The Triple Ratchet Protocol: A Bandwidth Efficient Hybrid-Secure Signal Protocol

Benedikt Auerbach
PQShield
Vienna, Austria
benedikt.auerbach@pqshield.com

Yevgeniy Dodis*
New York University
New York, USA
dodis@cs.nyu.edu

Daniel Jost
New York University
New York, USA
daniel.jost@cs.nyu.edu

Shuichi Katsumata
PQShield
AIST
Tokyo, Japan
shuichi.katsumata@pqshield.com

Thomas Prest
PQShield
Paris, France
thomas.prest@pqshield.com

Rolfe Schmidt
Signal Messenger
Albuquerque, USA
rolfe@signal.org

1 Executive Summary

Secure Messaging apps have seen growing adoption, and are used by billions of people daily. However, due to imminent threat of a "Harvest Now, Decrypt Later" attack, secure messaging providers must react know in order to make their protocols *hybrid-secure*: at least as secure as before, but now also post-quantum (PQ) secure. Since many of these apps are internally based on the famous Signal's Double-Ratchet (DR) protocol, making Signal hybrid-secure is of great importance.

In fact, Signal and Apple already put in production various Signal-based variants with certain levels of hybrid security: PQXDH (only on the initial handshake), and PQ3 (on the entire protocol), by adding a *PQ-ratchet* to the DR protocol. Unfortunately, due to the large communication overheads of the Kyber scheme used by PQ3, real-world PQ3 performs this PQ-ratchet approximately every 50 messages. As we observe, the effectiveness of this amortization, while reasonable in the best-case communication scenario, quickly deteriorates in other still realistic scenarios; causing *many consecutive* (rather than 1 in 50) re-transmissions of the same Kyber public keys and ciphertexts (of combined size 2272 bytes!).

In this presentation, we will talk about a new Signal-based, hybrid-secure secure messaging protocol, which significantly reduces the communication complexity of PQ3. We call our protocol "the *Triple Ratchet*" (TR) protocol. First, TR uses *erasure codes* to make the communication inside the PQ-ratchet provably balanced. This results in much better *worst-case* communication guarantees of TR, as compared to PQ3. Second, we design a novel "variant" of Kyber, called Katana, with significantly smaller combined length of ciphertext and public key (which is the relevant efficiency measure for "PQ-secure ratchets"). For 192 bits of security, Katana improves this key efficiency measure by over 37%: from 2272 to 1416 bytes. In doing so, we identify a critical security flaw in prior suggestions to optimize communication complexity of lattice-based PQ-ratchets, and fix this flaw with a novel proof relying on the recently introduced hint-MLWE assumption.

This protocol has been developed with the Signal team, and they are actively evaluating bringing a variant of it into production in a future iteration of the Signal protocol.

*Designated speaker

2 Extended Abstract

The Signal Protocol, used by Signal, WhatsApp, Google RCS, and Facebook Messenger to protect the communications of billions of people worldwide, has widely been considered to be a benchmark for secure messaging. At its core, it uses a famous *Double Ratchet* protocol [16] to provide important security properties called forward secrecy (FS) and post-compromise security (PCS). Signal (and the Double Ratchet protocol) has been widely deployed with heavily scrutinized open source implementations, and has been formally analyzed in [1, 3, 5, 7, 8, 12].

2.1 Post-Quantum Security

While this gives us confidence in the protocol today, these security guarantees are contingent on Diffie-Hellman (DH) assumptions for elliptic curves that can be broken by a quantum computer using Shor's algorithm [19]. This is not only a future threat, since protocol transcripts collected today can be recorded and saved until a quantum computer is available, then decrypted in a Harvest Now, Decrypt Later (HNDL) attack. Motivated by these concerns, the work by Alwen et al. [1] showed how to generalize the Signal protocol to work with any key encapsulation mechanism (KEM). As a result, one could potentially replace the DH-based Signal with a post-quantum variant; for example, using recently standardized Kyber (i.e., ML-KEM) [18]. Unfortunately, the resulting protocol is not sufficient for practical use, for two reasons. First, we do not want to lose the original DH-based security of Signal. Thus, practically relevant post-quantum extensions of Signal should provide what is called *hybrid* security, and meaningfully combine the DH-based Double Ratchet with some post-quantum variant. Second, the use of Kyber has noticeable costs in the communication complexity, making it often impractical in the real world.

2.2 PQXDH and PQ3

As a result, the industry transition to post-quantum Signal has been somewhat slower. First, Signal Messenger recently deployed PQXDH [13], an update to the X3DH [17] handshake component of the Signal Protocol, and formally verified that the updated protocol provides HNDL protection without removing any of the previous DH-based security guarantees [4]. Since this was only an update

to the initial protocol handshake, it does not provide any post-quantum PCS, one of the key features of the original Double Ratchet protocol.

