Beiträge aus der Informationstechnik

Mobile Nachrichtenübertragung
Nr. 91

Friedrich Pauls

Aspects of Latency Optimization for Hash-based Digital Signatures

VOGT

Dresden 2021

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Zugl.: Dresden, Techn. Univ., Diss., 2021

Die vorliegende Arbeit stimmt mit dem Original der Dissertation „Aspects of Latency Optimization for Hash-based Digital Signatures“ von Friedrich Pauls überein.

Gesetzt vom Autor

ISBN 978-3-95947-049-0

Jörg Vogt Verlag
Niederwaldstr. 36
01277 Dresden
Germany

Phone: +49-(0)351-31403921
Telefax: +49-(0)351-31403918
e-mail: info@vogtverlag.de
Internet : www.vogtverlag.de

Technische Universität Dresden

Aspects of Latency Optimization for Hash-based Digital Signatures

Friedrich Pauls

von der Fakultät Elektrotechnik und Informationstechnik der Technischen Universität Dresden

zur Erlangung des akademischen Grades

Doktoringenieur
(Dr.-Ing.)

genehmigte Dissertation

Vorsitzender:	Prof. Dr. rer. nat. Stefan Mannsfeld
Gutachter:	Prof. Dr.-Ing. Dr. h.c. Gerhard Fettweis Prof. Dr. sc.techn. Andreas Herkersdorf
Tag der Einreichung:	15.01.2021
Tag der Verteidigung:	12.05.2021

Friedrich Pauls
Aspects of Latency Optimization for Hash-based Digital Signatures
Dissertation, 12.05.2021

Vodafone Stiftungslehrstuhl für Mobile Kommunikationssysteme
Institut für Nachrichtentechnik
Fakultät Elektrotechnik und Informationstechnik
Technische Universität Dresden
01062 Dresden, Germany

Abstract

The Tactile Internet will enable interactive real-time applications in industry and society. Proposed applications require, first, an end-to-end latency of about one millisecond, and second, stringent security. Data authentication is an important pillar for this security objective. Digital signatures provide data authentication, but with quantum computers looming at the horizon, most currently employed signature schemes will become unsecure. Hash-based digital signatures have been proposed as post-quantum secure alternative. These schemes are computationally expensive, leading to processing times that contradict the low-latency requirements of the Tactile Internet. This thesis develops a framework to understand, assess, and minimize the end-to-end latency of selected hash-based signatures.

We argue for the use of the eXtended Merkle Signature Scheme (XMSS), a post-quantum secure algorithm with well-established security margins. For different architectural classes, we establish a performance baseline and assess the gap between our targeted latency budget of $100\,\mu$s for data authentication and state-of-the-art XMSS implementations. High-end desktop processors utilizing vector extensions are above the target by a factor of 33. A hardware accelerator for embedded applications employing a hardware/software co-design is above target by a factor of 270. Pure software implementations on embedded processors miss by three orders of magnitude.

A thorough bottleneck analysis using our enhanced call graph method identifies computational intensive parts of the algorithm, and allows us to focus our later optimizations on the critical processing steps: the authentication path computation, the production of hashes and hash chain nodes, and the processing of hash chains. A comprehensive data dependency analysis studies which XMSS operations can be processed independently. We find that the authentication path computation can be taken out of the critical latency path altogether. By exploiting this, our developed *EarlyAuth* scheme improves the latency significantly and achieves a speedup of about 4.23 as compared to the XMSS reference implementation.

