IACR Transactions on Cryptographic Hardware and Embedded Systems
ISSN XXXX-XXXX, Vol. 0, No. 0, pp. 1–23. DOI:XXXXXXXX

Low-Stack HAETAE for Memory-Constrained Microcontrollers

YoungBeom Kim[1], Christine van Vredendaal[2], Gustavo Banegas[3] and Seog Chung Seo[1*]

[1] Kookmin University, Seoul, Republic of Korea, {darania,scseo}@kookmin.ac.kr
[2] NXP Semiconductors, The Netherlands, christine.cloostermans@nxp.com
[3] LIX, CNRS, Inria, École Polytechnique, Institut Polytechnique de Paris, France
gustavo.souza-banegas@inria.fr

Abstract. We present a low-stack implementation of the module-lattice signature scheme HAETAE, targeting microcontrollers with 8 kB–16 kB of available SRAM. On such devices, peak stack usage is often the binding constraint, and HAETAE's hyperball-based sampler, large transient polynomial vectors, and variable-length signature payloads (hint and high-bits arrays) pose a particular challenge. To address this we introduce (i) *Rejection-aware pass decomposition*, (ii) *Component-level early rejection*, and (iii) *Reverse-order streaming entropy coding* using range Asymmetric Numeral Systems (rANS). Combined with streamed matrix generation, a two-pass hyperball sampler with streaming Gaussian backend, and row-streamed verification, these techniques bring signing stack from 71 kB–141 kB in the reference implementation down to 5.7 kB–5.9 kB, key generation to 4.7 kB–5.7 kB, and verification to 4.7 kB across all three security levels. Our pure C implementation covers all three security levels (HAETAE-2/3/5), whose optimization paths differ due to the public-key domain (d>0 vs. d=0) and rejection structure. We implement our optimization on a Nucleo-L4R5ZI and compare to the reference `pqm4` (for HAETAE-2 and -3) and a recently published memory-optimized implementation (targeting HAETAE-5 only). We reduce HAETAE-2, -3, and -5 stack by respectively 75, 86 and 8 % for key generation, 92, 95 and 26 % for signature generation, and 86, 91 and 22 % for verification. Depending on the parameter set, this impacts performance by at most a factor 1.9 and 2.6 for key and signature generation respectively, relative to the C reference. Verification at all security levels fits within 8 kB of RAM (signature buffer + stack) and is 2.99–4.25× faster than ML-DSA m4fstack at each comparable security level. We additionally validate portability under RIOT-OS on ARM Cortex-M4 and RISC-V targets, as well as on physical hardware spanning 8-bit AVR, 16-bit MSP430, and 32-bit Cortex-M0+.

Keywords: HAETAE · KpqC · lattice-based cryptography · small memory signatures · constrained devices

1 Introduction

Embedded systems rely on public-key authentication schemes for firmware updates, secure boot, device attestation, and device provisioning. With the transition to post-quantum cryptography, signature schemes based on module lattices are leading candidates for standardization. On low-end ARM Cortex-M and RISC-V microcontrollers, the stack available to an application may be only 8 kB–16 kB once the real-time operating system

*Corresponding author.

Licensed under Creative Commons License CC-BY 4.0. CC BY

(RTOS), network stack, and I/O drivers are loaded. For this class of targets, peak stack usage is often the binding constraint rather than code size or cycle count.

HAETAE [CCD+23, CCD+24], a module-lattice signature scheme presented at CHES'24, achieves shorter signatures and keys than ML-DSA [NIS24] by sampling ephemeral vectors from a hyperball rather than a hypercube. This compactness reduces communication and persistent storage costs, making HAETAE attractive for constrained deployments. For its attractive qualities, it has been identified as a winner in the Korean domestic post-quantum cryptography competition, KpqC [Kor].

Although HAETAE offers smaller public keys and signatures than ML-DSA at comparable security levels (Table 1), the reference implementation requires 71 kB to 141 kB of peak stack during the signing operation, depending on the security level, far exceeding the RAM budget of typical embedded targets. The root cause is that multiple large polynomial vectors used in sampling, challenge derivation, and hint computation are kept live simultaneously. The hyperball sampler amplifies the problem: because the scaling factor depends on the squared norm of the full Gaussian sample, a one-pass implementation must buffer the entire sample before producing any output.

This issue extends beyond signing. In modern embedded systems, key generation may also be performed at runtime for ephemeral credentials or key rotation, and verification must complete within tight stack budgets on sensor nodes that only verify signatures. A practical low-memory implementation should therefore address key generation, signing, and verification alike.

