Efficient Algorithms for Supersingular Isogeny Diffie-Hellman

Craig Costello(✉), Patrick Longa, and Michael Naehrig

Microsoft Research, Redmond, USA
{craigco,plonga,mnaehrig}@microsoft.com

Abstract. We propose a new suite of algorithms that significantly improve the performance of supersingular isogeny Diffie-Hellman (SIDH) key exchange. Subsequently, we present a full-fledged implementation of SIDH that is geared towards the 128-bit quantum and 192-bit classical security levels. Our library is the first constant-time SIDH implementation and is up to 2.9 times faster than the previous best (non-constant-time) SIDH software. The high speeds in this paper are driven by compact, inversion-free point and isogeny arithmetic and fast SIDH-tailored field arithmetic: on an Intel Haswell processor, generating ephemeral public keys takes 46 million cycles for Alice and 52 million cycles for Bob, while computing the shared secret takes 44 million and 50 million cycles, respectively. The size of public keys is only 564 bytes, which is significantly smaller than most of the popular post-quantum key exchange alternatives. Ultimately, the size and speed of our software illustrates the strong potential of SIDH as a post-quantum key exchange candidate and we hope that these results encourage a wider cryptanalytic effort.

Keywords: Post-quantum cryptography · Diffie-Hellman key exchange · Supersingular elliptic curves · Isogenies · SIDH

1 Introduction

Post-quantum Cryptography. The prospect of a large scale quantum computer that is capable of implementing Shor's algorithm [43] has given rise to the field of post-quantum cryptography (PQC). Its goal is to develop and ultimately deploy cryptographic primitives that resist cryptanalysis by both classical *and* quantum computers. Recent developments in quantum computing (see, e.g., [16,23,34]) have helped catalyze government and corporate action in this arena. For example, in April 2015, the National Institute of Standards and Technology (NIST) held a "Workshop on Cybersecurity in a Post-Quantum World", reaching out to academia and industry to discuss potential future standardization of PQC. Later, in August 2015, the National Security Agency (NSA) released a major policy statement that announced plans to "transition to quantum resistant algorithms in the not too distant future" [35]. In February 2016, NIST published a draft "Report on Post-Quantum Cryptography" [11], which

M. Robshaw and J. Katz (Eds.): CRYPTO 2016, Part I, LNCS 9814, pp. 572–601, 2016.
DOI: 10.1007/978-3-662-53018-4_21

emphasizes the need to start working towards the deployment of post-quantum cryptography in our information security systems, and outlines NIST's plans to "initiate a standardization effort in post-quantum cryptography".

In terms of public-key PQC, there are four well-known and commonly cited classes of cryptographic primitives that are believed to remain secure in the presence of a quantum computer: code-based cryptography, lattice-based cryptography, hash-based cryptography, and multivariate cryptography. Specific examples for each of these are McEliece's code-based encryption scheme [29]; Hoffstein, Pipher and Silverman's lattice-based encryption scheme "NTRU" [21]; Merkle's hash-tree signatures [30]; and Patarin's "HFE^{v-}" signature scheme [38]. A positive trait shared by all of these examples is a resistance to decades of attempted classical and quantum cryptanalysis which has inspired widespread confidence in their suitability as a post-quantum primitive. However, most of these examples also share the trait of having enormous public key and/or signature sizes, particularly when compared to traditional primitives based on the hardness of integer factorization or (elliptic curve) discrete logarithm computation.

Supersingular Isogeny Diffie-Hellman. In this paper, we study a different primitive that does not fall into any of the above classes, but is currently believed to offer post-quantum resistance: the supersingular isogeny Diffie-Hellman (SIDH) key exchange protocol proposed by Jao and De Feo in 2011 [22]. The SIDH key exchange protocol is more than a decade younger than all of the above schemes, so its security is yet to withstand the tests of time and of a wide cryptanalytic effort. Nevertheless, the current picture of its security properties looks promising. The best known classical and quantum attacks against the underlying problem are both exponential in the size of the underlying finite field, and their complexities make current SIDH key sizes significantly smaller than their post-quantum key exchange and/or encryption counterparts[1].

Our Contributions. We present a full-fledged, high-speed implementation of (unauthenticated) ephemeral SIDH that currently provides 128 bits of quantum security and 192 bits of classical security. This implementation uses 48-byte private keys to produce 564-byte ephemeral Diffie-Hellman public keys, is written in C and includes an optimized version of the field arithmetic written in assembly. To our knowledge, our library (see [14]) presents the first SIDH software that runs in *constant-time*, i.e., that is designed to resist timing [26] and cache-timing [37] attacks. On x64 platforms, our implementation runs up to 2.9 times faster than the (previously fastest) implementation of SIDH by Azarderakhsh et al. [2]. Note that this performance comparison does not take into account the fact that the implementation from [2] is not protected against timing attacks. The main technical contributions that lead to these improvements are:

[1] An exception here is NTRUEncrypt [21], which has comparable public key sizes – see https://github.com/NTRUOpenSourceProject/ntru-crypto.

Projective Curve Coefficients. A widely-deployed technique in traditional ECC involves avoiding inversions by working with elliptic curve points in projective space. Following Jao and De Feo [22], we also employ this technique to work efficiently with points in $\mathbb{P}^1$ by making use of the fast arithmetic associated with the Kummer varieties of Montgomery curves. A crucial difference in this work, however, is that we also work projectively with the curve coefficients; unlike traditional ECC where the curve is fixed, every SIDH key exchange requires computations on many different isogenous curves. In Sect. 3 we show that the Montgomery model also allows all of the necessary isogeny arithmetic to be performed efficiently in $\mathbb{P}^1$. This gives rise to more compact algorithms, significantly simplifies the overall computation, and means that key generation and shared secret computations only require one and two field inversions, respectively.

Prime Selection and Tailor-Made Montgomery Multiplication. We select a prime with form $p = \ell_A^{e_A} \ell_B^{e_B} f - 1$, where $\ell_A = 2$, $\ell_B = 3$, and the bit lengths of 2^{e_A} and 3^{e_B} are slightly smaller than a multiple of 64. This supports efficient arithmetic on a wide range of platforms and allows access to a large variety of optimizations such as the efficient use of vector instructions, Karatsuba multiplication, and lazy reduction. Moreover, it is well-known that primes of a special form can lead to faster algorithms for computing modular arithmetic in comparison with general-purpose algorithms. In this work, we note the special shape of these *SIDH-friendly* primes and modify the popular Montgomery multiplication algorithm to speed up modular arithmetic.

Ground Field Scalar Multiplications for Key Generation. Secure key generation in the SIDH protocol requires the definition of two independent cyclic subgroups of a fixed order (see Sect. 2). Jao and De Feo [22, Sect. 4.1] propose that generators of these two groups can be computed by multiplying random curve points by an appropriate cofactor, and that their linear independence can be checked via the Weil pairing. In Sect. 4 we employ a well-known technique from the pairing literature [42, Sect. 5] to work with two advantageous choices of torsion subgroups: the *base-field* and *trace-zero* subgroups. These choices allow the initial scalar multiplications that are required during key generation to be performed entirely over the base field. While these scalar multiplications only constitute a small fraction of the overall key generation time, and therefore the overall speedup from this technique is only moderate, a more visible benefit is the significant decrease in the size of the public parameters – see Sect. 6. We discuss possible security implications of this choice in Sect. 4.

Several of the above choices not only aid efficiency, but also the overall simplicity and compactness of the SIDH scheme. Choosing to unify points with their inverses and to unify Montgomery curves with their quadratic twists (see Sect. 3) effectively compresses the elements that are sent over the wire, i.e., the public keys, by a factor of two. Moreover, our software never requires the computation of square roots.

The timings we present in Sect. 7 reveal that high-security SIDH key exchange is more efficient than it was previously known to be. Our constant-time software shows that, if confidence in the security of SIDH warrants real-world deployment

in the future, the same level of side-channel protection can be achieved in the SIDH setting as in traditional number-theoretic schemes. We therefore hope that this paper encourages a wider cryptanalytic effort on the problems underlying the security of SIDH (see Sect. 2). Moreover, even if cryptanalytic improvements are made in the future, the huge difference between current SIDH key sizes and those of other PQC primitives suggest that the problem could remain of interest to practitioners. So long as the best known attacks remain exponential with a reasonable exponent (see the discussion below), it is reasonable to suggest that elliptic curves could offer the same benefit in post-quantum cryptography that they did in classical cryptography.

Beyond the efficiency improvements above, we present several techniques that help to bridge the gap between the theoretical SIDH scheme in [22] and its real-world deployment. Of particular importance are the contributions discussed in the following two paragraphs.

A Strong ECDH + SIDH Hybrid. Given the uncertainty surrounding the arrival date of large-scale quantum computers (as well as the time it takes for new primitives to be thoroughly cryptanalyzed, standardized and deployed), many real-world cryptographers are hastily pushing for deployment of post-quantum primitives sooner rather than later. Subsequently, a proposal that is gaining popularity in the PQC community is the deployment of *hybrid* schemes, i.e., schemes where a long-standing classically-secure primitive $\mathcal{P}$ is partnered alongside a newer post-quantum candidate $\mathcal{Q}$ (cf. [5]). The simple reasoning here is that, even if further cryptanalysis weakens $\mathcal{Q}$'s resistance to classical computers, the hybrid scheme $\mathcal{P}+\mathcal{Q}$ is likely to remain classically secure; conversely, $\mathcal{P}$'s presumed weakness against a quantum computer does not affect the post-quantum security of $\mathcal{P}+\mathcal{Q}$. Taking such a prudent measure in the case of SIDH, which is much newer than other post-quantum primitives, seems especially wise. In Sect. 8 we present a possibility to partner SIDH public keys alongside traditional elliptic curve Diffie-Hellman (ECDH) public keys that are extremely strong. In particular, while our proposed SIDH parameters respectively offer around 128 and 192 bits of security against the best known quantum and classical attacks, the proposed hybrid offers around 384 bits of classical security based on the elliptic curve discrete logarithm problem (ECDLP). While this might seem like overkill, we show that this partnering is a very natural choice and comes at a relatively small cost: compared to a standalone SIDH, the size of the public keys and the overall runtime in our SIDH + ECDH hybrid increase by no more than 17 % and 13 %, respectively, and there is almost no additional code required to include ECDH in the scheme.

Public Key Validation. The security of unauthenticated ephemeral key exchange is modeled using passive adversaries, in which case we can assume that both parties' public keys are honestly generated. As was pointed out in April 2015 by a group at the NSA [24], in static key exchange when private keys are reused, validating public keys in the case of isogeny-based cryptography becomes both necessary and non-trivial. The suggested indirect public key validation procedure described in [24] is costly and requires one party to reveal their secret key, such

that only the other party can reuse theirs. In Sect. 9 we detail a form of direct validation for the public keys used in our scheme, and show how to achieve this validation efficiently in our compact framework.

SIDH History and Security. Beginning with an unpublished preprint with Rostovtsev in early 2006 [40], and then in a series of Russian papers that culminated in his thesis [45], Stolbunov proposed a Diffie-Hellman-like cryptosystem based on the difficulty of computing isogenies between ordinary (i.e., non-supersingular) elliptic curves. The best algorithm to solve this problem on a classical computer runs in exponential time and is due to Galbraith and Stolbunov [18].

In late 2010, however, Childs et al. [12] gave a quantum algorithm that computes isogenies between ordinary curves in subexponential time, assuming the Generalized Riemann Hypothesis (GRH). Subsequently, in late 2011, Jao and De Feo [22] put forward SIDH, which is instead based on the difficulty of computing isogenies between supersingular elliptic curves. This problem is immune to the quantum attack in [12], since this attack crucially relies on the endomorphism ring being commutative, which is not the case for a supersingular curve whose endomorphism ring is isomorphic to an order in a quaternion algebra [44, Sect. V.3.1].

Given two isogenous supersingular elliptic curves defined over a field of characteristic p, the *general* supersingular isogeny problem is to construct an isogeny between them. The best known classical algorithm for this problem is due to Delfs and Galbraith [15] and requires $\tilde{O}(p^{1/2})$ bit operations, while the best known quantum algorithm is due to Biasse et al. [6] and requires $\tilde{O}(p^{1/4})$ bit operations. The problems underlying SIDH (see Sect. 2) are not general in that the degree of the isogeny, which is smooth and in $O(\sqrt{p})$, is known and public. As is discussed by De Feo et al. [17, Sect. 5.1][2], this specialized problem can be viewed as an instance of the *claw problem*, and the optimal asymptotic classical and quantum complexities for the claw problem are known to be $O(p^{1/4})$ and $O(p^{1/6})$, respectively [47,52]. Currently, this approach yields the best known classical and quantum attacks against SIDH.