As software optimizations are insufficient to achieve the required performance, we design an application-specific integrated processor based on the Cadence Tensilica LX6 architecture. Using a hardware/software-codesign approach, we develop a generic accelerator for the underlying hash function which constitutes the main bottleneck of the XMSS signing and verification procedures. Although a speedup of 183 over the reference implementation could be achieved, with $440\,\mu$s for signing alone, this generic architecture fails to meet our latency objective. Our instruction histogram analysis reveals that memory accesses require 51 percent of all cycles of the hash chain node computation. To mitigate these memory loads, we refine

the architecture to take the specifics of XMSS into account. This improved design employs a result shift mechanism, a hardware padding generator, and tailored buffers designed to minimize memory operations. Although the cycles for memory loads are reduced to 7 percent, the performance is still insufficient. Further analysis reveals that the hash permutation function emerges as the new bottleneck and requires over 79 percent of the chain node computation cycles. We alleviate this by introducing a hardware unrolling scheme and optimize the trade-offs with respect to area-time and area-time-energy products. Our fastest architecture achieves a minimal total processing time for signing and verification of $265\,\mu$s. Together with the *EarlyAuth* scheme, latencies below $100\,\mu$s become viable.
Finally, we develop an end-to-end latency model that not only takes the signing and verification times into account, but also the data transmission times of message and signature. With this model, we analyze achievable end-to-end latencies and perform XMSS parameter optimizations. In addition, we derive a lower bound for the achievable XMSS end-to-end latency as a tool for design space exploration. Our derived latency efficiency metric relates the latency achieved by an actual design to the lower bound. This allows for comparison of different XMSS implementations.

We present a comprehensive set of analysis tools, implementation methods, and models, that aim to optimize the end-to-end latency of data authentication from an application perspective. Only the combination of all developed software and hardware optimizations makes it possible to achieve our latency objective. We improve the state of the art by up to two orders of magnitude and widen the envelope of viable secure applications for the Tactile Internet.

Kurzfassung

Das Taktile Internet wird interaktive Echtzeitanwendungen in Industrie und Gesellschaft ermöglichen. Vorgeschlagene Anwendungen erfordern erstens, eine Ende-zu-Ende-Latenz von etwa einer Millisekunde und zweitens, eine sehr hohe Sicherheit. Die Datenauthentifizierung ist ein wichtiger Grundpfeiler für dieses Sicherheitsziel. Datenauthentifizierung wird mittels digitaler Signaturen sichergestellt, jedoch wird ein Großteil der derzeit verwendeten Signaturverfahren durch die sich abzeichnenden zukünftigen Quantencomputer gebrochen werden. Hash-basierte digitale Signaturen gelten als post-quanten-sichere Alternative, da Quantencomputer hier ihre Vorteile nicht ausspielen können. Diese Verfahren sind rechenintensiv, was zu Verarbeitungszeiten führt, die im Widerspruch zu den Latenzanforderungen des taktilen Internets stehen. Die vorliegende Arbeit entwickelt ein Framework, um die Ende-zu-Ende-Latenz zu verstehen, zu bewerten und sie für ausgewählte Hash-basierte Signaturen zu minimieren.

Wir plädieren für die Verwendung des eXtended Merkle Signature Scheme (XMSS), einem post-quanten-sicheren Algorithmus mit gut erforschten Sicherheitsmargen. Für verschiedene Architekturklassen ermitteln wir einen Performanz-Referenzwert und bewerten die Lücke zwischen unserem angestrebten Latenzbudget von $100\,\mu\mathrm{s}$ für die Datenauthentifizierung und dem Stand dem Stand der Technik von XMSS-Implementierungen. High-End-Desktopprozessoren, die Vektorerweiterungen nutzen, liegen um den Faktor 33 über dem Zielwert. Ein Hardwarebeschleuniger für eingebettete Anwendungen, der ein Hardware/Software-Co-Design verwendet, liegt um den Faktor 270 über der Ziellatenz. Reine Softwareimplementierungen für eingebetteten Prozessoren verfehlen das Ziel um drei Größenordnungen.

Eine ausführliche Engpassanalyse identifiziert mithilfe unserer erweiterten Call-Graph-Methode rechenintensive Teile des Algorithmus und ermöglicht es uns, unsere späteren Optimierungen auf die kritischen Verarbeitungsschritte zu konzentrieren: die Berechnung des Authentifizierungspfads, die Erzeugung von Hashes und Hashkettenknoten und die Verarbeitung von Hashketten.