To address this issue, Apple recently deployed PQ3 [2], — a protocol similar to Signal, — that continuously adds Kyber-768 freshly shared secrets to the "root secrets" of the Double Ratchet protocol. Simplifications of the resulting PQ3 protocol have been analyzed by [20] and machine verified by [15], but they do not fully capture what is done in the real world. Concretely, [20] only models Kyber public keys and ciphertexts as being sent with *every* asymmetric ratchet message. As we mentioned above, this is quite expensive, and Apple decided to perform a post-quantum ratchet approximately every 50 messages (or whenever they have not sent a fresh Kyber public key within a week), in order to amortize the large communication cost of Kyber keys and ciphertexts [10]. Heuristically (and somewhat oversimplifying), this means that users have 50 "cheap" epochs (which do not help with post-quantum PCS), followed by 1 "expensive" epoch (which gives post-quantum PCS, but at a much slower rate than DH-based PCS).[1]

2.3 Communication Efficiency of PQ3

While the deployment of PQ3 was an amazing, and greatly celebrated advance of post-quantum cryptography in the real-world, there are at least two avenues where it can be substantially improved in terms of its communication efficiency. (And we address these deficiencies in this work, as our main contribution.)

First, while PQ3's "amortization trick" might provide a reasonable trade-off in the best-case scenario, when the communication pattern between the users is roughly balanced, the effectiveness of this amortization quickly deteriorates in less balanced, but *still realistic* real-world scenarios. This is because each of Signal's sending epochs lasts roughly until the peer responds (and advances the public ratchet). So it might be possible — and certainly happens from time to time — that the "expensive epoch" happens exactly when one of the users is offline for an extended period of time,[2] resulting in *many consecutive re-transmissions repeating the same (long!) Kyber public keys and ciphertexts.*

Second, we already mentioned that Kyber's public key and ciphertext (and each "expensive epoch" message in PQ3 sends both) is much larger than the single DH group element sent by classical Signal. Concretely, (1088+1184=2272) bytes compared to 32 bytes, which is 71 times longer! Thus, any concrete efficiency improvement over using the generic (post-quantum) KEM advocated by [1] will likely result in much faster PCS. For example, it allows reduction of the number 50 in PQ3's heuristic amortization, while maintaining similar communication complexity. In that regard, [1, 9, 14] already described lattice-based protocols (either directly for Kyber, or equivalent variants over other rings) which seemingly achieve this goal. Unfortunately, the protocol of [9] achieves almost no saving (less than 2%, as noticed by the authors) as compared to using the generic Kyber, while the protocols of [1, 14] contain a critical subtle security flaw (which we found in this work) invalidating these analyses. Thus, prior to this work we did not have

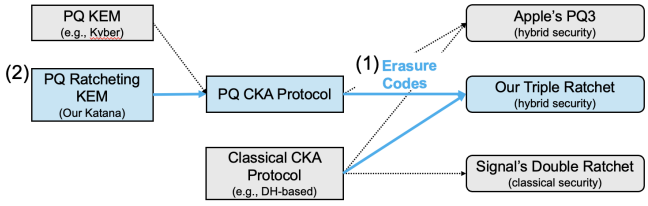

Figure 1: Different types of PQ KEM can be compiled into a PQ CKA protocol with different security and efficiency profiles. A classical CKA protocol leads to Signal's Double Ratchet protocol. This classical CKA and a PQ CKA based on Kyber, combined in a natural manner, leads to Apples's PQ3. Instead, using a PQ CKA based on Katana and performing erasure encoding for the hybrid composition leads to our Triple Ratchet protocol. The blue boxes and arrows indicate our construction.

optimized variants of Kyber which would significantly reduce the communication complexity of post-quantum Signal or its variants.

2.4 Our Contributions

In this work, we provide a practical hybrid-secure Double Ratchet protocol called the *Triple Ratchet* protocol.[3] Our name is taken from the fact that we use (i) a post-quantum public ratchet, (ii) a classical public ratchet, and (iii) symmetric ratchet. Compared to PQ3, it addresses both of the communication deficiencies mentioned above. An overview of our result is given in Figure 1. At a high level, our work consists of two technical contributions.

Contribution (1). First, it uses *erasure codes* to evenly distribute the communication inside the "post-quantum" ratchet (i.e., PQ CKA protocol in Figure 1), without any amortization heuristics. This is illustrated as Item (1) in Figure 1. At a high level, instead of sending one long message every 50 epochs, we encode the resulting message using an erasure code, and send a fresh chunk of this encoding with every message. For example, we could set parameters so that the long message will be decoded from *any* 50 chunks. Then, in a fully balanced setting we would still achieve PCS in 50 epochs and same communication as PQ3, but without any amortization. However, we start getting big savings in the unbalanced cases, when some epochs are long-lasting. For such epochs, PQ3's strategy could be viewed as using a hugely inefficient *repetition code*, leading to a big communication penalty; e.g., a factor of up to 50 in our "PQ3-inspired" example. We detail this and give an overview of some of the technical challenges we resolved in our presentation.