Organization . In Sect. 2 we recall the key concepts from [17] that are needed in SIDH. In Sect. 3 we show that all isogeny and point computations can be performed in $\mathbb{P}^1$; here we derive all of the lower-level functions that are called during the key generation and shared secret operations. In Sect. 4 we fix the underlying isogeny class used in our software, describe the high-level key exchange operations, and discuss other implementation choices. In Sect. 5 we detail the special field arithmetic that is tailored towards our chosen prime (as well as many other well-chosen SIDH-friendly primes).

We give a summary of the scheme in Sect. 6 and present performance results of our implementation in Sect. 7. In Sect. 8 we describe our proposal for a strong

[2] This is an extended version of the original SIDH paper by Jao and De Feo [22].

hybrid key exchange scheme that combines classical ECDH with post-quantum SIDH, and in Sect. 9 we show how to efficiently validate SIDH public keys in static key exchange settings. We conclude the paper in Sect. 10.

To promote future implementations of SIDH, we have endeavored to make this paper as self-contained as possible. Essentially, all functions that are needed to implement SIDH are described in Sect. 3. High level functions can be found in the appendix of the full version [13]. All other details can be found in the released code [14].

2 Diffie-Hellman Key Exchange from Supersingular Elliptic Curve Isogenies

This section sets the stage by introducing notation, giving some basic properties of torsion subgroups and isogenies, and recalling the supersingular isogeny Diffie-Hellman key exchange protocol. This is all described in a similar fashion by De Feo et al. in [17, Sect. 2].

Smooth Order Supersingular Elliptic Curves. SIDH uses isogeny classes of supersingular elliptic curves with smooth orders so that rational isogenies of exponentially large (but smooth) degree can be computed efficiently as a composition of low degree isogenies. Fix two small prime numbers ℓ_A and ℓ_B, an integer cofactor f, and let p be a prime of the form $p = \ell_A^{e_A}\ell_B^{e_B}f \pm 1$. It is then easy to construct a supersingular elliptic curve E defined over $\mathbb{F}_{p^2}$ of order $(\ell_A^{e_A}\ell_B^{e_B}f)^2$ [9].

For $\ell \in \{\ell_A, \ell_B\}$ and $e \in \{e_A, e_B\}$ the corresponding exponent, we have that the full ℓ^e-torsion group on E is defined over $\mathbb{F}_{p^2}$, i.e. $E[\ell^e] \subseteq E(\mathbb{F}_{p^2})$. Since ℓ is coprime to p, $E[\ell^e] \cong (\mathbb{Z}/\ell^e\mathbb{Z}) \times (\mathbb{Z}/\ell^e\mathbb{Z})$ [44, III. 6.4]. Let $P, Q \in E[\ell^e]$ be two points that generate $E[\ell^e]$ such that the above isomorphism is given by $(\mathbb{Z}/\ell^e\mathbb{Z}) \times (\mathbb{Z}/\ell^e\mathbb{Z}) \to E[\ell^e]$, $(m, n) \mapsto [m]P + [n]Q$. Roughly speaking, the SIDH secret keys are degree ℓ^e isogenies of the base curve E, which are in one-to-one correspondence with the cyclic subgroups of order ℓ^e that form their kernels. A point $[m]P + [n]Q$ has full order ℓ^e if and only if at least either m or n are not divisible by ℓ. There are $\ell^{2e-2}(\ell^2 - 1)$ such points. Since distinct cyclic subgroups only intersect in points of order less than ℓ^e and all full-order points in a single subgroup are coprime multiples of one such point, it follows that there are $\ell^{e-1}(\ell + 1)$ distinct cyclic subgroups of order ℓ^e.

Computing Large Degree Isogenies. Given a cyclic subgroup $\langle R \rangle \subseteq E[\ell^e]$ of order ℓ^e, there is a unique isogeny ϕ_R of degree ℓ^e, defined over $\mathbb{F}_{p^2}$ with kernel $\langle R \rangle$ [44, III. 4.12], mapping E to an isogenous elliptic curve $E/\langle R \rangle$. The isogeny ϕ_R can be computed as the composition of e isogenies of degree ℓ which in turn can be computed by using Vélu's formulas [49]. As described in [17, Sect. 4.2.2], we can start with $E_0 := E$ and $R_0 := R$ and then iteratively compute $E_{i+1} = E_i/\langle [\ell^{e-i-1}]R_i \rangle$ for $0 \leq i < e$ as follows. Each iteration computes the

degree-ℓ isogeny $\phi_i : E_i \to E_{i+1}$ whose kernel is the cyclic group $\langle [\ell^{e-i-1}]R_i \rangle$ of order ℓ, before applying the isogeny to compute $R_{i+1} = \phi_i(R_i)$. The point R_i is an (ℓ^{e-i})-torsion point and so $[\ell^{e-i-1}]R_i$ has order ℓ. Thus, the composition $\phi_R = \phi_{e-1} \circ \cdots \circ \phi_0$ has degree ℓ^e, which together with $(\phi_{e-1} \circ \cdots \circ \phi_0)(R) = R_e = \mathcal{O}$ shows that $\ker(\phi_R) = \langle R \rangle$, and therefore that $\phi = \phi_{e-1} \circ \cdots \circ \phi_0$.

There are two obvious ways of computing ϕ using the above decomposition. One of them follows directly from the description above: in each iteration, one first computes the scalar multiplication $[\ell^{e-i-1}]R_i$ to obtain a point of order ℓ, then uses Vélu's formulas to compute ϕ_i, and evaluates it at R_i to obtain the next point R_{i+1}. Jao and De Feo [22, Fig. 2] call this the multiplication-based strategy because it is dominated by the number of scalar multiplications by ℓ that are needed to obtain the ℓ-torsion points. The second obvious approach is called the isogeny-based method [22, Fig. 2] because it is dominated by the number of isogeny evaluations. It requires only one loop of scalar-multiplications that stores all ℓ-multiples of R, i.e., all intermediate results $Q_i = [\ell^i]R$ for $0 \leq i < e$. The point Q_{e-1} has order ℓ and can be used to obtain the isogeny ϕ_0 as above. One then replaces all Q_i for $0 \leq i \leq (e-2)$ by $\phi_0(Q_i)$. At this point Q_{e-2} has order ℓ and is used to obtain ϕ_1. This is repeated until one obtains ϕ_{e-1} and hence the composition ϕ.

De Feo et al. [17, Sect. 4.2.2] demonstrate that both of these methods are rather wasteful and that there is a much more efficient way to schedule the multiplications-by-ℓ and ℓ-isogeny evaluations. We briefly touch on this in Sect. 4, and defer the finer details to the full version [13].

SIDH Key Exchange. This paragraph recalls the SIDH key exchange protocol from [17, Sect. 3.2]. The public parameters are the supersingular curve $E_0/\mathbb{F}_{p^2}$ whose group order is $(\ell_A^{e_A}\ell_B^{e_B}f)^2$, two independent points P_A and Q_A that generate $E_0[\ell_A^{e_A}]$, and two independent points P_B and Q_B that generate $E_0[\ell_B^{e_B}]$. To compute her public key, Alice chooses two secret integers $m_A, n_A \in \mathbb{Z}/\ell_A^{e_A}\mathbb{Z}$, not both divisible by ℓ_A, such that $R_A = [m_A]P_A + [n_A]Q_A$ has order $\ell_A^{e_A}$. Her secret key is computed as the degree $\ell_A^{e_A}$ isogeny $\phi_A \colon E_0 \to E_A$ whose kernel is R_A, and her public key is the isogenous curve E_A together with the image points $\phi_A(P_B)$ and $\phi_A(Q_B)$. Similarly, Bob chooses two secret integers $m_B, n_B \in \mathbb{Z}/\ell_B^{e_B}\mathbb{Z}$, not both divisible by ℓ_B, such that $R_B = [m_B]P_B + [n_B]Q_B$ has order $\ell_B^{e_B}$. He then computes his secret key as the degree $\ell_B^{e_B}$ isogeny $\phi_B \colon E_0 \to E_B$ whose kernel is R_B, and his public key is E_B together with $\phi_B(P_A)$ and $\phi_B(Q_A)$. To compute the shared secret, Alice uses her secret integers and Bob's public key to compute the degree $\ell_A^{e_A}$ isogeny $\phi'_A \colon E_B \to E_{BA}$ whose kernel is the point $[m_A]\phi_B(P_A) + [n_A]\phi_B(Q_A) = \phi_B([m_A]P_A + [n_A]Q_A) = \phi_B(R_A)$. Similarly, Bob uses his secret integers and Alice's public key to compute the degree $\ell_B^{e_B}$ isogeny $\phi'_B \colon E_B \to E_{AB}$ whose kernel is the point $[m_B]\phi_A(P_B) + [n_B]\phi_A(Q_B) = \phi_A(R_B)$. It follows that E_{BA} and E_{AB} are isomorphic, so Alice and Bob can compute a shared secret as the common j-invariant $j(E_{BA}) = j(E_{AB})$.

Security Under SSDDH. In [17, Sect. 5], De Feo et al. give a number of computational problems related to SIDH and discuss their complexity. In [17, Sect. 6], they prove that SIDH is *session-key secure* in the authenticated-links adversarial model of Canneti and Krawczyk [10] under the Supersingular Decision Diffie-Hellman (SSDDH) problem, which we recall as follows. With the public parameters as above, one is given a tuple sampled with probability 1/2 from either $(E_A, E_B, \phi_A(P_B), \phi_A(Q_B), \phi_B(P_A), \phi_B(Q_A), E_{AB})$ or from $(E_A, E_B, \phi_A(P_B), \phi_A(Q_B), \phi_B(P_A), \phi_B(Q_A), E_C)$, where $E_{AB} \cong E_0/\langle [m_A]P_A + [n_A]Q_A, [m_B]P_B + [n_B]Q_B \rangle$, $E_C \cong E_0/\langle [m'_A]P_A + [n'_A]Q_A, [m'_B]P_B + [n'_B]Q_B \rangle$, and the values m'_A, n'_A, m'_B and n'_B are chosen randomly from the same respective distributions as m_A, n_A, m_B and n_B. The SSDDH problem is to determine from which distribution the tuple is sampled.

3 Projective Points and Projective Curve Coefficients

In this section we present one of our main technical contributions by showing that, just as the Montgomery form allows point arithmetic to be carried out efficiently in $\mathbb{P}^1$, in the context of SIDH it also allows isogeny arithmetic to be carried out in $\mathbb{P}^1$. This gives rise to fast, inversion-free point-and-isogeny operations that significantly boost the performance of SIDH. In comparison to the software[3] accompanying [17] that computes at least one inversion per isogeny computation, and therefore $O(\ell)$ inversions per round of the protocol, our software only requires one inversion during key generation and two inversions during the computation of the shared secret.

Montgomery Curves. Over a field K, a Montgomery curve [33] is defined by the two constants $(a, b) \in \mathbb{A}^2(K)$ as $E_{(a,b)} \colon by^2 = x^3 + ax^2 + x$. Unlike traditional ECC, in this work the defining curve does not stay fixed, but changes as we *move around* an isogeny class. As we discuss further below, it is therefore convenient to work projectively both with points on curves and with the curve coefficients themselves. Let $(A \colon B \colon C) \in \mathbb{P}^2(K)$ with $C \in \bar{K}^\times$ be such that $a = A/C$ and $b = B/C$. Then $E_{(a,b)}$ can alternatively be written as $E_{(A \colon B \colon C)} \colon By^2 = Cx^3 + Ax^2 + Cx$. The K-rational points on $E_{(a,b)}$ or $E_{(A \colon B \colon C)}$ are contained in $\mathbb{P}^2(K)$, so as usual we use the notation $(X \colon Y \colon Z) \in \mathbb{P}^2(K)$ with $Z \neq 0$ to represent all points $(x, y) = (X/Z, Y/Z)$ in $\mathbb{A}^2(K)$, and the point at infinity is $\mathcal{O} = (0 \colon 1 \colon 0)$. The j-invariants of the curves given by these models are $j(E_{a,b}) = \frac{256(a^2-3)^3}{a^2-4}$ and $j(E_{(A \colon B \colon C)}) = \frac{256(A^2-3C^2)^3}{C^4(A^2-4C^2)}$.