Eine umfassende Datenabhängigkeitsanalyse untersucht, welche XMSS-Operationen unabhängig voneinander verarbeitet werden können. Wir stellen fest, dass die Berechnung des Authentifizierungspfads aus dem kritischen Latenzpfad heraus genommen werden kann. Indem wir dies ausnutzen, verbessert unser entwickeltes *EarlyAuth*-Schema die Latenzzeit signifikant und erreicht einen Speedup von etwa 4,23 im Vergleich zur XMSS-Referenzimplementierung.

Da Softwareoptimierungen allein nicht ausreichen, um die geforderte Leistung zu erreichen, entwerfen wir auf Basis der Cadence Tensilica LX6 Architektur einen anwendungsspezifischen integrierten Prozessor. Mithilfe eines Hardware/Software-

Codesign-Ansatzes entwickeln wir einen generischen Beschleuniger für die zugrundeliegende Hashfunktion, die den Engpass der XMSS-Signier- und Verifikationsoperationen darstellt. Obwohl ein Speedup von 183 gegenüber der Referenzimplementierung erreicht werden konnte, erfüllt diese generische Architektur, mit $440\,\mu$s für das Signieren allein, nicht unser Latenzziel. Unsere Instruktionshistogrammanalyse zeigt, dass Speicherzugriffe 51 Prozent aller Taktzyklen der Hashkettenknotenberechnung benötigen. Um diese Speicherbelastung zu verringern, verfeinern wir die Architektur unter Ausnutzung spezifischer Eigenheiten von XMSS. Dieses verbesserte Design verwendet einen Result-Shift-Mechanismus, einen Hardware-Padding-Generator und maßgeschneiderte Datenpuffer, um die Speicheroperationen zu minimieren. Obwohl die Zyklen für Speicherzugriffe auf 7 Prozent reduziert werden, ist die Leistung immer noch unzureichend. Eine weitere Analyse zeigt, dass sich die Hash-Permutationsfunktion als neuer Engpass herausstellt und über 79 Prozent aller Taktzyklen für die Berechnung der Kettenknoten benötigt. Dies können wir durch Einführung eines Hardware-Unrolling-Schemas abmildern und optimieren sich ergebende Parameter durch Abwägungen der Fläche$\times$Zeit- und Fläche$\times$Zeit$\times$Energie-Produkte. Unsere schnellste Architektur erreicht eine minimale Gesamtverarbeitungszeit für Signierung und Verifizierung von $265\,\mu$s. Zusammen mit dem *EarlyAuth*-Schema werden Latenzen unter $100\,\mu$s realisierbar. Zudem entwickeln wir ein Ende-zu-Ende-Latenz-Modell, das nicht nur die Signier- und Verifikationszeiten berücksichtigt, sondern auch die Datenübertragungszeiten von Nachricht und Signatur miteinbezieht. Mit diesem Modell analysieren wir die erreichbaren Ende-zu-Ende-Latenzen und führen Optimierungen der XMSS-Parameter durch. Darüber hinaus leiten wir eine untere Schranke für die erreichbare Ende-zu-Ende-Latenz von XMSS ab, die als Werkzeug für die Entwurfsraumexploration dient. Unsere abgeleitete Latenz-Effizienz-Metrik setzt die durch ein tatsächliches Design erreichte Latenz zur unserer hergeleiteten theoretischen unteren Schranke in Beziehung. Dies ermöglicht den Vergleich verschiedener XMSS-Implementierungen.

Wir präsentieren einen umfassenden Satz von Analysewerkzeugen, Implementierungsmethoden und Modellen, die darauf abzielen, die Ende-zu-Ende-Latenz der Datenauthentifizierung aus Anwendungsperspektive zu optimieren. Erst durch Kombination aller entwickelten Software- und Hardwareoptimierungen ist es möglich, unser Latenzziel zu erreichen. Wir verbessern den Stand der Technik um bis zu zwei Größenordnungen und erweitern so den Bereich der realisierbaren sicheren Anwendungen für das Taktile Internet.