Related work. Bos, Renes, and Sprenkels [BRS22] showed that ML-DSA can fit within a few kilobytes of stack by trading recomputation for peak memory, streaming seed-derived objects, and reusing lifetime-disjoint buffers; similar strategies have been applied to FrodoKEM [BBC+23]. For HAETAE, however, only streaming matrix generation and sparse challenge multiplication carry over directly; the hyperball sampler, variable-length rANS encoding, and hint computation remain unaddressed by [BRS22]. To the best of our knowledge, [HCK+26] is the state-of-the-art memory-optimized HAETAE implementation on Cortex-M4. Following the strategy of [BRS22], they achieve 5,212 B/8,092 B/6,220 B (for respectively key generation/signing/verification) for HAETAE-5, approaching the minimum-state estimate we contribute in Section 3. However, they target HAETAE-5 only, leaving HAETAE-2/3 (i.e. $d>0$ and different public-key domain and rejection structure) unaddressed. Additionally, their verification stack (6,220 B) plus the signature (2,948 B) totals 9,168 B, exceeding the 8 kB budget of many constrained devices.

1.1 Contributions

In this work we aim to execute HAETAE on memory-constrained devices that only have up to 8 kB of SRAM. More concretely, we contribute

- **A low-stack HAETAE implementation with support for all HAETAE-2/3/5 security levels.** Our pure C implementation achieves 4.7 kB–5.7 kB key generation stack, 5.7 kB–5.9 kB signing stack, and 4.7 kB verification stack across all levels, with `.data` = 0 and `.bss` = 0. Our implementation is constant-time with the exception of the rejection-sampling loop inherent in the Fiat–Shamir-with-aborts paradigm.

- **Novel stack optimization techniques.** We introduce (i) *Rejection-aware pass decomposition*, which confines large buffers to non-overlapping lifetime phases; (ii) *Component-level early rejection*, which short-circuits the response computation when a partial norm exceeds the bound; and (iii) *Reverse-order streaming entropy coding*, which eliminates full hint and high-bits staging arrays. Combined

with streamed matrix generation, a two-pass hyperball sampler with a fully streaming Gaussian backend, and row-streamed verification, these techniques cover key generation, signing, and verification across all three security levels.

- **Evaluation and practical applicability.** We evaluate our techniques on the `pqm4` framework and validate portability under RIOT-OS on Cortex-M4 (nRF52840) and RISC-V (ESP32-C6) targets. On 8 kB devices, verification at all HAETAE security levels fits within budget (signature buffer + stack), and is 2.99–4.25× faster than ML-DSA m4fstack [BRS22] at each comparable security level. On 12 kB devices, all HAETAE levels support full signing (secret key + signature + stack), whereas ML-DSA-65 and ML-DSA-87 exceed this budget.

Our code is available here: https://github.com/smallhaetae/haetae_lowstack.

2 Background

In this section, we briefly explain the mathematical and algorithmic background of HAETAE, covering its algebraic setting, parameter sets, and core operations.

2.1 HAETAE

HAETAE is a module-lattice signature scheme operating in the polynomial ring

$$R_q \;=\; \mathbb{Z}_q[X]/(X^N+1),$$

with q a prime modulus. Vectors of ℓ (resp. k) polynomials are called `polyvecl` (resp. `polyveck`) and occupy $\ell \cdot N \cdot 4$ (resp. $k \cdot N \cdot 4$) bytes when stored with 32-bit coefficients. Polynomial multiplication in R_q is performed efficiently via the Number Theoretic Transform (NTT), which maps a polynomial to its evaluation at the $2N$-th roots of unity modulo q and reduces multiplication to a pointwise product in $\mathcal{O}(N \log N)$ time.

Table 1: Selected HAETAE parameters. $|\texttt{poly}| = N \times 4 = 1{,}024$ B.

Parameter	Description	HAETAE-2	HAETAE-3	HAETAE-5
N	polynomial degree	256	256	256
q	modulus	64 513	64 513	64 513
(k, ℓ)	module dimensions	$(2, 4)$	$(3, 6)$	$(4, 7)$
τ	challenge weight	58	80	128
d	PK truncation	1	1	0
α	$\mathbf{z}_1$ compression	256	256	256
α_h	$\mathbf{h}$ compression	512	512	256
$\lvert pk\rvert$	public key (bytes)	992	1 472	2 080
$\lvert sk\rvert$	secret key (bytes)	1 408	2 112	2 752
$\lvert sig\rvert$	signature (bytes)	1 474	2 349	2 948

Parameter sets. HAETAE is specified at three security levels. Table 1 lists the parameters referenced in this paper. All three sets share the polynomial degree N=256 and modulus q=64,513. The module dimensions (k, ℓ) determine the sizes of the public matrix $\mathbf{A} \in R_q^{k\times\ell}$ and the secret/response vectors, and are the primary driver of memory usage. The truncation parameter d governs the public-key format: d=1 (HAETAE-2/3) stores a rounded representation, while d=0 (HAETAE-5) stores the NTT-domain image directly (Section 2.1).