Kummer Varieties and Points in $\mathbb{P}^1$. Following [33], viewing the x-line $\mathbb{P}^1$ as the Kummer variety of $E_{(a,b)}$ allows for particularly efficient arithmetic in $E_{(a,b)}/\langle \pm 1 \rangle \cong \mathbb{P}^1$. Let $x \colon E_{(a,b)} \setminus \{\mathcal{O}\} \to \mathbb{P}^1$, $(X \colon Y \colon Z) \mapsto (X \colon Z)$. For the points $P, Q \in E_{(a,b)} \setminus \{\mathcal{O}\}$ and $m \in \mathbb{Z}$, Montgomery [33] gave efficient formulas

[3] See https://github.com/defeo/ss-isogeny-software/.

for computing the doubling function $\texttt{xDBL}\colon (x(P), a) \mapsto x([2]P)$, the function $\texttt{xADD}\colon (x(P), x(Q), x(Q-P)) \mapsto x(Q+P)$ for *differential additions*, and the function $\texttt{xDBLADD}\colon (x(P), x(Q), x(Q-P), a) \mapsto (x([2]P), x(Q-P))$ for the merging of the two. These are all ingredients in the Montgomery ladder function to compute the $\mathbb{Z}$-action on $E_{(a,b)}/\langle \pm 1 \rangle \cong \mathbb{P}^1$, i.e., $\texttt{LADDER}\colon (x(P), a, m) \mapsto x([m]P)$. We also make use of the Montgomery tripling function $\texttt{xTPL}\colon (x(P), a) \mapsto x([3]P)$ on $E_{(a,b)}/\{\pm 1\}$, which is taken from [17].

We note that the xADD function works identically for $E_{(a,b)}$ and $E_{(A\colon B\colon C)}$, while the other functions on $E_{(a,b)}$ that involve a can be trivially modified to work on $E_{(A\colon B\colon C)}$ by substituting $a = A/C$ and avoiding the inversion by carrying the denominator C through to the projective output. All of these functions are summarized in Table 1. Conveniently, all of these subroutines are only needed to work entirely in only one of $E_{(A\colon B\colon C)}$ and $E_{(a,b)}$.

During the computations of shared secrets, we found it advantageous to employ the function $\texttt{LADDER\_3\_pt}\colon (x(P), x(Q), x(Q-P), a, m) \mapsto x(P + [m]Q)$, which is precisely the "three point ladder" given by De Feo et al. [17, Algorithm 1].

Minimizing the Number of Inversions via Curves in $\mathbb{P}^1$. Observe that all of the functions mentioned above on $E_{(a,b)}/\{\pm 1\}$ (resp. $E_{(A\colon B\colon C)}/\{\pm 1\}$) depend entirely on a (resp. A and C) and are independent of b (resp. B). This is because, for a fixed $a = A/C$ and up to isomorphism, there are only two curves found by varying b (resp. B) over K: the curve E and its non-trivial quadratic twist. Indeed, an elliptic curve and its twist are unified under the quotient by $\{\pm 1\}$, i.e., have the same Kummer variety, so it is no surprise that the Kummer arithmetic is independent of the Montgomery b (resp. B) coefficient. Moreover, we see above that the j-invariant is also independent of b (resp. B).

Our implementation profits significantly from these observations, and the choice of Montgomery form provides two advantages in parallel. The first is the well-known Montgomery-style *point* arithmetic that unifies points and their inverses by ignoring the Y coordinate to work with $(X\colon Z) \in \mathbb{P}^1$; the second is new *isogeny* arithmetic that unifies curves and their quadratic twists by ignoring the B coefficient to instead work only with $(A\colon C) \in \mathbb{P}^1$. In this way all point operations and isogeny computations are performed in $\mathbb{P}^1$, meaning that only one inversion is required (at the very end) when generating public keys or computing shared secrets. In the latter case, the inversion is computed during the j-invariant function $\texttt{j\_inv}\colon (A, C) \mapsto j(E_{(A\colon B\colon C)})$, while in the former case we use a 3-way simultaneous inversion [33] to normalize all of the components of the public key prior to transmission; see Table 1 for more details on these functions.

Projective Three Isogenies. Let $x(P) = (X_3\colon Z_3) \in \mathbb{P}^1$ be such that P has order 3 in $E_{(A\colon C)}$. Let $E'_{(A'\colon C')} = E_{(A\colon C)}/\langle P \rangle$, $\phi\colon E_{(A\colon C)} \to E'_{(A'\colon C')}$, $Q \in E_a \setminus \ker(\phi)$, and write $x(Q) = (X\colon Z) \in \mathbb{P}^1$ with $x(\phi(Q)) = (X'\colon Z') \in \mathbb{P}^1$. Our goal is to derive two sets of explicit formulas: the first set computes the isogenous curve $E_{(A'\colon C')}$ from $(X_3\colon Z_3)$ and $E_{(A\colon C)}$, while the second set is used to evaluate the corresponding isogeny by computing $(X'\colon Z')$ from

the additional input $(X\colon Z)$. The projective version of [17, Eq. (17)] gives $(A'\colon C') = \big((AX_3Z_3 + 6(Z_3^2 - X_3^2))X_3\colon CZ_3^3\big)$, which can be computed in $6\mathbf{M} + 2\mathbf{S} + 5\mathbf{a}$[4]. However, it is possible to do much better by using $Z_3 \neq 0$ and the fact that X_3/Z_3 is a root of the 3-division polynomial $\psi_3(x) = 3x^4 + 4(A/C)x^3 + 6x^2 - 1$ on $E_{(A\colon C)}$. This yields the alternative expression $(A'\colon C') = \big(Z_3^4 + 18X_3^2Z_3^2 - 27X_3^4\colon 4X_3Z_3^3\big)$, which is independent of the coefficients of $E_{(A\colon C)}$ and can be computed in $3\mathbf{M} + 3\mathbf{S} + 8\mathbf{a}$; see the function `get_3_isog` in Table 1. For the evaluation of the isogeny, we modify the map in [17, Eq. (17)] to give $(X'\colon Z') = \big(X(X_3X - Z_3Z)^2\colon Z(Z_3X - X_3Z)^2\big)$. This costs $6\mathbf{M} + 2\mathbf{S} + 2\mathbf{a}$; see the function `eval_3_isog` in Table 1.

Projective Four Isogenies. We now let $x(P) = (X_4\colon Z_4) \in \mathbb{P}^1$ be such that P has exact order 4 in $E_{(A\colon C)}$, and leave all other notation and definitions as above. As is discussed in [17, Sect. 4.3.2], there are some minor complications in the derivation of 2- and 4-isogenies, either because a direct application of Vélu's formulas [49] for a 2-isogeny do not preserve the Montgomery form, or because repeated application of the 4-isogeny resulting from Vélu's formulas is essentially degenerate. For our purposes, i.e., in the case of 4-isogenies (overall, we found using 4-isogenies to be significantly faster than using 2-isogenies), the latter problem is remedied by application of the simple isomorphism in [17, Eq. (15)]. When building the 4^e isogenies as a composition of 4-isogenies, this isomorphism is needed in every 4-isogeny computation except for the very first one, and we derive explicit formulas for both of these cases.

Note that for the very first 4-isogeny $\phi_0\colon E_{(A\colon C)} \to E_{(A'\colon C')}$ computed in the public key generation phase, the curve $E_{(A\colon C)}$ is that which is specified in the system parameters; and, for the first 4-isogeny in the shared secret computation, $E_{(A\colon C)}$ is the curve that is received as part of a public key sent over the wire. In both cases the curve is normalized so that $A = a$ and $C = 1$. In this case we use [17, Eq. (20)] directly, which gives $(A'\colon C') = (2(a+6)\colon a-2)$, and projectivize the composition of [17, Eqs. (19) and (21)] to give $(X'\colon Z') = \big((X+Z)^2(aXZ + X^2 + Z^2)\colon (2-a)XZ(X-Z)^2\big)$. This costs $4\mathbf{M} + 2\mathbf{S} + 9\mathbf{a}$; see the function `first_4_isog` in Table 1.

For the general 4-isogeny, we projectivized the composition of the above isogeny with the isomorphism in [17, Eq. (15)], making some modifications as follows. We made use of the `xDBL` function to parameterize the point of order 2 in [17, Eq. (15)] in terms of the point $(X_4\colon Z_4)$ of order 4. For the isogeny evaluation function, we again found it advantageous to simplify under the applicable component of the 4-division polynomial $\psi_4(x, y) = 4y(x-1)(x+1)\hat{\psi}_4(x)$, which is $\hat{\psi}_4(x) = x^4 + 2(A/C)x^3 + 6x^2 + 2(A/C)x + 1$ and which vanishes at X_4/Z_4. For the computation of the isogenous curve, we get $(A'\colon C') = \big(2(2X_4^4 - Z_4^4)\colon Z_4^4\big)$, and for the evaluation of the isogeny, we get the image

[4] As usual, **M**, **S** and **a** represent the costs of field multiplications, squarings, and additions, respectively. We always count multiplications by curve coefficients as full multiplications, since these coefficients change within an isogeny class and thus we cannot expect any savings by treating them differently to generic elements.

point $(X' : Z')$ where $X' = X\left(2X_4Z_4Z - X(X_4^2 + Z_4^2)\right)(X_4X - Z_4Z)^2$ and $Z' = Z\left(2X_4Z_4X - Z(X_4^2 + Z_4^2)\right)(Z_4X - X_4Z)^2$. Since each 4-isogeny is evaluated at multiple points, during the above computation of the isogenous curve, we also compute and store five values that can be (re)used in the evaluation: $\mathbf{c} = [X_4^2 + Z_4^2, X_4^2 - Z_4^2, 2X_4Z_4, X_4^4, Z_4^4]$.

The computation of the isogenous curve and of the five values in $\mathbf{c}$ above costs $5\mathbf{S} + 7\mathbf{a}$, and on input of $\mathbf{c}$ and $Q = (X : Z)$, the isogeny evaluation costs $9\mathbf{M} + 1\mathbf{S} + 6\mathbf{a}$; see the functions `get_4_isog` and `eval_4_isog` in Table 1.

Summary of Subroutines. All of the point and isogeny operations are summarized in Table 1. We note that the input $\mathbf{c} \in K^5$ into the `eval_4_isog` function is the same tuple of constants output from `get_4_isog`, as described above.

Table 1. Summary of the subroutines used in our SIDH implementation. Here the points P and Q are on the curve $E_{(a,b)} = E_{(A:\,B:\,C)}$, and $E' = E_{(A':\,B':\,C')}$ is used to denote the isogenous curve. We use $n = \log_2 m - 1$ to count operations in loops. For a more detailed table, see the full version [13].

Function	Input (s)	Output (s)	M	S	a	I
`j_inv`	(A, C)	$j(E)$	3	4	8	1
`xDBLADD`	$\left(x(P), x(Q), x(Q-P), \frac{a+2}{4}\right)$	$(x([2]P), x(Q+P))$	6	4	8	-
`xADD`	$(x(P), x(Q), x(Q-P))$	$x(Q+P)$	3	2	6	-
`xDBL`	$\left(x(P), A+2C, 4C\right)$	$x([2]P)$	4	2	4	-
`xDBLe`	$(x(P), A, C, e)$	$x([2^e]P)$	4e	2e	4e	-
`LADDER`	$\left(x(P), a, m\right)$	$x([m]P)$	$5n$	$4n$	$9n$	-
`LADDER_3_pt`	$(x(P), x(Q), x(Q-P), a, m)$	$x(P + [m]Q)$	$9n$	$6n$	$14n$	-
`xTPL`	$\left(x(P), A+2C, 4C\right)$	$x([3]P)$	8	4	8	-
`xTPLe`	$(x(P), A, C, e)$	$x([3^e]P)$	8e	4e	8e	-
`get_3_isog`	$x(P)$	(A', C')	3	3	8	-
`eval_3_isog`	$(x(P), x(Q))$	$x(\phi(Q))$	6	2	2	-
`first_4_isog`	$(x(Q), a)$	$(x(\phi_0(Q)), A', C')$	4	2	9	-
`get_4_isog`	$x(P)$	$(A', C', \mathbf{c})$	-	5	7	-
`eval_4_isog`	$(\mathbf{c}, x(Q))$	$x(\phi(Q))$	9	1	6	-
`secret_pt`	$(P, Q = \tau(P), m)$	$x(P + [m]Q)$	$5n$	$4n$	$9n$	-
`distort_and_diff`	x_P	$x(\tau(P) - P)$	-	1	2	-
`get_A`	(x_P, x_Q, x_{Q-P})	A	4	1	7	1
`inv_3_way`	(z_1, z_2, z_3)	$(z_1^{-1}, z_2^{-1}, z_3^{-1})$	6	-	-	1

4 Parameters and Implementation Choices

Prime Field and Isogeny Class. From here on, the field K is fixed as $K = \mathbb{F}_{p^2}$, where $p := 2^{372} \cdot 3^{239} - 1$, and $\mathbb{F}_{p^2} = \mathbb{F}_p(i)$ for $i^2 = -1$. In terms of the

notation from Sect. 2, this means that $\ell_A = 2$, $\ell_B = 3$, $e_A = 372$, $e_B = 239$ and $f = 1$. We searched for primes of the form $2^{e_A}3^{e_B}f - 1$ with a bit length close to (but no larger than) 768, aiming to strike a balance $\ell_A^{e_A} \approx \ell_B^{e_B}$ to ensure that one side of the key exchange is not appreciably easier to attack than the other (more on this below), and to balance the computational costs for Alice and Bob. We originally searched with no restriction on the cofactor f, but did not find an example of another prime that would perform as fast as ours and where the overall security was increased enough to warrant $f \neq 1$. Given the best known classical and quantum attack complexities (see Sect. 1), choosing a prime close to 768 bits aims to reach a claim of 192 bits of classical security and 128 bits of quantum security. The arithmetic advantages of this prime choice are detailed in Sect. 5.