Acknowledgement

The contributions of this work are based on the research work carried out during my time as a research associate at the Vodafone Chair Mobile Communications Systems at the Technische Universität Dresden. First of all, I wish to express my gratitude to Professor Gerhard Fettweis for giving me the opportunity and for continuously supporting my work. Without your motivation, your encouragement and your expertise, it wouldn't have been possible to write this thesis. I also want to thank my second reviewer Professor Andreas Herkersdorf for the interest in my research topic and for the valuable comments.

I would like to express my deepest gratitude to Dr. Tim Hentschel for the support and continuous guidance throughout the writing of my thesis. In addition, I would also like to thank Dr. Emil Matus, and my colleagues and friends from the chair. You have always supported and motivated me in my research. Finally, I could not have completed this work without the support of my friends Paul, Martin, and Sebastian, who provided stimulating discussions, motivation as well as happy distractions to the rest my mind outside of my research.

I also truly appreciate the patience and the trust of my family.

Contents

Introduction 1

The need for secure post-quantum digital signatures and implementations thereof, is an accepted fact in science and industry [BL17; SIT18; SM19]. Moreover, applications for the Tactile Internet, require sub-millisecond latencies and strong security guarantees [Fet+14; Sim+16]. Here, a challenge arises, to bring together the computationally expensive post-quantum signature schemes, and the required sub-millisecond latencies of Tactile Internet applications. But to start from the beginning, we first introduce what signatures are, why they are important, and what role latency – the notion of a time delay between two events – plays. Second, we stress the importance of data authentication in our digital world and shortly explain the Tactile Internet. Motivated by this, we derive the objective of the thesis and give an outline.

1.1 Motivation

This thesis is about signatures and their latency. Before diving deeper into the technical details, the subsequent paragraphs give the reader an intuitive understanding of what signatures do, and why it is sometimes important to provide these in a timely manner.

A message delivers information from a sender to a recipient. Messages can take different forms. They can be physical documents, like a letter or a contract, or they can be entirely digital, for example, an email, a file, or a protocol message between interacting computer systems.
Ideally, a message is relevant, meaning that it provides value to the recipient. How much value a message provides, depends not only on its information content, but to a large degree on whether it is genuine, authentic, and that the sender consents to the content. A signature convinces a recipient that a message has those properties. In a broader sense, a signature can take many forms: handwritten, digital, or as a seal or watermark, for instance, known from bank notes or concert tickets. For the following examples, consider the difference in value with and without a valid signature: A concert ticket bought on Ebay, a bank loan contract, your new 3-year work contract, or a message from your boss directing you to issue a large financial transaction. In all cases, the value changes drastically depending on whether you can make sure that the document at hand is genuine, signed, or authentic. The latter example shows, that acting on a message which is not properly authenticated, can even put you at risk.

Another property influencing the value of a message is often its timeliness. What happens if a message is not delivered in a timely manner? Suppose you were able to acquire an expensive concert ticket to the very last concert of your favorite artist. The ticket is genuine. There is a union strike at the postal service and your ticket comes a day after the concert. Even though the ticket is genuine, what is it worth? Similarly, consider an autonomous vehicle that suddenly detects a dangerous situation – sudden ice, or a child running on the street. That vehicle now wants to signal following cars a warning, that it is going to make an emergency braking maneuver. Clearly, the faster the warning message is received and checked for validity, the more likely it is, that harm can be prevented. So, the value of a genuine message can degrade if it is not received within a particular time frame.

This work focuses on digital messages that need to be signed *and* delivered in time, in order to retain their value. The following section introduces why data authentication is important in our digital world and what challenges current digital signature schemes face. Thereafter, we shortly introduce the Tactile Internet, a new communication paradigm that requires both data authentication and ultra-low communication latencies.

Data Authentication – A Pillar of the Digital World

The digital world we live in today relies heavily on a variety of cryptography protocols, often unnoticed by the user. These protocols provide security to almost every digital transaction. This can be the connection to a website, a financial transaction, sending an email, a software update for your laptop or phone, or even a telephone call. Security encompasses both encryption, but first an foremost, data authentication. Digital signatures provide data authentication. They empower the receiver of a message to make sure that it comes from a trustworthy source (authenticity), and that it has not been modified in between (integrity). These properties are especially relevant in safety-critical use-cases, for example, in autonomous driving, or for emergency breaking messages.