Contribution (2). Second, we design a novel *Continuous Key Agreement* (CKA) protocol based on Kyber, which we call Katana-CKA, which could be used inside our Triple Ratchet protocol. This is illustrated as Item (2) in Figure 1. Recall, CKA was a generic building block used by [1] to abstract out the design of the Double Ratchet Protocol. [1] then presented a generic KEM-based CKA, where every message contained a KEM public key and ciphertext. When applied to Kyber at security level 192 bits, this gives CKA *messages of size 2272 bytes*. In contrast, for the same security level Katana-CKA uses *messages of size 1416 bytes*, saving over 37% over the generic construction.

[1] This heuristics is related to "on-demand" ratcheting suggested by [6].
[2] E.g., when using devices which are periodically turned off.
[3] This should not be confused with the protocol by [5] with the same name.

We notice that Katana-CKA is closely related to what previous works called "optimized" lattice-based CKA [1, 14], but instantiated with a carefully chosen variant of Kyber. As we mentioned, however, we identify a critical flaw in the previous analyses of this "optimized" KEM, and non-trivially fix them with a novel proof relying on the recently introduced hint-MLWE assumption [11].

In more detail, we first generalize the KEM-based CKA from [1] to work with what we call a *Ratcheted* KEM (RKEM). On a high level, RKEM abstracts KEM properties in a way which allows a freshly sampled ciphertext also be used as "part" of a different KEM public key. In essence, this is precisely why the original DH-based CKA of Signal saved a factor of 2 in communication, when compared to the generic KEM-based DH construction. And this is why RKEM is precisely fitted for the use inside a CKA. Once we define RKEM and show that it generically implies CKA, it allows us to focus on a cleaner RKEM primitive, which we then construct from the hint-MLWE assumption. We call the resulting RKEM Katana,[4] which explains the name Katana-CKA for our new CKA. We expand on our technique in the presentation.

Lastly, we wrap up by providing an efficiency analysis of our Triple Ratchet protocol by comparing it with Apple's PQ3 and a variant of our Triple Ratchet instantiated with Kyber (i.e., we use the standard PQ KEM to construct the PQ CKA protocol in Figure 1). The latter variant illustrates the effectiveness of only relying on erasure codes. The efficiency comparison is found in Table 1. Here, we assume a simple model of unbalanced communication where every sender has a probability p of sending another message before receiving all incoming messages, independent of previous events. In row one we use $p = 0$ to capture perfectly balanced communication. In row two we use $p = 0.5$ to conservatively approximate the sending behavior of two online parties using typing indicators and read receipts, and we see that at this point both TR instantiations have an advantage over PQ3. Finally, in row 3, we use $p = 0.9$ to approximate the behavior of a device that is offline for hours at a time, where PQ3 is more than 4 times as expensive as TR with Katana. The talk will include more discussion on our efficiency analysis and the tradeoff between security.

	PQ3	TR with Kyber-768	TR with Katana (λ = 192)
$p = 0$	8 144	11 270	8 722
$p = 0.5$	12 760	11 615	8 989
$p = 0.9$	49 688	14 375	11 125

Table 1: Expected communication cost in bytes to attain PCS for PQ3 and TR. See text for the parameter p. PQ3 is assumed to send two Kyber-768 encapsulation keys and ciphertexts every 50 messages. TR with Kyber-768 (resp. Katana) uses a post-quantum CKA based on Kyber-768 (resp. Katana with $\lambda = 192$). This includes base message cost of 36B for PQ3 and 46B for TR to account for the overhead of sending counters and DH keys but excludes the 64B signature used by PQ3 for fair comparison.

References

[1] J. Alwen, S. Coretti, and Y. Dodis. The double ratchet: Security notions, proofs, and modularization for the Signal protocol. In Y. Ishai and V. Rijmen, editors, *EUROCRYPT 2019, Part I*, volume 11476 of *LNCS*, pages 129–158, Darmstadt, Germany, May 19–23, 2019. Springer, Cham, Switzerland.

[2] Apple Security Engineering and Architecture (SEAR). iMessage with PQ3: The new state of the art in quantum-secure messaging at scale, 2 2024. Available at https://security.apple.com/blog/imessage-pq3/.