Our implementation works in the isogeny class of elliptic curves over $\mathbb{F}_{p^2}$ that contains the supersingular Montgomery curve $E_0/\mathbb{F}_{p^2} : y^2 = x^3 + x$. Every curve in this isogeny class has $(p+1)^2 = (2^{372} \cdot 3^{239})^2$ points and is also supersingular [44, Exercise 5.4 and 5.10(a)]. The curve E_0 is the public parameter that is the starting point for the key exchange protocol.

The Base-Field and Trace-Zero Torsion Subgroups. A valuable technique that was introduced by Verheul [50] and that has played a key role in the implementation of symmetric pairings on supersingular elliptic curves [42], is that of using a distortion map. Verheul showed that every supersingular elliptic curve has a distortion map [50]. For a prime power $\ell^e \mid \#E_0(\mathbb{F}_p)$, such a map connects the cyclic torsion subgroup $E_0(\mathbb{F}_p)[\ell^e]$ defined over the base field $\mathbb{F}_p$ with the trace-zero subgroup of $E_0(\mathbb{F}_{p^2})[\ell^e]$. The distortion map we use for E_0 is given by the endomorphism $\tau : E_0(\mathbb{F}_{p^2}) \to E_0(\mathbb{F}_{p^2}),\ (x, y) \mapsto (-x, iy)$.

An ℓ^e torsion point $P \in E_0(\mathbb{F}_p)$ is mapped to an ℓ^e-torsion point $\tau(P) \in E_0(\mathbb{F}_{p^2})$ and the Weil pairing $e_{\ell^e}(P, \tau(P)) \neq 1$ is non-trivial. It is easy to see that the trace of the image point is zero, namely $\mathrm{Tr}(\tau(P)) = \tau(P) + \pi_p(\tau(P)) = \mathcal{O}$, where π_p is the p-power Frobenius endomorphism on E_0. An advantage of using the trace-zero subgroup is that its points can be represented by two $\mathbb{F}_p$-elements only and are therefore half the size of a general curve point defined over $\mathbb{F}_{p^2}$.

Choosing Generator Points for Torsion Subgroups. We apply a similar idea in that we fix the public $\ell_A^{e_A}$-torsion points P_A, Q_A and $\ell_B^{e_B}$-torsion points P_B, Q_B as generators of the (respective) base field and trace-zero subgroups, chosen as follows. Let $P_A \in E_0(\mathbb{F}_p)[2^{372}]$ be the point given as $[3^{239}](z, \sqrt{z^3 + z})$, where z is the smallest positive integer such that $\sqrt{z^3 + z} \in \mathbb{F}_p$ and P_A has order 2^{372}. The point P_B is selected in the same way with order and cofactor swapped. We then take $Q_A = \tau(P_A)$ and $Q_B = \tau(P_B)$, which produces the following generators: $P_A = [3^{239}](11, \sqrt{11^3 + 11})$, $Q_A = \tau(P_A)$, $P_B = [2^{372}](6, \sqrt{6^3 + 6})$, and $Q_B = \tau(P_B)$.

In addition to the base field representations mentioned above, the simple relationship between the coordinates of Q_A and P_A and the coordinates of Q_B and P_B helps to further compactify the public parameters; see Sect. 6. However,

choosing $\{P_A, Q_A\}$ and $\{P_B, Q_B\}$ as the bases for generating isogeny kernels from the base-field and trace-zero torsion subgroups can have caveats. For example, in the case $\ell = \ell_A = 2$, one obtains the following lemma (the proof of which is in the full version [13]).

Lemma 1. *Let $E : y^2 = x^3 + x$ be a supersingular elliptic curve defined over $\mathbb{F}_p$, $p > 3$, $p \equiv 3 \pmod 4$, such that $\#E(\mathbb{F}_p) = 2^e \cdot N$ with N odd. Let $\mathbb{F}_{p^2} = \mathbb{F}_p(i)$, $i^2 = -1$, and let $E[\ell^e] \subseteq E(\mathbb{F}_{p^2})$. Let $P \in E(\mathbb{F}_p)[2^e]$ be any point of order 2^e and let $Q \in E(\mathbb{F}_{p^2})[2^e]$ be any point of order 2^e with $\mathrm{Tr}(Q) = Q + \pi_p(Q) = \mathcal{O}$. Then the order of $P + Q$ equals 2^{e-1}.*

In particular, Lemma 1 proves that any point of the form $P + [m]Q$ for odd m has order less than 2^e. Also note that if m is even, then the order of $P + [m]Q$ is 2^e because $[2^{e-1}](P + [m]Q) = [2^{e-1}]P \neq \mathcal{O}$. Furthermore, this means that the points P and Q do not generate the full 2^e-torsion subgroup, and strictly speaking, the two points are not independent[5].

In the following two paragraphs we show how Alice and Bob can choose their secret scalars to guarantee that the degrees of their isogenies are maximal, i.e., $\ell_A^{e_A}$ and $\ell_B^{e_B}$ respectively.

Sampling Full Order 2-Torsion Points. To sample a 2-torsion point R_A of full order, we sample a uniform random integer $m' \in \{1, 2, \ldots, 2^{e_A - 1} - 1 = 2^{371} - 1\}$ and set $R_A = P_A + [2m']Q_A$; R_A is guaranteed to have order 2^{e_A} by the above discussion. Because two distinct choices for m' lead to two distinct cyclic subgroups generated by the corresponding R_A, one can reach $2^{e_A - 1} - 1 = 2^{371} - 1$ distinct subgroups and thus isogenies with this sampling procedure. We have seen in Sect. 2 that there are $3 \cdot 2^{e_A - 1}$ distinct full order subgroups in $E_0[2^{e_A}]$, and thus our sampling procedure only reaches about one third of those.

Sampling Full Order 3-Torsion Points. To sample a 3-torsion point R_B of full order, we sample a uniform random integer $m' \in \{1, 2, \ldots, 3^{e_B - 1} - 1 = 3^{238} - 1\}$ and set $R_B = P_B + [3m']Q_B$. Since $[3^{e_B - 1}]R_B = [3^{e_B - 1}]P_B \neq \mathcal{O}$, R_B is guaranteed to have order 3^{e_B}. In this way, we reach $3^{238} - 1$ of the possible subgroups and corresponding isogenies. Since there are $4 \cdot 3^{e_B - 1}$ such subgroups in $E_0[3^{e_B}]$, we sample from about one quarter of those.

Strategies for Isogeny Computation and Evaluation. For computing and evaluating $\ell_A^{e_A}$- and $\ell_B^{e_B}$-isogenies, we closely follow the methodology described in [17, Sect. 4.2]. As already described in Sect. 2, such isogenies are composed of e_A isogenies of degree ℓ_A and e_B isogenies of degree ℓ_B, respectively. Figure 2 in [17] illustrates this computation with the help of a directed acyclic graph. In order to be able to evaluate the desired isogeny, one needs to compute all points that are

[5] Whenever we use the term independent for the points P and Q in what follows, we mean that the Weil pairing evaluated at P and Q is non-trivial.

represented by the final vertices, i.e., the leaves in the graph. As described earlier in Sect. 2, using the multiplication-based or isogeny-based methods to traverse this graph yields a simple but costly algorithm. De Feo et al. [17, Sect. 4.2.2] provide a discussion of how to obtain an optimal algorithm. They formally define the notion of a *strategy* for evaluating ϕ along a directed acyclic graph and show how to find an optimal strategy depending on the relative costs of scalar multiplication-by-ℓ and ℓ-isogeny evaluation. For the details on the optimal strategies for our chosen parameters, we refer to the full version [13].

5 Field Arithmetic

In this section, we describe the advantages of the chosen prime and optimizations to speed up the modular reduction inside SIDH, which were inspired by similar work on so-called Montgomery-friendly primes (e.g., see [19,27]). We remark that similar ideas can be easily applied to selecting primes and implementing their modular arithmetic at different security levels.

In our case, arithmetic is performed modulo the prime $p = 2^{372} \cdot 3^{239} - 1$. As described in Sect. 4, choosing an SIDH prime such that $\ell_A^{e_A} \approx \ell_B^{e_B}$ ensures a certain security strength across the whole key exchange scheme. Additionally, some implementations benefit from having a prime with a bit length slightly smaller than a multiple of a word size. Since 768 is the next multiple of 32 and 64 above the bit length of our prime, and $\log_2 p = 751 = 768 - 17$, the *extra room* available at the word boundaries enables the efficient use of other optimization techniques such as carry-handling elimination, and eases the efficient use of vector instructions. Working on a field of size slightly smaller than 2^{768} enables us to, e.g., use 12×64-bit limbs to represent field elements, whereas a prime slightly larger than 2^{768}, such as $p_{768} = 2^{387} \cdot 3^{242} - 1$ from [2], requires 13×64-bit limbs; the latter choice brings a relatively small increase in security at the expense of a significant increase in the cost of the modular arithmetic.

Since we work over $\mathbb{F}_{p^2}$, where $\mathbb{F}_{p^2} = \mathbb{F}_p(i)$ for $i^2 = -1$, we can leverage the extensive research done on the efficient implementation of such quadratic extension fields. In the context of pairings, high-speed implementations have exploited the combination of Karatsuba multiplication, lazy reduction, and carry-handling elimination; e.g., these techniques have been combined in optimized implementations on the curve BN254 [1]. Here we can follow a similar strategy since our field definition and underlying prime share several common traits with BN254, e.g., our prime being slightly smaller than a multiple of the word size enables the computation of several additions without carry-outs in the most significant word.

Efficient Modular Reduction. The cost of modular arithmetic (and, in particular, of modular multiplication) dominates the cost of the isogeny-based key exchange, so its efficient implementation is crucial for achieving high performance. At first glance, it would seem that SIDH primes prompt the use of generic Montgomery [32] or Barrett [3] reduction algorithms, which are relatively expensive in comparison with the efficient reduction of certain primes with special

form (e.g., pseudo-Mersenne primes). For example, Azarderakhsh et al. [2] use a generic Barrett reduction for computing the modular multiplication in their SIDH implementation. However, we note that primes of this form *do* have a special shape that is amenable to faster modular reduction. Consider the case of the well-known Montgomery reduction [32]: letting $R = 2^{768}$ and $p' = -p^{-1} \bmod R$, then one can compute the Montgomery residue $c = aR^{-1} \bmod p$ for an input $a < pR$, by using $c = (a + (ap' \bmod 2^{768}) \cdot p)/2^{768}$, which costs approximately $s^2 + s$ multiplications for a $2s$-limb value a. For $p = 2^{372} \cdot 3^{239} - 1$, however, this computation simplifies to $c = (a + (ap' \bmod 2^{768}) \cdot 2^{372} \cdot 3^{239})/2^{768}$.

Moreover, $p' = -p^{-1} \bmod 2^{768}$ also exhibits a special form which reduces the cost of computing $ap' \bmod 2^{768}$ (e.g., $p' - 1$ contains five 64-bit limbs or eleven 32-bit limbs of value 0). In total, the cost of computing c in this case is $s(s - \lfloor 372/w \rfloor)$ multiplications for a word-size w. For example, if $w = 64$ (i.e., $s = 12$), the theoretical speedup for the simplified modular reduction is about 1.85x when applying these optimizations.