Almost all traditional public key cryptography will become unsecure once large-scale quantum computers become viable. This includes currently used digital signature schemes like the Rivest-Shamir-Adleman (RSA) cryptosystem or the Elliptic Curve Digital Signature Algorithm (ECDSA) [BL17]. Cryptography algorithms that maintain their security properties against quantum computers are called *post-quantum secure*. Quantum computing seems to be always 10 years away from now. It is impossible to tell, when exactly quantum computers will be powerful enough to break current protocols. However, the hazard of such a machine is large

enough, and the time to migrate from current to post-quantum secure protocols is long enough, that the National Security Agency, the Central Security Service, and the National Academy of Sciences recommend to prepare for this transition [CSS16; SM19]. Additionally, proposed use cases, like fully autonomous driving, will also still need several years until deployed in large scale. Because of the stringent latency requirements, most security primitives need to be implemented with hardware support. This makes it infeasible to update devices and deployed infrastructure to post-quantum secure algorithms on a software basis.

One promising class of post-quantum secure data authentication schemes are hash-based digital signatures which have very reliable security estimates and are considered to be mature [NW17].

The Tactile Internet and the Need for Low-Latency

The paradigm of the Tactile Internet as evolution of the Internet of Things (IoT) will enable real-time interactive systems in a variety of different areas such as automation, transportation, gaming, education, and healthcare. Key requirements of the Tactile Internet are ultra-low latency, extremely high availability, reliability and security [Fet+14; Fet14]. It has been shown, that an end-to-end latency of about *one millisecond* is a requirement for applications in, for instance, autonomous driving, vehicle-to-infrastructure communication, control communications, or in other forms of human-to-machine and machine-to-machine communications [Sim+16].

1.2 Thesis Objective

The need for this work arises for two reasons: (1) Data authentication is a key requirement for most Tactile Internet applications, as they are often safety- and security-critical. However, currently used cryptographic algorithms will be broken, once large scale quantum computers arrive. Hash-based digital signatures schemes have been proposed as post-quantum secure alternative, but they demand for large signatures and high processing power. (2) Many Tactile Internet applications require end-to-end latencies of about *one millisecond* to become viable. The high processing demand of hash-based digital signature schemes contradict the low-latency objective and create a tension field which this thesis explores.

Since we are dealing with an end-to-end latency, the available millisecond must be reasonably split among different time consuming procedures: The acquisition of sensor data, processing at the receiver, message signing, modulation, signal

propagation, demodulation, message verification, processing at the receiver, and the final activation of an actor. Consequently, only a fraction of the total time can be used for data authentication, and we assume this budget to be $100\,\mu$s.

The central research question is to examine, whether it is feasible to provide post-quantum secure data authentication, using hash-based digital signatures, within an assumed time budget of $100\,\mu$s. To tackle this topic, we must approach the problem from different viewpoints. Thus, our study is based on the three research domains shown in Figure 1.1.

First, the domain of **applied cryptography** answers questions as to what authentication schemes are suitable, and which signature algorithms and hash functions to use.

Second, as state-of-the-art implementations do not achieve the required performance [OLC17; Wan+18], we must design an architecture that provides the required processing power and efficiency. We follow a **hardware/software-codesign** approach which takes the intertwining of software and hardware into account. This includes a bottleneck analysis, designing an Application-specific Instruction Set Processor (ASIP), iterative hardware/software optimization, and a design evaluation based on typical metrics, such as the area-time (AT) product.

Finally, an **application-oriented design space exploration** uses a modeling approach and puts our design into an end-to-end system context. The goal is to study the interaction between signing and verification times, and the data transmission time, which has not been studied before on that level. This helps to determine achievable end-to-end latencies and to optimize parameters globally. This optimization profits from a study of the relationships between algorithm and use-case parameters, such as the network data rate. Lastly, we explore limitations and provide a metric to compare different implementations and architectures.