Like ML-DSA [NIS24], HAETAE is constructed in the *Fiat–Shamir with Aborts* (FSwA) framework [Lyu09]. In FSwA, the signer commits to an ephemeral randomness $\mathbf{y}$, receives (or derives) a binary challenge c of low Hamming weight, and computes a masked response

Algorithm 1 KeyGen(1^λ) for $d > 0$

Ensure: (pk, sk)
1: $\text{seed} \leftarrow \{0,1\}^{\rho_0}$
2: $(\text{seed}_{\mathbf{A}}, \text{seed}_{sk}, K) \leftarrow H_{\text{gen}}(\text{seed})$
3: $(\mathbf{a}_{\text{gen}}, \widehat{\mathbf{A}}_{\text{gen}}) \leftarrow \mathsf{ExpandA}_d(\text{seed}_{\mathbf{A}})$
4: $(\text{counter}_{sk}, \text{flag}) \leftarrow (0, \text{true})$
5: **while** flag **do**
6: $(\mathbf{s}_{\text{gen}}, \mathbf{e}_{\text{gen}}) \leftarrow \mathsf{ExpandS}(\text{seed}_{sk}, \text{counter}_{sk})$
7: $\mathbf{b} \leftarrow \mathbf{a}_{\text{gen}} + \mathsf{NTT}^{-1}(\widehat{\mathbf{A}}_{\text{gen}} \circ \mathsf{NTT}(\mathbf{s}_{\text{gen}})) + \mathbf{e}_{\text{gen}} \bmod q$
8: $(\mathbf{b}_0, \mathbf{b}_1) \leftarrow (\mathsf{LowBits}^{pk}(\mathbf{b}), \mathsf{HighBits}^{pk}(\mathbf{b}))$
9: $(\mathbf{s}_1, \mathbf{s}_2) \leftarrow (\mathbf{s}_{\text{gen}}, \mathbf{e}_{\text{gen}} - \mathbf{b}_0)$
10: $\text{counter}_{sk} \leftarrow \text{counter}_{sk} + 1$
11: **if** $\mathcal{N}(\mathbf{s}_1, \mathbf{s}_2) \leq \gamma^2 n$ **then**
12: flag $\leftarrow$ false
13: **end if**
14: **end while**
15: $tr \leftarrow H(\text{seed}_{\mathbf{A}}, \mathbf{b}_1)$
16: **return** $(pk = (\text{seed}_{\mathbf{A}}, \mathbf{b}_1), sk = (\mathbf{s}_1, \mathbf{s}_2, K, tr, \text{seed}_{\mathbf{A}}, \mathbf{b}_1))$

Algorithm 2 KeyGen(1^λ) for $d = 0$

Ensure: (pk, sk)
1: $\text{seed} \leftarrow \{0,1\}^{\rho_0}$
2: $(\text{seed}_{\mathbf{A}}, \text{seed}_{sk}, K) \leftarrow H_{\text{gen}}(\text{seed})$
3: $(\text{counter}_{sk}, \text{flag}) \leftarrow (0, \text{true})$
4: **while** flag **do**
5: $(\mathbf{s}_{\text{gen}}, \mathbf{e}_{\text{gen}}) \leftarrow \mathsf{ExpandS}(\text{seed}_{sk}, \text{counter}_{sk})$
6: $(\mathbf{s}_1, \mathbf{s}_2) \leftarrow (\mathbf{s}_{\text{gen}}, \mathbf{e}_{\text{gen}})$
7: $\text{counter}_{sk} \leftarrow \text{counter}_{sk} + 1$
8: **if** $\mathcal{N}(\mathbf{s}_1, \mathbf{s}_2) \leq \gamma^2 n$ **then**
9: flag $\leftarrow$ false
10: **end if**
11: **end while**
12: $\widehat{\mathbf{A}}_{\text{gen}} \leftarrow \mathsf{ExpandA}_d(\text{seed}_{\mathbf{A}})$
13: $\widehat{\mathbf{b}} \leftarrow -2\left(\widehat{\mathbf{A}}_{\text{gen}} \circ \mathsf{NTT}(\mathbf{s}_{\text{gen}}) + \mathsf{NTT}(\mathbf{e}_{\text{gen}})\right) \bmod q$
14: $tr \leftarrow H(\text{seed}_{\mathbf{A}}, \widehat{\mathbf{b}})$
15: **return** $(pk = (\text{seed}_{\mathbf{A}}, \widehat{\mathbf{b}}), sk = (\mathbf{s}_1, \mathbf{s}_2, K, tr, \text{seed}_{\mathbf{A}}, \widehat{\mathbf{b}}))$

Figure 1: Implementation specification of HAETAE key generation [CCD+23, Kpq26]: $d > 0$ applies to HAETAE-2 and HAETAE-3, while $d = 0$ applies to HAETAE-5.

$\mathbf{z} = \mathbf{y} + (-1)^b(c \star \mathbf{s})$. An abort (rejection) step is performed to ensure that the distribution of $\mathbf{z}$ does not leak information about the secret $\mathbf{s}$. The key distinction between ML-DSA and HAETAE lies in the distribution from which the ephemeral $\mathbf{y}$ is drawn, as detailed in Section 2.2.