It is straightforward to extend the above optimizations to the different Montgomery reduction variants that exist in the literature. For our implementation, we adapted the Comba-based Montgomery reduction algorithm from [41]. Although merged multiplication/reduction algorithms, such as the coarsely integrated operand scanning (CIOS) Montgomery multiplication [25], offer performance advantages in certain scenarios, we prefer an implementation variant that consists of separate routines for integer multiplication and modular reduction. This approach enables the use of lazy reduction for the $\mathbb{F}_{p^2}$ arithmetic and allows easy-to-implement improvements in the integer multiplication, e.g., by using Karatsuba.

Algorithm 1 is based on the Montgomery reduction algorithm in product scanning form (a.k.a. Comba) presented in [41]. It has been especially tailored for efficient computation modulo the prime $p = 2^{372} \cdot 3^{239} - 1$ following the optimizations discussed above. As usual, given a radix-2^r field element representation using s limbs, the algorithm receives as input an operand $a < 2^{rs}p$ (e.g., the integer product of two Montgomery residues) and outputs the Montgomery residue $c = a \cdot 2^{-rs} \bmod p$. Here c is typically computed as $(a + (ap' \bmod 2^r) \cdot p)/2^r$ (s times) in a Comba-like fashion, where $p' = -p^{-1} \bmod 2^r$. However, as mentioned above, this expression simplifies to $(a + (a \bmod 2^r) \cdot \hat{p})/2^r$ where $\hat{p} = p + 1 = 2^{372} \cdot 3^{239}$, since $p' = 1$ for our prime. In addition, Algorithm 1 eliminates several multiplications due to the fact that the $\lfloor e_A/r \rfloor$ least significant limbs in $\hat{p}$ have value 0.

Since our scheme forces the availability of extra room in the radix-2^r representation (which is made possible by having the additional condition that $p < 2^{rs-2}$), there is no overflow in the most significant word during the computation of c in Algorithm 1 (i.e., its intermediate value can be held on exactly s r-bit registers). Moreover, if field elements are represented as elements in $[0, 2p - 1]$ (instead of the typical range $[0, p - 1]$), the output of Algorithm 1 remains bounded without the need of the conditional subtraction in Steps 19–20 [51].

Algorithm 1. Optimized Comba-based Montgomery reduction for the prime $p = 2^{372} \cdot 3^{239} - 1$.

Input: The prime $p = 2^{e_A} \cdot 3^{e_B} - 1$; the value $\hat{p} = p + 1$ containing $z = \lfloor e_A/r \rfloor$ 0-value terms in its r-bit representation, where $e_A = 372$, $e_B = 239$ and 2^r is the radix; the Montgomery constant 2^{rs} such that $2^{r(s-1)} \leq p < 2^{rs-1}$; and, the operand $a = (a_{2s-1}, ..., a_1, a_0)$ with $a < 2^{rs}p$ and $s = \lceil \log_2 p/r \rceil$.
Output: The Montgomery residue $c = a \cdot 2^{-rs} \bmod p$.

$(t, u, v) = 0$
for $i = 0$ **to** $s - 1$ **do**
  for $j = 0$ **to** $i - 1$ **do**
    if $j < i - z + 1$ **then**
      $(t, u, v) = c_j \times \hat{p}_{i-j} + (t, u, v)$
  $(t, u, v) = (t, u, v) + a_i$
  $c_i = v$
  $v = u,\ u = t,\ t = 0$
for $i = s$ **to** $2s - 2$ **do**
  if $z > 0$ **then**
    $z = z - 1$
  for $j = i - s + 1$ **to** $s - 1$ **do**
    if $j < s - z$ **then**
      $(t, u, v) = c_j \times \hat{p}_{i-j} + (t, u, v)$
  $(t, u, v) = (t, u, v) + a_i$
  $c_{i-s} = v$
  $v = u,\ u = t,\ t = 0$
$c_{s-1} = v + a_{2s-1}$
if $c \geq p$ **then**
  $c = c - p$
return c

Although typical values for r would be $w = 32$ or 64 to match w-bit architectures, some redundant representations might benefit from the use of $r < w$ in order to avoid additions with carries or to facilitate the efficient use of vector instructions. To this end, the chosen prime is very flexible and supports different efficient alternatives; for example, it supports the use of a 58-bit representation with $s = 13$ limbs when using 64-bit multipliers or the use of a 26-bit representation with $s = 29$ limbs when using 32-bit multipliers.

In our 64-bit implementation, we opted for a generic radix-2^{64} representation using $s = 12$ limbs, in which case the Montgomery constant is $2^{rs} = 2^{768}$. In this case, given that the initial and final loop iterations can be simplified in an unrolled implementation of Algorithm 1, the cost of the modular reduction is 83 multiplication instructions. This result almost halves the number of multiplication instructions compared to a naïve Montgomery reduction, which requires $12^2 + 12 = 156$ multiplication instructions (per reduction).

Inversions. Our SIDH implementation requires one modular inversion during key generation, and two modular inversions during the computation of the shared

secret. These inversions can be implemented using Montgomery inversion based on, e.g., the binary GCD algorithm. However, this method does not run in constant time by default, and therefore requires additional countermeasures to protect it against timing attacks (e.g., the application of input randomization). Since inversion is used scarcely in our software, we instead opted for the use of Fermat's little theorem, which inverts the field element a via the exponentiation $a^{p-2} \bmod p$ that uses a fixed addition chain. Our experiments showed that the cost of this exponentiation is around 9 times slower than (an average run of) the GCD-based method, however even the more expensive inversion only contributes to less than 1 % of the overall latency of each round of the protocol. Thus, our choice to compute each isolated inversion via a fixed exponentiation protects the implementation without impacting the performance in any meaningful way, and avoids the need for any additional randomness.

6 SIDH Implementation Summary

In this section we pull together all of the main ingredients from Sects. 2–5 to give a brief overview of the scheme and its implementation. For high-level Magma code that illustrates the entire SIDH protocol, see `SIDH.mag` in [14].

Public Parameters. Together with the curve $E_0 : y^2 = x^3 + x$ and the prime $p = 2^{372}3^{239} - 1$, the public parameters are $P_A = [3^{239}](11, \sqrt{11^3 + 11})$, $Q_A = \tau(P_A)$, $P_B = [2^{372}](6, \sqrt{6^3 + 6})$, and $Q_B = \tau(P_B)$. Given that all these square roots are in $\mathbb{F}_p$ (we choose the "odd" ones), and that Q_A and Q_B require no storage, this means that only 4 $\mathbb{F}_p$-elements (or 3004 bits) are required to fully specify the public generators. If we were to instead randomly choose extension field torsion generators without use of the distortion map, as is suggested in [17], then 16 $\mathbb{F}_p$ elements (or 12016 bits) would be required to specify the public generators.

Key Generation. On input of the public parameters above, and the secret key m_A chosen as in Sect. 4, Alice proceeds as in [13, Algorithm 3] (see [13, Algorithm 2] for the simple, but slower *multiplication-based* main loop). She calls the `secret_pt` function, which computes $P_A + [m_A]Q_A$ by calling `LADDER` to compute $x([m_A]Q_A)$, before recovering the corresponding y-coordinate using the Okeya-Sakurai strategy [36]; this allows the addition of P_A and $[m_A]Q_A$. All of these operations are performed over the ground field and we proceed by taking only $x(P_A + [m_A]Q_A)$ through the main loop.

We note that our implementation requires that Alice's secret isogeny is evaluated at both of the public parameters x_{P_B} and x_{Q_B}, as well as at the x-coordinate of the difference, $x_{Q_B - P_B}$; this allows Bob to kickstart the `three_pt_ladder` function (from [17, Algorithm 1]) during his shared secret phase. Conversely, Bob must also evaluate his secret isogeny at $x_{Q_A - P_A}$. In both cases, rather than setting x_{Q-P} as a public parameter, it can be computed on-the-fly from x_P,

since in this special instance, $x_{Q-P} = x_{\tau(P)-P} = i \cdot (x_P^2 + 1)/(2x_P)$. This is fed directly into our projective isogeny evaluation function, so we do not need $x_{Q-P} \in \mathbb{A}$, but can instead compute $x(Q-P) = (i(x_P^2+1) : 2x_P) \in \mathbb{P}^1$, which costs just one squaring and two additions in $\mathbb{F}_p$; this operation is performed with the `distort_and_diff` function.

At the conclusion of [13, Algorithm 3], Alice outputs her public key as $\mathrm{PK}_{\mathrm{Alice}} = [x_{\phi_A(P_B)}, x_{\phi_A(Q_B)}, x_{\phi_A(Q_B-P_B)}] \in \mathbb{F}_{p^2}^3$. Bob proceeds similarly, as shown in [13, Algorithm 5] (again, see [13, Algorithm 4] for a simpler, but slower multiplication-based approach), and outputs his public key as $\mathrm{PK}_{\mathrm{Bob}} = [x_{\phi_B(P_A)}, x_{\phi_B(Q_A)}, x_{\phi_B(Q_A-P_A)}] \in \mathbb{F}_{p^2}^3$.

Alice's fast key generation via [13, Algorithm 3], using the strategies for computing the isogeny trees as given in Sect. 4, requires 638 multiplications-by-4 and the evaluation of 1330 4-isogenies; calling the simpler [13, Algorithm 2] requires 17020 multiplications-by-4 an 744 4-isogeny evaluations. On Bob's side, the optimal strategy (i.e., fast key generation) requires 811 multiplications-by-3 and the evaluation of 1841 3-isogenies; the simpler version requires 28441 multiplications-by-3 and 956 3-isogeny evaluations. See Sect. 7 for the benchmarks and further discussion.

Remark 1. Observe that the public keys above only contain x-coordinates of points, and do not contain the Montgomery coefficient, a, that defines the isogenous curve E_a. This is because a can be recovered (on the other side) by exploiting the relation $a = \frac{(1-x_Px_Q-x_Px_{Q-P}-x_Qx_{Q-P})^2}{4x_Px_Qx_{Q-P}} - x_P - x_Q - x_{Q-P}$, which holds if x_P, x_Q and x_{Q-P} are the respective x-coordinates of three points P, Q and $Q-P$ on the Montgomery curve with coefficient a [19, Sect. A.2]. Here public key compression (i.e., dropping the a coefficient) is free, and decompression via the above equation amounts to $4\mathbf{M}+1\mathbf{S}+7\mathbf{a}+1\mathbf{I}$; see the function `get_A` in [14]. Compared to the overall shared secret computation, this decompression comes at a minor cost. In an earlier draft of this paper, we provided an option for a compression that instead transmitted the a coefficient, together with x_P, x_Q, and a *sign bit* that was used to choose the correct square root (during the recovery of x_{Q-P}). The above compression has the obvious advantage of saving the sign bit, and, more importantly, means that decompression only requires an inversion (instead of a square root). Since our software already required inversions, but did not use square roots anywhere else, the amount of additional code required to include this compression is minimal. We thank Luca De Feo and Ben Smith for pointing out this simpler compression.

Shared Secret. On input of $\mathrm{PK}_{\mathrm{Bob}} = [x_{\phi_B(P_A)}, x_{\phi_B(Q_A)}, x_{\phi_B(Q_A-P_A)}]$ and her secret key m_A, Alice first computes $a_B = \texttt{get\_A}(x_{\phi_B(P_A)}, x_{\phi_B(Q_A)}, x_{\phi_B(Q_A-P_A)})$, then calls [13, Algorithm 7] (again, see [13, Algorithm 6] for a more compact, but significantly slower main loop) to generate her shared secret. This starts by calling the `three_pt_ladder` function (from [17, Algorithm 1]) to compute $x(\phi_B(P_A) + [m_A]\phi_B(Q_A))$, which is used to generate the kernel of the isogeny

that is computed in the main loop. Finally, Alice uses the `j_inv` function to compute her shared secret. For Bob's analogous shared key generation, see [13, Algorithms 8–9].

Alice's fast key generation via [13, Algorithm 7], again using the strategies in Sect. 4, requires 638 multiplications-by-4 and the evaluation of 772 4-isogenies; calling the simpler [13, Algorithm 6] requires 17020 multiplications-by-4 and 186 4-isogeny evaluations. On Bob's side, the optimal strategy (i.e., fast key generation) requires 811 multiplications-by-3 and the evaluation of 1124 3-isogenies; the simpler version requires 28441 multiplications-by-3 and 239 3-isogeny evaluations. See Sect. 7 for the benchmarks and further discussion.

