution timeouts are held constant** across all cryptographic universes.
• Network routing, region placement, and availability zone selection are fixed to eliminate environmental variance.

These controls ensure that all observed differences in latency, cost, and scaling behavior can be causally attributed to cryptographic algorithm selection rather than environmental noise.

3.10 Summary

Control variables are enforced to achieve experimental validity and reproducibility between all experimental runs:

All cloud resources are provided via Infrastructure-as-Code (IaC), which ensures that the deployment configurations of experiments are identical.

Only one independent variable, cryptographic configuration. Substitutes All other software components, runtime dependencies, and container images are fixed.

Cold-start rate- Concurrency ceilings, memory allocation, CPU limits and execution timeouts are fixed across all cryptographic universes.

Network routing, region placement and availability zone selection are fixed to avoid environmental variance.

These checks are such that any differences in latency, cost and scaling behavior that are observed can be causally related to learning cryptographic algorithm choice instead of noise in the environment.

References

[1] **Montenegro, José A., Ruben Rios, and Javier Bonilla.**"Comparative Analysis of Post-Quantum Handshake Performance in QUIC and TLS Protocols." *Computer Networks*, vol. 275, 2026, article 111957, https://doi.org/10.1016/j.comnet.2025.111957.

[2] **Ukwuoma, Henry Chima, Arome, Gabriel, Thompson, Aderonke and Alese, Boniface Kayode**. "Post-quantum cryptography-driven security framework for cloud computing" *Open Computer Science*, vol. 12, no. 1, 2022, pp. 142-153. https://doi.org/10.1515/comp-2022-0235

[3] **Dimitrios Sikeridis, Panos Kampanakis, Michael Devetsikiotis.** "Post-Quantum Authentication in TLS 1.3: A Performance Study.", https://dx.doi.org/10.14722/ndss.2020.24203

[4]**Yaser Baseri, Abdelhakim Hafid, Arash Habibi Lashkari, "**Future-Proofing Cloud Security Against Quantum Attacks: Risk, Transition, and Mitigation Strategies", https://doi.org/10.48550/arXiv.2509.15653

[5] **Henrich, J., Heinemann, A., Wiesmaier, A., Schmitt, N**. (2023). Performance Impact of PQC KEMs on TLS 1.3 Under Varying Network Characteristics. In: Athanasopoulos, E., Mennink, B. (eds) Information Security. ISC 2023. Lecture Notes in Computer Science, vol 14411. Springer, Cham. https://doi.org/10.1007/978-3-031-49187-0_14

[6] **Parameswara Reddy Nangi, Chaithanya Kumar Reddy Nala Obannagari.** "Scalable End-to-End Encryption Management Using Quantum-Resistant Cryptographic Protocols for Cloud-Native Microservices Ecosystems", https://doi.org/10.63282/3050-9246.IJETCSIT-V4I1P116

[7] **Sai Srinivas Matta1, Manish Bolli.** "POST-QUANTUM CRYPTOGRAPHY FRAMEWORKS FOR SECURING GLOBAL CLOUD SYSTEMS", Doi: 10.63125/dhbvry98

[8] **Sosnowski, Markus, et al.** "The Performance of Post-Quantum TLS 1.3." *Companion of the 19th International Conference on Emerging Networking Experiments and Technologies (CoNEXT 2023)*, Association for Computing Machinery, 2023, pp. 19–27. DOI: 10.1145/3624354.3630585.

[9] **Jakkaraju, Venkata Thej Deep.** "Post-Quantum Cryptography Integration in CI/CD Pipelines: Future-Proofing Software Supply Chains." *Computer Fraud & Security*, 2024, pp. 457–467.

[10] Souvatzidaki, K., & Limniotis, K. (2025). Post-Quantum Key Exchange in TLS 1.3: Further Analysis on Performance of New Cryptographic Standards. *Cryptography*, *9*(4), 73. https://doi.org/10.3390/cryptography9040073

[11] **Panos Kampanakis, Will Childs-Klein.** "The impact of data-heavy, post-quantum TLS 1.3 on the Time-To-Last-Byte of real-world connections", Amazon Web Services, https://dx.doi.org/10.14722/madweb.2024.23010