HAETAE makes extensive use of HighBits and LowBits decompositions; we abbreviate these as HB and LB throughout, with superscripts (h, z_1, pk) indicating the decomposition type. For HAETAE-2/3 (d=1), the verification key stores only the high-order bits $\mathbf{b}_1 = \mathsf{HB}^{pk}(\mathbf{b})$ together with a public $\text{seed}_{\mathbf{A}}$ for expanding the matrix. The low-order bits $\mathbf{b}_0 = \mathsf{LB}^{pk}(\mathbf{b})$ are folded into the secret key. For HAETAE-5 (d=0), no rounding is applied; instead $\widehat{\mathbf{b}} = \mathsf{NTT}(-2\mathbf{b})$ is stored directly in the NTT domain.

The signature and key sizes of HAETAE are smaller than those of ML-DSA at comparable security levels, owing to the hyperball-based sampling strategy described below. At HAETAE-2 (targeting NIST security level 2), the signature size is 1 474 bytes, making it competitive with other lattice-based signatures while remaining efficient to verify.

Key Generation. Key generation expands a seed into the public matrix $\mathbf{A} \in R_q^{k \times \ell}$ and secret vectors $(\mathbf{s}_1, \mathbf{s}_2)$, then computes $\mathbf{b} = \mathbf{A}\mathbf{s}_1 + \mathbf{s}_2 \bmod q$. Unlike ML-DSA, HAETAE applies an explicit singular-value norm check $\mathcal{N}(\mathbf{s}_1, \mathbf{s}_2) \leq \gamma^2 N$ that rejects and resamples until the bound is satisfied; this rejection loop and its FFT-based workspace affect the key-generation stack peak. The implementation differs across security levels: for HAETAE-2/3 (d>0), $\mathbf{b}$ is rounded and the public key stores $\mathbf{b}_1 = \mathsf{HB}^{pk}(\mathbf{b})$, while for HAETAE-5 ($d$=0), $\widehat{\mathbf{b}} = \mathsf{NTT}(-2\mathbf{b})$ is stored directly in the NTT domain. The norm rejection also differs: HAETAE-2/3 interleave it with the matrix–vector multiplication, while HAETAE-5 performs it before expanding $\mathbf{A}$. See Algorithms 1 and 2 for the detailed specifications.

Signature generation. Unlike ML-DSA, which samples $\mathbf{y}$ uniformly from a hypercube, HAETAE draws $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ from a *discretized hyperball*: coefficients are sampled from a discrete Gaussian $\mathcal{D}_\sigma$, the vector is rescaled so that $\|(\mathbf{y}_1, \mathbf{y}_2)\|_2 \leq B_0\Lambda$, and a norm rejection test is applied. This yields shorter signatures but requires a more expensive and

Algorithm 3 HAETAE Signing

Require: $sk = (\mathbf{s}_1, \mathbf{s}_2, K, tr, \text{seed}_{\mathbf{A}}, \psi)$, M
Ensure: σ
$\widehat{\mathbf{A}} \leftarrow \mathsf{UnpackA}_d(\text{seed}_{\mathbf{A}}, \psi)$
$\mu \leftarrow H_{\mathsf{gen}}(tr, M)$
$\text{seed}_{ybb} \leftarrow H_{\mathsf{gen}}(K, \mu)$
$(\kappa, \sigma) \leftarrow (0, \perp)$
while $\sigma = \perp$ **do**
  $(\mathbf{y}_1, \mathbf{y}_2, b, b', \kappa) \leftarrow \mathsf{ExpandYbb}(\text{seed}_{ybb}, \kappa)$
  $\mathbf{w} \leftarrow \mathsf{NTT}^{-1}(\widehat{\mathbf{A}} \circ \mathsf{NTT}(\lfloor \mathbf{y}_1 \rceil)) + 2 \cdot \lfloor \mathbf{y}_2 \rceil \bmod q$
  $\mathbf{w}' \leftarrow \mathsf{fromCRT}(\mathbf{w}, \lfloor y_{1,1} \rceil)$
  $\mathbf{w}'_1 \leftarrow \mathsf{HB}^h(\mathbf{w}')$
  $\rho \leftarrow H(\mathbf{w}'_1, \mathsf{LSB}(\lfloor y_{1,1} \rceil), \mu)$
  $c \leftarrow \mathsf{SBC}_\tau(\rho)$
  $\widehat{c} \leftarrow \mathsf{NTT}(c)$
  $z_{1,1} \leftarrow y_{1,1} + (-1)^b \cdot c$
  $(\mathbf{z}_1)_{2..\ell} \leftarrow (\mathbf{y}_1)_{2..\ell} + (-1)^b \mathsf{NTT}^{-1}(\widehat{c} \circ \widehat{\mathbf{s}}_1)$
  $\mathbf{z}_2 \leftarrow \mathbf{y}_2 + (-1)^b \mathsf{NTT}^{-1}(\widehat{c} \circ \widehat{\mathbf{s}}_2)$
  if $\|(\mathbf{z}_1, \mathbf{z}_2)\|_2 < B'$ **and then** $(\|2(\mathbf{z}_1, \mathbf{z}_2) - (\mathbf{y}_1, \mathbf{y}_2)\|_2 > B$ **or** $b' = 1)$
    $\mathbf{h} \leftarrow \mathbf{w}'_1 - \mathsf{HB}^h(\mathbf{w}' - 2\lfloor \mathbf{z}_2 \rceil) \bmod^+ \frac{2(q-1)}{\alpha_h}$
    $\sigma \leftarrow \mathsf{PackSig}(\mathsf{HB}^{z_1}(\lfloor \mathbf{z}_1 \rceil), \mathsf{LB}^{z_1}(\lfloor \mathbf{z}_1 \rceil), \mathbf{h}, c)$
  end if
end while