7 SIDH Performance

To evaluate the performance of the proposed supersingular isogeny system and the different optimizations, we wrote a software library supporting ephemeral SIDH key-exchange. The software is mostly written in the C language and has been designed to facilitate the addition of specialized code for different platforms and applications. The first release of the library comes with a fully portable C implementation supporting 32- and 64-bit platforms and two optional x64 implementations of the field arithmetic: one implementation based on intrinsics (which is, e.g., supported on Windows OS by Visual Studio) and one implementation written in x64 assembly (which is, e.g., supported on Linux OS using GNU GCC and clang compilers). The latter two optional modules are intended for high-performance applications. All of the software is publicly available in [14].

In Table 2, we present the performance of our software using the x64 assembly implementation in comparison with the implementation proposed by [2]. Results for the implementation in [2] were obtained by benchmarking their software[6] on the same Intel Sandy Bridge and Haswell machines, running Ubuntu 14.04 LTS. Note that the results in Table 2 differ from what was presented in Table 3 in [2]. The differences might be due to the use of overclocking (i.e., TurboBoost technology). For our comparisons, we disabled TurboBoost for a more precise and fair comparison.

Table 2 shows that the total cost of computing one Diffie-Hellman shared key (adding Alice's and Bob's individual costs together) using our software is, on both platforms, over 2.8 times faster than the software from [2]. These results are due to the different optimizations discussed throughout this work, the most prominent two being (i) the elimination of inversions during isogeny computations by working with projective curve coefficients, and (ii) the faster modular arithmetic triggered by the selected prime and the tailor-made Montgomery reduction for SIDH primes. It is important to note that, in particular, the advantage over [2] is not even larger because the numerous inversions used during the isogeny computations in [2] are not computed in constant time. Making such inversions constant-time would significantly degrade their performance (see the related paragraph in Sect. 5).

[6] See http://djao.math.uwaterloo.ca/thesis-code.tar.bz2.

Table 2. Performance results (expressed in millions of clock cycles) of the proposed SIDH implementation in comparison with the implementation by Azarderakhsh et al. [2] on x64 platforms. Benchmark tests were taken with Intel's TurboBoost disabled and the results were rounded to the nearest 10^6 clock cycles. Benchmarks were done on a 3.4 GHz Intel Core i7-2600 Sandy Bridge and a 3.4 GHz Intel Core i7-4770 Haswell processor running Ubuntu 14.04 LTS.

Operation	This work		Prior work [2]	
	Sandy Bridge	Haswell	Sandy Bridge	Haswell
Alice's keygen	50	46	165	149
Bob's keygen	57	52	172	152
Alice's shared key	47	44	133	118
Bob's shared key	55	50	137	122
Total	207	192	608	540

Remark 2. In Sect. 4 we discussed several specialized choices that were made for reasons unrelated to performance, e.g., in the name of simplicity and/or compactness. We stress that, should future cryptanalysis reveal that these choices introduce a security vulnerability, the performance of SIDH and the performance improvements in Sects. 3 and 5 are unlikely to be affected (in any meaningful way) by reverting back to the more general case(s). In particular, if it turns out that sampling from a fraction of the possible 2- and 3-torsion subgroups gives an attacker some appreciable advantage, then modifying the code to sample from the full set of torsion subgroups is merely an exercise, and the subsequent performance difference would be unnoticeable. Similarly, if any of (i) starting on a subfield curve (see [13, Remark 2]), (ii) using of the base-field and trace-zero subgroups, or (iii) using the distortion map, turns out to degrade SIDH security, then the main upshot of reverting to randomized public generators or starting on a curve minimally defined over $\mathbb{F}_{p^2}$ would be the inflated public parameters (see Sect. 6); the slowdown during key generation would be minor and the shared secret computations would be unchanged.

8 BigMont: A Strong ECDH + SIDH Hybrid

We now return to the discussion (from Sect. 1) of a hybrid scheme. Put simply, and in regards to both security and suitability, at present there is not enough confidence and consensus within the PQC community to warrant the standalone deployment of one particular post-quantum key exchange primitive. Subsequently, there is interest (cf. [5]) in deploying classical primitives alongside post-quantum primitives in order to hedge one's bets until a confidence-inspiring PQC key exchange standard arrives. This is particularly interesting in the case of SIDH, whose security has (because of its relatively short lifespan) received less cryptanalytic scrutiny than its post-quantum counterparts.

In this section we discuss how traditional ECDH key exchange can be included alongside SIDH key exchange at the price of a very small overhead. The main benefit of our approach is its simplicity; while SIDH could be partnered with ECDH on any of the standardized elliptic curves, this would mean that a lot more code needs to be written and/or maintained. In particular, it is often the case that the bulk of the code in high-speed ECC implementations relates to the underlying field arithmetic. Given that none of the fields underlying the standardized curves are SIDH-friendly[7], such a partnership would require either a generic implementation that would be much less efficient, or two unrelated implementations of field arithmetic. Our proposal avoids this additional complexity by performing ECDH on an elliptic curve defined over the same ground field as the one used for SIDH.

For $p = 2^{372}3^{239} - 1$, recall that our SIDH software works with isogenous curves $E_a/\mathbb{F}_{p^2}\colon y^2 = x^3 + ax^2 + x$ whose group orders are of the form $\#E_a = 2^i \cdot 3^j$, meaning that elliptic curve discrete logarithms are easy on all such curves by the Pohlig-Hellman algorithm [39]. However, there are also (exponentially many) ordinary curves of the form $E_a/\mathbb{F}_{p^2}$ that *are* cryptographically secure. In particular, over the base field $\mathbb{F}_p$, we can hope to find $a \in \mathbb{F}_p$ such that $E_a/\mathbb{F}_p$ and its quadratic twist $E'_a/\mathbb{F}_p$ are cryptographically strong, i.e., such that $E_a/\mathbb{F}_p$ is *twist-secure* [4].

Since $p \equiv 3 \bmod 4$, we searched for such a curve in exactly the same way as, e.g., Hamburg's Goldilocks curve [20] was found. Namely, since the value $(a+2)/4$ is the constant that appears in Montgomery's ladder computation [33], we searched for the value of a that gave rise to the smallest absolute value of $(a + 2)/4$ (when represented as an integer in $[0, p)$), and such that $\#E_a$ and $\#E'_a$ are both 4 times a large prime. For p as above, the first such value is $a = 624450$; to make a clear distinction between curves in the supersingular isogeny class and the strong curve used to perform ECDH, we (re)label this curve as $M_a/\mathbb{F}_p\colon y^2 = x^3 + ax^2 + x$ with $a = 624450$. The trace t_{M_a} of the Frobenius endomorphism on M_a (see [13]) gives $\#M_a = p + 1 - t_{M_a} = 4r_a$ and $\#M'_a = p + 1 + t_{M_a} = 4r'_a$, where r_a and r'_a are both 749-bit primes.

Following [4], every element in $\mathbb{F}_p$ corresponds to the x-coordinate of a point on either M_a or on M'_a. Together with the fact that Montgomery's `LADDER` function correctly computes underlying scalar multiplications independently of the quadratic twist, M_a being twist-secure allows us to treat all $\mathbb{F}_p$ elements as valid public keys and to perform secure ECDH without the need for any point validation.

The ECDH secret keys are integers in $[0, r_a)$. To ensure an easy constant-time `LADDER` function, we search for the smallest $\alpha \in \mathbb{N}$ such that αr_a and $(\alpha + 1)r_a - 1$ are the same bit length, which is $\alpha = 3$; accordingly, secret keys are parsed into $[3r_a, 4r_a)$ prior to the execution of scalar multiplications via `LADDER`. Subsequently, for $m \in [0, r_a)$ and $x(P) \in \mathbb{P}^1(\mathbb{F}_p)$, computing $x([m]P) =$ `LADDER`$(x(P), m, a)$ requires 1 call to `xDBL` and 750 calls to `xDBLADD` (see Table 1 for the operation counts of these functions, but note that here we can take

[7] Nor are any of the fields large enough to support highly (quantum-)secure SIDH.

Table 3. Comparison of standalone SIDH versus hybrid SIDH + ECDH. Timing benchmarks were taken on a 3.4 GHz Intel Core i7-4770 Haswell processor running Ubuntu 14.04 LTS with TurboBoost disabled and results rounded to the nearest 10^6 clock cycles. For simplicity, the bit-security of the primitives was taken to be the target security level and is not intended to be precise.

Comparison		Standalone SIDH	Hybrid SIDH + ECDH
≈ bit-security (hard problem)	Classical	192 (SSDDH)	384 (ECDHP)
	PQ	128 (SSDDH)	128 (SSDDH)
Public key size		564	658
Speed (cc $\times 10^6$)	Alice's keygen	46	52
	Bob's keygen	52	58
	Alice's shared key	44	50
	Bob's shared key	50	57

advantage of the fixed, small constant a). As all of these computations take place over the ground field, the total time taken to compute ECDH public keys and shared secrets is only a small fraction of the total time taken to compute the analogous SIDH keys – see Table 3.

From an implementation perspective, partnering SIDH with ECDH as above is highly advantageous because the functions required to compute $x([m]P) = \texttt{LADDER}(x(P), m, a)$ are already available from our Montgomery SIDH framework. In particular, the key generation (see Sect. 6) already has a tailored Montgomery `LADDER` function that works entirely over the base field, i.e., on the starting curve E_0, so computing ECDH keys is as simple as calling pre-existing functions on input of a different constant.

Though the speed overhead incurred by adding ECDH to SIDH in this way is small (see Table 3), choosing to use such a large elliptic curve group makes concatenated keys larger than they would be if a smaller elliptic curve was used for ECDH. For example, suppose we were to instead use the curve currently recommended in Suite B [35], Curve P-384, and (noting that uncompressed Curve P-384 points are larger than our proposed ECDH public keys) were to compress ECDH public keys as an x-coordinate and a sign bit. The total public key size with SIDH-compressed keys would then be 612 bytes, instead of the 658 bytes reported in Table 3. Though this difference is noticeable, it must be weighed up against the cost of the extensive additional code required to support Curve P-384, which would almost certainly share nothing in common with the existing SIDH code. Moreover, the simplicity of adding ECDH to SIDH as we propose is not the only reason to justify slightly larger public keys; the colossal 384-bit security achieved by M_{624450} also puts it in a position to tolerate the possibility of significant future advancements in ECDLP attacks. Due to the complexity of the ECDLP on M_{624450} in comparison with all of the elliptic curves in the standards, we dub this curve "BigMont".

In Table 3 we compare hybrid SIDH + ECDH versus standalone SIDH. The take-away message is that for a less than 1.17x increase in public key sizes

and less than 1.13x increase in the overall computing cost, we can increase the classical security of the key exchange from 192 bits (based on the relatively new SSDDH problem) to 384 bits (based on the long-standing ECDLP).

9 Validating Public Keys

Recall from Sect. 2 that De Feo et al. [17] prove that SIDH is *session-key secure* (under SSDDH) in the authenticated-links adversarial model [10]. This model assumes perfectly authenticated links which effectively forces adversaries to be passive eavesdroppers; in particular, it assumes that public keys are correctly generated by honest users. While this model can be suitable for key exchange protocols that are instantiated in a truly ephemeral way, in real-world scenarios it is often the case that (static) private keys are reused. This can incentivize malicious users to create faulty public keys that allow them to learn information about the other user's static private key, and in such scenarios validating public keys becomes a mandatory practical requirement.

In traditional elliptic curve Diffie-Hellman (ECDH), validating public keys essentially amounts to checking that points are on the correct and cryptographically secure curve [7]. Such *point validation* is considered trivial in ECDH, since checking that a point satisfies a curve equation requires only a handful of field multiplications and additions, and this is negligible compared to the overall cost (e.g., of a scalar multiplication).

In contexts where SIDH private keys are reused, public key validation is equally as important but is no longer as trivial. In April 2015, a group from the NSA [24] pointed out that "direct public key validation is not always possible for [...] isogeny based schemes" before describing more complicated options that validate public keys *indirectly*. In this section we describe ways to directly validate various properties of our public keys that, in particular, work entirely in our compact framework, i.e., without the need of y-coordinates or of the Montgomery b coefficient that fixes the quadratic twist.

Recall from Sect. 6 that an honest user generates public keys of the form $\text{PK} = \left[x_P, x_Q, x_{Q-P}\right] \in \mathbb{F}_{p^2}^3$, where $P = (x_P, y_P)$ and $Q = (x_Q, y_Q)$ are of the same order ℓ^e on a Montgomery curve E_a that is $\mathbb{F}_{p^2}$-isogenous to E_0, and are such that $Q \neq [\lambda]P$ for any $\lambda \in \mathbb{Z}$; the algorithms we describe below will only deem a purported public key as valid if this is indeed the case. Recall from Remark 1 that the three x-coordinates in the public key are immediately used to recover the Montgomery a coefficient that was dropped during compression; this coefficient must also be considered as part of the public key during validation.