[3] K. Bhargavan, A. Bichhawat, Q. H. Do, P. Hosseyni, R. Küsters, G. Schmitz, and T. Würtele. DY*: A modular symbolic verification framework for executable cryptographic protocol code. In *2021 IEEE European Symposium on Security and Privacy*, pages 523–542, Vienna, Austria, Sept. 6–10, 2021. IEEE Computer Society Press.

[4] K. Bhargavan, C. Jacomme, F. Kiefer, and R. Schmidt. Formal verification of the PQXDH post-quantum key agreement protocol for end-to-end secure messaging. In D. Balzarotti and W. Xu, editors, *USENIX Security 2024*, Philadelphia, PA, USA, Aug. 14–16, 2024. USENIX Association.

[5] A. Bienstock, J. Fairoze, S. Garg, P. Mukherjee, and S. Raghuraman. A more complete analysis of the Signal double ratchet algorithm. In Y. Dodis and T. Shrimpton, editors, *CRYPTO 2022, Part I*, volume 13507 of *LNCS*, pages 784–813, Santa Barbara, CA, USA, Aug. 15–18, 2022. Springer, Cham, Switzerland.

[6] A. Caforio, F. B. Durak, and S. Vaudenay. Beyond security and efficiency: On-demand ratcheting with security awareness. In J. Garay, editor, *PKC 2021, Part II*, volume 12711 of *LNCS*, pages 649–677, Virtual Event, May 10–13, 2021. Springer, Cham, Switzerland.

[7] R. Canetti, P. Jain, M. Swanberg, and M. Varia. Universally composable end-to-end secure messaging. In Y. Dodis and T. Shrimpton, editors, *CRYPTO 2022, Part II*, volume 13508 of *LNCS*, pages 3–33, Santa Barbara, CA, USA, Aug. 15–18, 2022. Springer, Cham, Switzerland.

[8] K. Cohn-Gordon, C. Cremers, B. Dowling, L. Garratt, and D. Stebila. A formal security analysis of the Signal messaging protocol. *Journal of Cryptology*, 33(4):1914–1983, Oct. 2020.

[9] N. Drucker and S. Gueron. Continuous key agreement with reduced bandwidth. In *International Symposium on Cyber Security Cryptography and Machine Learning*, pages 33–46. Springer, 2019.

[10] F. Jacobs. Designing imessage pq3: Quantum-secure messaging at scale. Invited talk at the Real World Crypto Symposium 2025, 2024.

[11] D. Kim, D. Lee, J. Seo, and Y. Song. Toward practical lattice-based proof of knowledge from hint-MLWE. In H. Handschuh and A. Lysyanskaya, editors, *CRYPTO 2023, Part V*, volume 14085 of *LNCS*, pages 549–580, Santa Barbara, CA, USA, Aug. 20–24, 2023. Springer, Cham, Switzerland.

[12] N. Kobeissi, K. Bhargavan, and B. Blanchet. Automated verification for secure messaging protocols and their implementations: A symbolic and computational approach. In *2017 IEEE European Symposium on Security and Privacy, EuroS&P 2017, Paris, France, April 26-28, 2017*, pages 435–450. IEEE, 2017.

[13] E. Kret and R. Schmidt. The pqxdh key agreement protocol, 2023. Available at https://signal.org/docs/specifications/pqxdh/.

[14] J. Lee, J. Kwon, and J. S. Shin. Efficient continuous key agreement with reduced bandwidth from a decomposable kem. *IEEE Access*, 11:33224–33235, 2023.

[15] F. Linker, R. Sasse, and D. Basin. A formal analysis of apple's iMessage PQ3 protocol. Cryptology ePrint Archive, Paper 2024/1395, 2024.

[16] M. Marlinspike and T. Perrin. The double ratchet algorithm, 2016. Available at https://signal.org/docs/specifications/doubleratchet/.

[17] M. Marlinspike and T. Perrin. The x3dh key agreement protocol, 2016. Available at https://signal.org/docs/specifications/x3dh/.

[18] P. Schwabe, R. Avanzi, J. Bos, L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, J. M. Schanck, G. Seiler, D. Stehlé, and J. Ding. CRYSTALS-KYBER. Technical report, National Institute of Standards and Technology, 2022. available at https://csrc.nist.gov/Projects/post-quantum-cryptography/selected-algorithms-2022.

[19] P. W. Shor. Algorithms for quantum computation: Discrete logarithms and factoring. In *35th FOCS*, pages 124–134, Santa Fe, NM, USA, Nov. 20–22, 1994. IEEE Computer Society Press.

[20] D. Stebila. Security analysis of the iMessage PQ3 protocol. Cryptology ePrint Archive, Report 2024/357, 2024.

[4] Similar to Kyber, Katana is a certain type of an ancient (Japanese) sword.