Algorithm 4 HAETAE Verification

Require: $pk = (\text{seed}_{\mathbf{A}}, \psi)$, M, σ
Ensure: Accept or Reject
$\widehat{\mathbf{A}} \leftarrow \mathsf{UnpackA}_d(\text{seed}_{\mathbf{A}}, \psi)$
$(\mathsf{HB}^{z_1}(\lfloor \mathbf{z}_1 \rceil), \mathsf{LB}^{z_1}(\lfloor \mathbf{z}_1 \rceil), \mathbf{h}, c) \leftarrow \mathsf{UnpackSig}(\sigma)$
$\tilde{\mathbf{z}}_1 \leftarrow \mathsf{HB}^{z_1}(\lfloor \mathbf{z}_1 \rceil) \cdot 256 + \mathsf{LB}^{z_1}(\lfloor \mathbf{z}_1 \rceil)$
$w' \leftarrow \mathsf{LSB}(\tilde{z}_{1,1} - c)$
$\tilde{\mathbf{w}} \leftarrow \widehat{\mathbf{A}} \circ \mathsf{NTT}(\tilde{\mathbf{z}}_1) \bmod q$
$\tilde{\mathbf{w}}' \leftarrow \mathsf{fromCRT}(\tilde{\mathbf{w}}, w')$
$\tilde{\mathbf{w}}'_1 \leftarrow \tilde{\mathbf{h}} + \mathsf{HB}^h(\tilde{\mathbf{w}}') \bmod^+ \frac{2(q-1)}{\alpha_h}$
$\tilde{\mathbf{z}}_2 \leftarrow [\tilde{\mathbf{w}}'_1 \cdot \alpha_h + w'\mathbf{j} - \tilde{\mathbf{w}}' \bmod 2q] \,/\, 2$
$\tilde{\mu} \leftarrow H_{\mathsf{gen}}(\text{seed}_{\mathbf{A}}, \psi, M)$
return $(c = \mathsf{SBC}_\tau(H(\tilde{\mathbf{w}}'_1, w', \tilde{\mu})))$
  $\wedge\ (\|(\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2)\| < B'')$

Note: These two algorithms specify the implementation of HAETAE signing (left) and verification (right) [CCD+23, Kpq26]. M: message, σ: signature, HB: HighBits, LB: LowBits, SBC: SampleBinaryChallenge.

memory-intensive sampler (Section 4.4).

In the signature operation, the signer derives $\mu = H(pk, M)$ and repeats the steps of sampling $\mathbf{y}$, computing $\mathbf{w} = \mathbf{A}\lfloor \mathbf{y}_1 \rceil \bmod 2q$, deriving challenge c, and forming $\mathbf{z} = \mathbf{y} + (-1)^b c \star \mathbf{s}$. If $\mathbf{z}$ passes the norm bounds, the hint $\mathbf{h} = \mathsf{HB}^h(\mathbf{w}') - \mathsf{HB}^h(\mathbf{w}' - 2\lfloor \mathbf{z}_2 \rceil)$ is computed and the signature $\sigma = (\mathsf{HB}^{z_1}(\lfloor \mathbf{z}_1 \rceil), \mathsf{LB}^{z_1}(\lfloor \mathbf{z}_1 \rceil), \mathbf{h}, c)$ is encoded. See Algorithm 3 for the detailed specification. A naïve implementation keeps $\mathbf{y}$, $\mathbf{w}$, $\mathbf{z}$, and encoding buffers live simultaneously, totaling $(\ell+2k) \cdot |\mathtt{poly}|$ of ring objects (e.g. over 12 kB for HAETAE-5). The hyperball sampler amplifies this: the scaling factor depends on the norm of the full Gaussian sample, so a one-pass implementation must store all $(\ell+k) \cdot N$ coefficients before producing any output.

Verification. Verification (Algorithm 4) parses the signature, reconstructs $\tilde{\mathbf{w}} = \mathbf{A}\tilde{\mathbf{z}}_1 \bmod 2q$, recomputes the challenge from $\tilde{\mathbf{h}} + \mathsf{HB}^h(\tilde{\mathbf{w}}')$ and the message digest, and accepts if the recomputed challenge matches and all norm bounds hold. There is no rejection loop or secret-dependent sampler, but materializing the matrix product, decoded signature, and hint buffers simultaneously can still require substantial stack space.