Public key validation must check that the (underlying) points P and Q are of the full order ℓ^e. If not, then an SIDH-like analogue of the Lim-Lee [28] small subgroup attack becomes a threat; e.g., an attacker could send x_Q where Q has small order q and guess the shared secret (i.e., the kernel $\langle P + [m]Q\rangle$) to learn $m \bmod q$. In addition, the procedure must also assert that $Q \neq [\lambda]P$ (or equivalently, that $P \neq [\lambda]Q$) for some $\lambda \in \mathbb{Z}$; if this assertion is not made, then a malicious user can simply send a public key where $Q = [\lambda]P$, which ultimately

forces the shared secret to be independent of the honest party's private key. Such capabilities could be catastrophic if the authentication mechanism does not detect them.

The validation procedure we describe below guards against all of these attacks by asserting that P and Q both have order ℓ^e, and that the Weil pairing $e_{\ell^e}(P,Q)$ has the maximum possible order, namely the same order as the Weil pairing of the corresponding public parameters[8]; this means that the points P and Q generate as much of the ℓ^e torsion as is possible (according to the definition of the public parameters). This second assertion can be made in a very simple way, thanks to an observation by Ben Smith, who pointed out the following (using [31, Lemma 16.2]). If the points P and Q are in $E[mn]$, then the n-th power of the Weil pairing $e_{mn}(P,Q)$ can be computed as $e_{mn}(P,Q)^n = e_m([n]P,[n]Q)$, which allows us to efficiently check that the order of the Weil pairing is as it should be[9].

The application of the above validation procedure (to the three x-coordinates in a public key) is different for Alice and Bob, so we now describe these cases separately. We then discuss how both parties validate that the curve E_a corresponds to a supersingular curve in the correct isogeny class, and conclude the section with performance benchmarks for the validation process. All of the procedures described below can be found in the file `Validate.mag` [14].

Alice's Validation of Bob's Public Key. Alice must determine whether Bob's transmission $[x_P, x_Q, x_R] \in \mathbb{F}_{p^2}^3$ passes the tests described above. Recall from Sect. 4 that a consequence of Lemma 1 is that if the public parameters P_A and Q_A are chosen from the base field and trace-zero subgroups, then they do not form a basis for the full $\ell_A^{e_A}$-torsion. In particular, the order of the Weil pairing $e_{\ell_A^{e_A}}(P_A,Q_A)$ in our case is $\ell_A^{e_A-1} = 2^{371}$; although this order is less than $\ell_A^{e_A}$, it is as large as is possible when the two basis elements are chosen from these particular torsion subgroups.

If Bob's public key is honestly generated, then x_P and x_Q correspond to points P and Q whose Weil pairing also has order $\ell_A^{e_A-1}$; indeed, checking that this is the case ensures that we maximize the number of torsion subgroups that are spanned by $P + [2m']Q$. Let a be computed from x_P, x_Q and x_R as in Remark 1, and let $m = 4$ and $n = 2^{370}$ so that $mn = \ell_A^{e_A} = 2^{372}$. We assert that the exact order of $e_{\ell_A^{e_A}}(P,Q)$ is $\ell_A^{e_A-1}$ by showing that $e_{\ell_A^{e_A}}(P,Q)^{\ell_A^{e_A-2}}$ is non-trivial, making use of the identity above which gives $e_{\ell_A^{e_A}}(P,Q)^{\ell_A^{e_A-2}} = e_{mn}(P,Q)^n, = e_m([n]P,[n]Q) = e_4\left([2^{370}]P,[2^{370}]Q\right)$. Together with the assertion that P and Q both have exact order 2^{372}, the assertion that the Weil pairing $e_4\left([2^{370}]P,[2^{370}]Q\right)$ is non-trivial completes the validation of x_P and x_Q. If indeed P and Q have order 2^{372}, the points $P' = [2^{370}]P$ and $Q' = [2^{370}]Q$

[8] We thank Steven Galbraith and David Jao, who independently pointed out that the Pohlig-Hellman algorithm [39] can also be used to efficiently check whether P and Q are dependent.

[9] A prior version of this paper made a weaker assertion using a more elaborate computation.

have exact order 4. In that case, $e_4(P', Q') \neq 1$ if, and only if, $x(P') \neq x(Q')$. This can be seen by an elementary proof using [8, Theorem IX.10(5.)] and [8, Corollary IX.11] together with the fact that $Q' \in \langle P' \rangle$ implies $x(P') = x(Q')$. All of these checks can be performed entirely with x-coordinates as follows. We compute $x(P') = x([2^{370}]P) = \texttt{xDBLe}(x(P), a, 370)$ and $x(Q') = x([2^{370}]Q) = \texttt{xDBLe}(x(Q), a, 370)$. Next, we assert that $x(P') \neq x(Q')$, which is done projectively via a cross-multiplication. To check that P has full order 2^{372}, we then use two more calls to `xDBL` to assert that $(X : Z) = x([2]P')$ has $Z \neq 0$ and that $(\tilde{X} : \tilde{Z}) = x([4]P')$ has $\tilde{Z} = 0$; we do exactly the same for Q. If any of these checks fail, the public key is deemed invalid and rejected.

The assertion that x_R is the correct *difference* x_{Q-P} on E_a is implicit from the computation of a during decompression, and from the combined validation of x_P, x_Q and a. Validating that a indeed corresponds to a supersingular curve in the correct isogeny class is performed in the same way for Alice and Bob, so we postpone it until after describing Bob's validation.

Bob's Validation of Alice's Public Key. Bob must determine whether Alice's transmission $[x_P, x_Q, x_R] \in \mathbb{F}_{p^2}^3$ passes the tests described above. In this case our choice of the base field and trace-zero subgroups does not impede the possibility of the Weil pairing having full order; indeed, the public generators P_B and Q_B are such that the order of $e(P_B, Q_B)$ is $\ell_B^{e_B}$. Thus, honest public keys also give rise to the Weil pairing $e_{\ell_B^{e_B}}(P, Q)$ having order $\ell_B^{e_B}$. To make use of the identity above, we set $m = 3$ and $n = 3^{238}$ so that $mn = \ell_B^{e_B} = 3^{239}$, which gives $e_{\ell_B^{e_B}}(P, Q)^{\ell_B^{e_B - 1}} = e_{mn}(P, Q)^n = e_m([n]P, [n]Q) = e_3\left([3^{238}]P, [3^{238}]Q\right)$. Together with the assertion that P and Q both have exact order 3^{239}, the assertion that the Weil pairing $e_3\left([3^{238}]P, [3^{238}]Q\right)$ is non-trivial completes the validation of x_P and x_Q. If $P' = [3^{238}]P$ and $Q' = [3^{238}]Q$ have order 3, then $e_3(P', Q') \neq 1$ if, and only if, $x(P') \neq x(Q')$. This follows directly from [8, Corollary IX.11]. Again, we perform all of these checks using only x-coordinates as follows. We compute $x(P') = x([3^{238}]P) = \texttt{xTPLe}(x(P), a, 238)$ and $x(Q') = x([2^{238}]Q) = \texttt{xTPLe}(x(Q), a, 238)$ and assert that $x(P') \neq x(Q')$, which is again done projectively via a cross-multiplication. To check that P has full order 3^{239}, we assert that $(X : Z) = x(P')$ has $Z \neq 0$, and use one more call to `xTPL` to assert that $(\tilde{X} : \tilde{Z}) = x([3]P')$ has $\tilde{Z} = 0$; again, we do the same for Q. If any of these checks fail, the public key is deemed invalid and rejected.

Validating the Curve. We now show how to validate that a (i.e., the curve coefficient that is computed during the decompression of Alice or Bob's public key) corresponds to a Montgomery curve E_a that is a member of the correct supersingular isogeny class. The validation has two steps: we firstly assert that $j(E_a) \notin \mathbb{F}_p$ so that E_a is not a subfield curve, then we assert that E_a is in the correct supersingular isogeny class.

The first step is easy and totals a handful of multiplications in $\mathbb{F}_p$ (see the full version [13]); the less trivial step is to validate that E_a is supersingular.

To do this, we make use of Sutherland's probabilistic algorithm [46, Algorithm 1], which (for our purposes) says to pick a random point $P \in E_a(\mathbb{F}_{p^2})$, and to check whether $[p-1]P = \mathcal{O}$ or $[p+1]P = \mathcal{O}$. If this is the case, then E_a is supersingular with overwhelming probability: the probability that this test would pass if E_a was actually an ordinary curve is at most $8p/(p-1)^2 < 1/2^{747}$ [46, Proposition 1].

We now point out that E_a being supersingular is equivalent to either E_a or its quadratic twist, E'_a, belonging to the correct isogeny class. Namely, by [44, V.5.10(a)], E_a is supersingular if and only if its trace, t_{E_a}, satisfies $t_{E_a} \equiv 0 \bmod p$. Together with [48, Theorem 1], and recalling that $-2p \leq t_{E_a} \leq 2p$ [44, V.1.1], this means that there are (at most) 5 possible isogeny classes of supersingular elliptic curves, those which are described by $t_{E_a} \in \{-2p, -p, 0, p, 2p\}$. Since $p \equiv 3 \bmod 4$, there are only two possibilities for t_{E_a} that correspond to a Montgomery curve, i.e., two possible t_{E_a} such that $4 \mid \#E_a$ [33], namely $t_{E_a} = -2p$ and $t_{E_a} = 2p$. These traces respectively correspond to curves with $\#E_a = (p+1)^2$ that are in the correct isogeny class, and to curves with $\#E'_a = (p-1)^2$ that are in the isogeny class containing all of their non-trivial quadratic twists.

In our case we are trying to validate that a corresponds to a curve with $\#E_a = (p+1)^2$, so at first glance it would seem that the best route is to pick a random point $P \in E_a(\mathbb{F}_{p^2})$ and to assert that $[p+1]P = \mathcal{O}$. However, generating such a random point requires a square-root computation, and it turns out that we can (again) avoid the need for a square root altogether. For a given a, recall from Sect. 8 (or, in turn, from [4]) that elements in $\mathbb{F}_{p^2}$ are either the x-coordinate of a point on $E_a/\mathbb{F}_{p^2}$ or the x-coordinate of a point on $E'_a/\mathbb{F}_{p^2}$. This means that if E_a is supersingular, every element in $\mathbb{F}_{p^2}$ is the x-coordinate of a point whose order divides either $p-1$ or $p+1$. This gives us a way to quickly assert (with overwhelming probability) that a corresponds to a supersingular Montgomery curve in the correct isogeny class. With the Montgomery `LADDER` function as described in Sect. 3, we simply take a random element r in $\mathbb{F}_{p^2}$, compute $(X : Z) = \texttt{LADDER}((r : 1), a, p+1)$ and $(X' : Z') = \texttt{LADDER}((r : 1), a, p-1)$, and ensure that $Z \cdot Z' = 0$; otherwise, we reject the public key as invalid. We can compute a condition equivalent to $Z \cdot Z' = 0$ using only one call to the `LADDER` function as follows. The condition $\mathcal{O} \in \{[p-1]P, [p+1]P\}$ is equivalent to the condition $x(P) = x([p]P)$, which can be checked by computing $(X : Z) = \texttt{LADDER}(x(P), a, p)$ with $x(P) = (x_P : 1)$ and checking that $Z \cdot x_P = X$. However, calling `LADDER` to compute $x([p]P)$ directly is undesirable; given that $p + 1 = 2^{\ell_A} 3^{\ell_B}$, it is instead preferable to write a tailored ladder (consisting only of `xDBL` and `xTPL` operations) that computes a scalar multiplication by $p+1$. We do this by noting that the condition $x(P) = x([p]P)$ is equivalent to the condition that either $x([p+1]P) = x([2]P)$ or $[p+1]P = \mathcal{O}$ is satisfied.

The Price of Our Public Key Validation Procedure. On our target platforms, i.e., a 3.4 GHz Intel Core i7-2600 Sandy Bridge and a 3.4 GHz Intel Core i7-4770 Haswell processor running Ubuntu 14.04 LTS, the validation of Alice's public key costs (according to the above procedure) around 23 million

and 21 million clock cycles, respectively. Similarly, the validation of Bob's public key costs around 20 million and 18 million clock cycles, respectively. Referring back to Table 2, this means that both Alice and Bob's validation procedures cost between 0.39 and 0.43 times their key generation and shared secret computations.