As a result, this thesis develops a framework of analysis tools, implementation methods, and models, that aim to optimize the end-to-end latency of post-quantum secure data authentication procedures for the Tactile Internet.

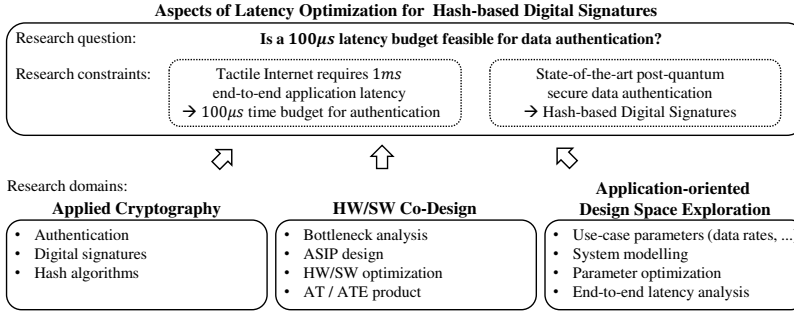

Fig. 1.1. Central research question, constraints and domains.

1.3 Thesis Overview

Chapter 2 provides a brief introduction to digital signatures followed by a qualitative comparison of post-quantum digital signature schemes. Next, we explain the principle of hash-based digital signatures and why they are viable candidates for our intended use-case. Finally, we introduce the eXtended Merkle Signature Scheme (XMSS) which is a particular instance of a hash-based digital signature scheme. We argue that XMSS is a suitable candidate for our use case, and why we have chosen the SHA-3 hash function family for our implementation.

Chapter 3 gives an overview of prior art and establishes a performance baseline by comparing current XMSS implementations in three architectural classes. It follows an assessment of the performance gap between our latency objective and current state-of-the-art designs. Finally, a thorough bottleneck analysis using our enhanced call graph method identifies computational intensive parts of the algorithm.

Chapter 4 analyses the internal dependencies of the XMSS signing and verification procedures on a functional level. It allows to identify operations that can be scheduled in a manner to improve the end-to-end latency from an application perspective. In addition, the dependency analysis helps to focus our later optimizations on functions that are within the critical latency path. Further, we develop and evaluate a scheme that uses a favorable processing sequence to optimize the latency.

Chapter 5 focuses on the acceleration of critical functions which have been identified in the previous chapters. It explains what an ASIP is, the method of hardware acceleration we use, and why we use it. As the hash function is the major bottleneck of the XMSS operation, first, we introduce our generic hash function accelerator

that supports the KECCAK hash function family, which we use for the XMSS algorithm. Next, we further refine this architecture by exploiting specifics of the XMSS algorithm, in particular, the processing of the hash chain node computations. We introduce a hardware padding generator, a results shift mechanism, and tailored buffers, which minimize memory operations and thus improve the overall latency. Thereafter, the permutation function remains as the major bottleneck. We optimize this by introducing a hardware unrolling scheme and optimize trade-offs with respect to area-time (AT) and area-time-energy (ATE) products. It follows a study on the influence the data memory width has on the latency. Finally, a system level performance analysis gives insights on achievable processing times for key, generation, signing, verification, and single hash performance.

Chapter 6 develops an end-to-end latency model which accounts not only for signing and verification times, but also for the data transmission time of message and signature. This is necessary for two reasons: First, because of the large signatures, the data transmission time is a relevant part in our targeted latency range. Second, the signature size is a function of the XMSS parameters, and thus the model helps to analyze the achievable end-to-end performance and to optimize relevant XMSS parameters. First, we develop models for the XMSS signing and verification times and fit model parameters to our developed hardware architecture. Second, we introduce our model that estimates the data transmission time. Finally, the models are combined and provide estimates on achievable end-to-end latencies. Thereafter, we develop a lower bound for achievable end-to-end latency as a tool for a parameter or design space exploration. Furthermore, we derive a latency efficiency metric which relates the actual achievable latencies to the lower bound. This allows to rate and compare different XMSS implementations.

Chapter 7 highlights conclusions and gives recommendations for future work.