2.2 Applicable ML-DSA Memory Optimization

Due to the similarity of the schemes, the memory-optimization techniques of [BRS22] can be applied to the operations common between HAETAE and ML-DSA.

Both schemes are module-lattice Fiat–Shamir signatures with the same high-level shape: seed expansion to generate a public matrix $\mathbf{A}$; computation of a public commitment involving $\mathbf{A}$ and short secret vectors; a sign-then-reject loop in which an ephemeral sample is masked by a low-weight challenge; and verification by recomputing the challenge from

the response and the public key. Both also rely on NTT-based polynomial arithmetic and HB/LB decompositions to compress intermediate values. This structural similarity means that the *streaming* and *liveness-reduction* ideas from [BRS22] carry over to HAETAE with appropriate adaptation.

The primary algorithmic difference is the distribution from which the ephemeral randomness $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ is drawn. ML-DSA samples $\mathbf{y}$ *uniformly from a hypercube*: each coefficient is drawn independently and uniformly from $\{-\gamma_1+1, \ldots, \gamma_1\}$. HAETAE instead samples $\mathbf{y}$ from a *discrete Gaussian conditioned on a hyperball*: coefficients are drawn from a discrete Gaussian, the vector is rescaled so that its Euclidean norm lies in $[\,0, B_0\Lambda\,]$, and a Euclidean norm rejection test is applied. This hyperball sampling yields shorter signatures and keys because it more efficiently fills the acceptance region, but at the price of a more expensive and memory-intensive sampler (see Section 4.4).

In ML-DSA, the public key is $(\rho, \mathbf{t}_1)$, where $\mathbf{t}_1$ is the high-order part of $\mathbf{A}\mathbf{s}_1 + \mathbf{s}_2$. The matrix seed ρ is stored explicitly, and the secret key additionally holds $\mathbf{t}_0$ (the low-order residue) to enable efficient signing. In HAETAE, the public key is $(\text{seed}_{\mathbf{A}}, \mathbf{b}_1)$, where $\mathbf{b}_1$ is the high-order part of a related lattice commitment $\mathbf{b}$. The HAETAE secret key is more compact; it stores a single short vector $\mathbf{s}$ of augmented dimension $k + \ell + 1$ and a nonce K—because the lattice relation allows the signer to recover the full signing basis on-the-fly.

3 Minimum-State Memory Estimates for HAETAE

Before describing the concrete implementation techniques in Section 4, we analyze the minimum-state memory footprint of each HAETAE routine by counting only the dominant ring objects (polynomials and polynomial vectors) while ignoring constant-size overheads such as hash states, scalar variables, and call frames. This analysis clarifies the relationship between the number of simultaneously live large ring objects and the achievable peak stack. We call these *minimum-state estimates* rather than lower bounds: they count only dominant ring objects, so implementation techniques that pack state more tightly can fall below them (Section 4).

All three HAETAE parameter sets use N=256; a single polynomial with 32-bit coefficients occupies $|\texttt{poly}| = 1\,024$ B. The public matrix $\mathbf{A} \in R_q^{k\times\ell}$ has k rows and ℓ columns; a `polyveck` (k polynomials) and `polyvecl` (ℓ polynomials) scale linearly. For HAETAE-5 (k=4, ℓ=7) these are 4 096 B and 7 168 B respectively. We summarize the HAETAE workflow in Algorithm 5; it is intentionally schematic since our implementation keeps the mathematics unchanged while reorganizing the order in which objects are materialized. We use the HB/LB abbreviations introduced in Section 2.1.

For key generation, the reference implementation materializes $\mathbf{A}$, $\mathbf{s}_1$, and $\mathbf{b}$ simultaneously ($\approx (k\ell+\ell+k) \cdot |\texttt{poly}|$). Since both $\mathbf{A}$ and $\mathbf{s}_1$ are deterministically expandable from seeds, the matrix–vector product can be streamed with $2\cdot|\texttt{poly}|$ (one row accumulator plus one sampling scratch). However, the singular-value norm check $\mathcal{N}(\mathbf{s}_1, \mathbf{s}_2) \le \gamma^2 N$ requires an FFT-based computation with a 2 048 B workspace, giving a total minimum-state estimate of $2\cdot|\texttt{poly}| + 2\,048 = 4\,096$ B.

For signature generation, the reference implementation keeps $\mathbf{y}_1$ (ℓ polys) together with $\mathbf{y}_2$ and $\mathbf{w}$ ($2k$ polys) live simultaneously, totaling $\approx (\ell+2k) \cdot |\texttt{poly}|$ (e.g. 15 360 B for HAETAE-5). With standard streaming, $\mathbf{y}$ and $\mathbf{w}$ can be regenerated row by row from seeds, but the full hint buffer $\mathbf{h} \in R^k$, the high-bits array $\mathsf{HB}^{z_1}(\mathbf{z}_1) \in \mathbb{Z}^{\ell\times N}$, and the challenge polynomial c must still be materialized before the rANS encoder can process them. In particular, c remains live throughout the rejection test and the packing phase, coexisting with the hint and streaming buffers. This gives a minimum-state estimate of $(k+3) \cdot |\texttt{poly}| + \ell\cdot N$ bytes (e.g. $7\,168 + 1\,792 = 8\,960$ B for HAETAE-5).