Unlike public key validation in some other contexts, e.g., point validation in ECC, the compute time of the above SIDH public key validation is non-negligible compared to the compute time of each round of the key exchange. Nevertheless, in scenarios where static keys are desirable, the above overhead might be preferred over changes in the protocol description, e.g., the *indirect* validation proposed in [24].

10 Conclusion

We presented several new algorithms that have given rise to more efficient SIDH key exchange. We built a software library around a supersingular isogeny class determined by a fixed base curve that was chosen to target 128 bits of quantum security, and showed that these techniques give rise to a factor speedup of up to 2.9x over the previous fastest SIDH software. To our knowledge, our SIDH key exchange software is the first such implementation to run in constant time, and offers a range of additional benefits, such as compactness. In addition, we introduced two new techniques that bridge the gap between theoretical and real-world deployment of SIDH key exchange: the ECDH+SIDH hybrid and efficient algorithms for validating properties of public keys. The speed of our software (and the size of the public keys it generates) highlights the potential that SIDH currently offers as a candidate for post-quantum key exchange.

Acknowledgements. This paper has been significantly improved due to the feedback we received on a previous version. We are especially thankful to Ben Smith who pointed out a much simpler and faster method of our public key validation (see Sect. 9). We thank Luca De Feo and Ben Smith for pointing out a simplified compression of public keys (see Sect. 6). We thank Luca De Feo, Steven Galbraith and David Jao for their useful feedback, and the anonymous reviewers for their comments.

References

1. Aranha, D.F., Karabina, K., Longa, P., Gebotys, C.H., López, J.: Faster explicit formulas for computing pairings over ordinary curves. In: Paterson, K.G. (ed.) EUROCRYPT 2011. LNCS, vol. 6632, pp. 48–68. Springer, Heidelberg (2011)
2. Azarderakhsh, R., Fishbein, D., Jao, D.: Efficient implementations of a quantum-resistant key-exchange protocol on embedded systems. Technical report (2014). http://cacr.uwaterloo.ca/techreports/2014/cacr2014-20.pdf
3. Barrett, P.: Implementing the Rivest Shamir and Adleman public key encryption algorithm on a standard digital signal processor. In: Odlyzko, A.M. (ed.) CRYPTO 1986. LNCS, vol. 263, pp. 311–323. Springer, Heidelberg (1987)

4. Bernstein, D.J.: Curve25519: new Diffie-Hellman speed records. In: Yung, M., Dodis, Y., Kiayias, A., Malkin, T. (eds.) PKC 2006. LNCS, vol. 3958, pp. 207–228. Springer, Heidelberg (2006)
5. Bernstein, D.J.: The post-quantum internet. Invited talk at PQCrypto 2016, February 2016. https://cr.yp.to/talks/2016.02.24/slides-djb-20160224-a4.pdf
6. Biasse, J., Jao, D., Sankar, A.: A quantum algorithm for computing isogenies between supersingular elliptic curves. In: Meier, W., Mukhopadhyay, D. (eds.) INDOCRYPT 2014. LNCS, vol. 8885, pp. 428–442. Springer, Berlin (2014)
7. Biehl, I., Meyer, B., Müller, V.: Differential fault attacks on elliptic curve cryptosystems. In: Bellare, M. (ed.) CRYPTO 2000. LNCS, vol. 1880, p. 131. Springer, Heidelberg (2000)
8. Blake, I.F., Seroussi, G., Smart, N.P. (eds.): Advances in Elliptic Curve Cryptography. London Mathematical Society Lecture Notes Series, vol. 317. Cambridge University Press, Cambridge (2004)
9. Bröker, R.: Constructing supersingular elliptic curves. J. Comb. Number Theory **1**(3), 269–273 (2009)
10. Canetti, R., Krawczyk, H.: Analysis of key-exchange protocols and their use for building secure channels. In: Pfitzmann, B. (ed.) EUROCRYPT 2001. LNCS, vol. 2045, pp. 453–474. Springer, Heidelberg (2001)
11. Chen, L., Jordan, S., Liu, Y.-K., Moody, D., Peralta, R., Perlner, R., Smith-Tone, D.: Report on post-quantum cryptography. NISTIR 8105, DRAFT (2016). http://csrc.nist.gov/publications/drafts/nistir-8105/nistir_8105_draft.pdf
12. Childs, A.M., Jao, D., Soukharev, V.: Constructing elliptic curve isogenies in quantum subexponential time. J. Math. Cryptology **8**(1), 1–29 (2014)
13. Costello, C., Longa, P., Naehrig, M.: Efficient algorithms for supersingular isogeny Diffie-Hellman (full version). Cryptology ePrint Archive, Report 2016/413 (2016). http://eprint.iacr.org/
14. Costello, C., Longa, P., Naehrig, M.: SIDH Library (2016). https://www.microsoft.com/en-us/research/project/sidh-library/
15. Delfs, C., Galbraith, S.D.: Computing isogenies between supersingular elliptic curves over $\mathbb{F}_p$. Des. Codes Crypt. **78**(2), 425–440 (2016)
16. Devoret, M.H., Schoelkopf, R.J.: Superconducting circuits for quantum information: an outlook. Science **339**(6124), 1169–1174 (2013)
17. De Feo, L., Jao, D., Plût, J.: Towards quantum-resistant cryptosystems from supersingular elliptic curve isogenies. J. Math. Cryptology **8**, 209–247 (2014)
18. Galbraith, S.D., Stolbunov, A.: Improved algorithm for the isogeny problem for ordinary elliptic curves. Appl. Algebra Eng. Commun. Comput. **24**(2), 107–131 (2013)
19. Hamburg, M.: Fast and compact elliptic-curve cryptography. IACR CryptologyePrint Archive, 2012:309 (2012)
20. Hamburg, M.: Ed448-Goldilocks, a new elliptic curve. Cryptology ePrint Archive, Report 2015/625 (2015). http://eprint.iacr.org/
21. Hoffstein, J., Pipher, J., Silverman, J.H.: NTRU: a ring-based public key cryptosystem. In: Buhler, J.P. (ed.) ANTS 1998. LNCS, vol. 1423, pp. 267–288. Springer, Heidelberg (1998)
22. Jao, D., De Feo, L.: Towards quantum-resistant cryptosystems from supersingular elliptic curve isogenies. In: Yang, B.-Y. (ed.) PQCrypto 2011. LNCS, vol. 7071, pp. 19–34. Springer, Heidelberg (2011)

23. Kelly, J., Barends, R., Fowler, A.G., Megrant, A., Jeffrey, E., White, T.C., Sank, D., Mutus, J.Y., Campbell, B., Chen, Y., Chen, Z., Chiaro, B., Dunsworth, A., Hoi, I.-C., Neill, C., O'Malley, P.J.J., Quintana, C., Roushan, P., Vainsencher, A., Wenner, J., Cleland, A.N., Martinis, J.M.: State preservation by repetitive error detection in a superconducting quantum circuit. Nature **519**, 66–69 (2015)
24. Kirkwood, D., Lackey, B.C., McVey, J., Motley, M., Solinas, J.A., Tuller, D.: Failure is not an option: standardization issues for post-quantum key agreement. Talk at NIST Workshop on Cybersecurity in a Post-Quantum World, April 2015. http://www.nist.gov/itl/csd/ct/post-quantum-crypto-workshop-2015.cfm
25. Koc, C.K., Acar, T., Kaliski, B.S.: Analyzing and comparing Montgomery multiplication algorithms. IEEE Micro **16**(3), 26–33 (1996)
26. Kocher, P.C.: Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In: Koblitz, N. (ed.) CRYPTO 1996. LNCS, vol. 1109, pp. 104–113. Springer, Heidelberg (1996)
27. Lenstra, A.K.: Generating RSA moduli with a predetermined portion. In: Ohta, K., Pei, D. (eds.) ASIACRYPT 1998. LNCS, vol. 1514, pp. 1–10. Springer, Heidelberg (1998)
28. Lim, C.H., Lee, P.J.: A key recovery attack on discrete log-based schemes using a prime order subgroup. In: Kaliski Jr., B.S. (ed.) CRYPTO 1997. LNCS, vol. 1294, pp. 249–263. Springer, Heidelberg (1997)
29. McEliece, R.J.: A public-key cryptosystem based on algebraic coding theory. Coding Thv **4244**, 114–116 (1978)
30. Merkle, R.C.: Secrecy, authentication, and public key systems. Ph.D. thesis, Stanford University (1979)
31. Milne, J.S.: Abelian Varieties. In: Cornell, G., Silverman, J.H. (eds.) Arithmetic Geometry, pp. 103–150. Springer, New York (1986)
32. Montgomery, P.L.: Modular multiplication without trial division. Math. Comput. **44**(170), 519–521 (1985)
33. Montgomery, P.L.: Speeding the Pollard and elliptic curve methods of factorization. Math. Comput. **48**(177), 243–264 (1987)
34. Mosca, M.: Cybersecurity in an era with quantum computers: will we be ready? Cryptology ePrint Archive, Report 2015/1075 (2015). http://eprint.iacr.org/
35. National Security Agency (NSA): Cryptography today, August 2015. https://www.nsa.gov/ia/programs/suiteb_cryptography/
36. Okeya, K., Sakurai, K.: Efficient elliptic curve cryptosystems from a scalar multiplication algorithm with recovery of the y-coordinate on a Montgomery-form elliptic curve. In: Koç, Ç.K., Naccache, D., Paar, C. (eds.) CHES 2001. LNCS, vol. 2162, pp. 126–141. Springer, Heidelberg (2001)
37. Page, D.: Theoretical use of cache memory as a cryptanalytic side-channel. Technical report CSTR-02-003, Department of Computer Science, University of Bristol (2002). http://www.cs.bris.ac.uk/Publications/Papers/1000625.pdf
38. Patarin, J.: Hidden fields equations (HFE) and isomorphisms of polynomials (IP): two new families of asymmetric algorithms. In: Maurer, U.M. (ed.) EUROCRYPT 1996. LNCS, vol. 1070, pp. 33–48. Springer, Heidelberg (1996)
39. Pohlig, S.C., Hellman, M.E.: An improved algorithm for computing logarithms over GF(p) and its cryptographic significance. IEEE Trans. Inf. Theory **24**(1), 106–110 (1978)
40. Rostovtsev, A., Stolbunov, A.: Public-key cryptosystem based on isogenies. Cryptology ePrint Archive, Report 2006/145 (2006). http://eprint.iacr.org/
41. Scott, M.: Fast machine code for modular multiplication (1995). Manuscript, available for download at ftp://ftp.computing.dcu.ie/pub/crypto/fastmodmult2.ps

42. Scott, M.: Computing the Tate pairing. In: Menezes, A. (ed.) CT-RSA 2005. LNCS, vol. 3376, pp. 293–304. Springer, Heidelberg (2005)
43. Shor, P.W.: Algorithms for quantum computation: discrete logarithms and factoring. In: 35th Annual Symposium on Foundations of Computer Science, Proceedings, pp. 124–134. IEEE (1994)
44. Silverman, J.H.: The Arithmetic of Elliptic Curves. Graduate Texts in Mathematics, 2nd edn. Springer, New York (2009)
45. Stolbunov, A.: Cryptographic schemes based on isogenies. Ph.D. thesis, Norwegian University of Science and Technology (2012). http://www.item.ntnu.no/_media/people/personalpages/phd/anton/stolbunov-crypthographic_schemes_based_on_isogenies-phd_thesis_2012.pdf
46. Sutherland, A.V.: Identifying supersingular elliptic curves. LMS J. Comput. Math. **15**, 317–325 (2012)
47. Tani, S.: Claw finding algorithms using quantum walk. Theor. Comput. Sci. **410**(50), 5285–5297 (2009)
48. Tate, J.: Endomorphisms of abelian varieties over finite fields. Inventiones Math. **2**(2), 134–144 (1966)
49. Vélu, J.: Isogénies entre courbes elliptiques. CR Acad. Sci. Paris Sér. AB **273**, A238–A241 (1971)
50. Verheul, E.R.: Evidence that XTR is more secure than supersingular elliptic curve cryptosystems. J. Cryptology **17**(4), 277–296 (2004)
51. Walter, C.D.: Montgomery exponentiation needs no final subtractions. Electron. Lett. **35**(21), 1831–1832 (1999)
52. Zhang, S.: Promised and distributed quantum search. In: Wang, L. (ed.) COCOON 2005. LNCS, vol. 3595, pp. 430–439. Springer, Heidelberg (2005)