Lastly, for verification, a column-streamed approach accumulates the matrix product $\tilde{\mathbf{w}} = \widehat{\mathbf{A}} \circ \mathsf{NTT}(\tilde{\mathbf{z}}_1)$ into a full `polyveck` buffer while sweeping $\tilde{\mathbf{z}}_1$ once. The challenge

Algorithm 5 Schematic HAETAE (Key generation, Signing, Verification)

procedure KEY GENERATION(1^λ)
  Sample seeds ρ, κ; sample $(\mathbf{s}_1, \mathbf{s}_2)$
  Reject if $\mathcal{N}(\mathbf{s}_1, \mathbf{s}_2) > \gamma^2 N$ ▷ FFT-based singular-value norm
  Expand $\rho \mapsto \mathbf{A}$; compute $\mathbf{b} \leftarrow \mathbf{A}\mathbf{s}_1 + \mathbf{s}_2 \bmod q$; output (pk, sk)
end procedure
procedure SIGNING(sk, M)
  Derive $\mu \leftarrow H(pk, M)$
  repeat
    Sample $(\mathbf{y}_1, \mathbf{y}_2)$; compute $\mathbf{w} \leftarrow \mathbf{A}\lfloor \mathbf{y}_1 \rceil \bmod 2q$
    $c \leftarrow \mathsf{SampleBinaryChallenge}_\tau(H(\mathsf{HB}^h(\mathbf{w}'), \mathsf{LSB}(\lfloor y_{1,1} \rceil), \mu))$
    $\mathbf{z} \leftarrow \mathbf{y} + (-1)^b c \star \mathbf{s}$
  until $\mathbf{z}$ passes rejection tests
  $\mathbf{h} \leftarrow \mathsf{HB}^h(\mathbf{w}') - \mathsf{HB}^h(\mathbf{w}' - 2\lfloor \mathbf{z}_2 \rceil)$ ▷ post-acceptance
  $\sigma \leftarrow (\mathsf{HB}^{z_1}(\lfloor \mathbf{z}_1 \rceil), \mathsf{LB}^{z_1}(\lfloor \mathbf{z}_1 \rceil), \mathbf{h}, c)$
end procedure
procedure VERIFICATION(pk, M, σ)
  Parse σ into $(\mathsf{HB}^{z_1}, \mathsf{LB}^{z_1}, \mathbf{h}, c)$; reconstruct $\tilde{\mathbf{z}}_1$
  $\tilde{\mathbf{w}} \leftarrow \mathbf{A}\tilde{\mathbf{z}}_1 \bmod 2q$; recompute c' from $(\tilde{\mathbf{h}} + \mathsf{HB}^h(\tilde{\mathbf{w}}'), \mu)$
  Accept iff $c'{=}c$ and norm bounds hold
end procedure

Table 2: Minimum-state estimates vs. measured stack on `pqm4` (HAETAE-5, bytes). Estimates count dominant ring objects and staging buffers; constant-size overheads (hash states, scalars, call frames) are excluded. Full performance results are in Table 3.

	Minimum-state estimate	[HCK+26]	Ours
Key generation	$2 \times 1\,024 + 2\,048 = 4\,096$	5 212	**4 776**
Signing	$(k{+}3) \times 1\,024 + \ell{\cdot}N = 8\,960$	8 092	**6 004**
Verification	$(k{+}2) \times 1\,024 = 6\,144$	6 220	**4 840**

polynomial c must also remain live for the final comparison, giving a minimum-state estimate of $(k{+}2) \cdot |\texttt{poly}|$ (e.g. 6 144 B for HAETAE-5).

Table 2 summarizes the minimum-state estimates for HAETAE-5; these estimates count dominant ring objects and staging buffers while excluding seeds, hash states, and other implementation-specific overheads. The same analysis applies to HAETAE-2/3 with adjusted (k, ℓ) values, though differences in the public-key domain ($d{>}0$ requires an additional rounding step and $\mathbf{a}$-vector) and the placement of the rejection loop affect the concrete figures. The prior work of [HCK+26] reports HAETAE-5 stack usage close to these estimates: 5,212 B (Key generation), 8,092 B (Signing), and 6,220 B (Verification). In the next section we introduce additional techniques, including rejection-aware pass decomposition, reverse-order streaming entropy coding, and row-streamed verification, that push the stack footprint below these minimum-state estimates for signing and verification. Section 5.1 confirms that the resulting implementation achieves lower stack usage while retaining competitive cycle counts relative to [HCK+26].

4 Low-footprint HAETAE Memory Techniques

In this section, we describe the main techniques we use to reduce the memory footprint of HAETAE key generation, signing, and verification. Our approach is guided by the principle of trading increased computation time for reduced peak stack footprint, which is essential for predictable execution on memory-constrained microcontrollers. We also leverage the structure of HAETAE's algorithms, such as the use of seed-derived objects and sparse challenges, to enable streaming and in-place computation strategies that minimize stack usage. We use a single 32-bit (`int32_t`) representation for all working polynomials.

A 16-bit representation would halve |poly| to 512 B, and the reduced coefficients of R_q do fit in 16 bits ($q = 64{,}513$); the intermediate values do not. Signing computes $\mathbf{w} = \mathbf{A}\lfloor \mathbf{y}_1 \rceil \bmod 2q$, and $2q = 129{,}026$ already exceeds 16 bits. We still use narrower types when they suffice: verification keeps the high bits $\mathsf{HB}^{z_1}(\mathbf{z}_1)$ in an `int8_t`$[N]$ buffer and the hint in a `uint16_t`$[N]$ buffer (Section 4.3). Narrowing the polynomial coefficients would also not reduce the peak much, since it is set by the sampler's 2 048 B fixed-point FFT scratch rather than the coefficient width.

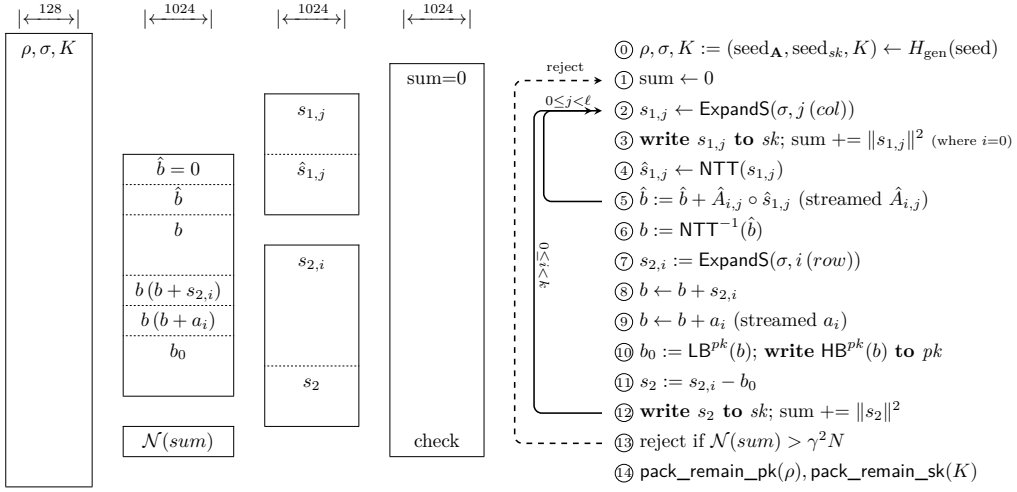

Figure 2: Streamed memory layout of HAETAE-2,3 key generation (d>0). Two polynomial slots (`poly b, s`) and one norm accumulator (`sum[N]`) are reused across the streamed $\widehat{\mathbf{A}}$ expansion. Dashed arrow: rejection loop.

4.1 Low-Stack Key Generation via Streaming

When key generation for digital signatures on constrained devices is required to be performed at runtime, it can no longer be circumvented with a one-time provisioning step. Practical applications include generating ephemeral keys for firmware integrity verification and implementing key rotation to mitigate compromise.

The baseline key generation is specified in Algorithms 1 and 2. We describe our stack optimizations in two parts. First we present the streaming techniques applied to HAETAE-2/3 ($d > 0$), which involve rounding and an additional $\mathbf{a}$-vector not present in the d=0 case; this streaming of key generation for $d > 0$ is a new contribution not covered in previous work. Figure 2 illustrates the resulting memory layout. We then describe our HAETAE-5 ($d = 0$) implementation, which follows the same high-level approach as [HCK+26], but further compresses the stack through caller-level union analysis.

Streaming matrix–vector multiplication. Key generation computes $\mathbf{b} = \mathbf{a} + \mathbf{A}\mathbf{s}_1 + \mathbf{s}_2 \pmod q$, where $\mathbf{A} \in R_q^{k\times\ell}$ is derived from a seed and $\mathbf{s}_1 \in R_q^{\ell}$, $\mathbf{s}_2 \in R_q^{k}$ are short secret vectors. The reference implementation materializes $\mathbf{A}$ and $\widehat{\mathbf{s}}_1$ as full polynomial vectors, causing large stack pressure. We reorganize the computation so that $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{A}$ are all generated and consumed on-the-fly (steps ②–⑫ in Figure 2): for each column j we sample $\mathbf{s}_{1,j}$ into a single scratch polynomial (②), apply the NTT in place (④), and for each row i generate $\mathbf{A}_{i,j}$ from the seed and immediately accumulate the pointwise product into $\mathbf{b}_i$ (⑤). The secret vectors are written to the secret key as they are produced (③, ⑫), and a norm accumulator (`int32_t sum[N]`) tracks the singular-value check incrementally (⑬); $\mathbf{s}_1$ norms are accumulated only at i=0 to avoid redundant computation in the inner loop.

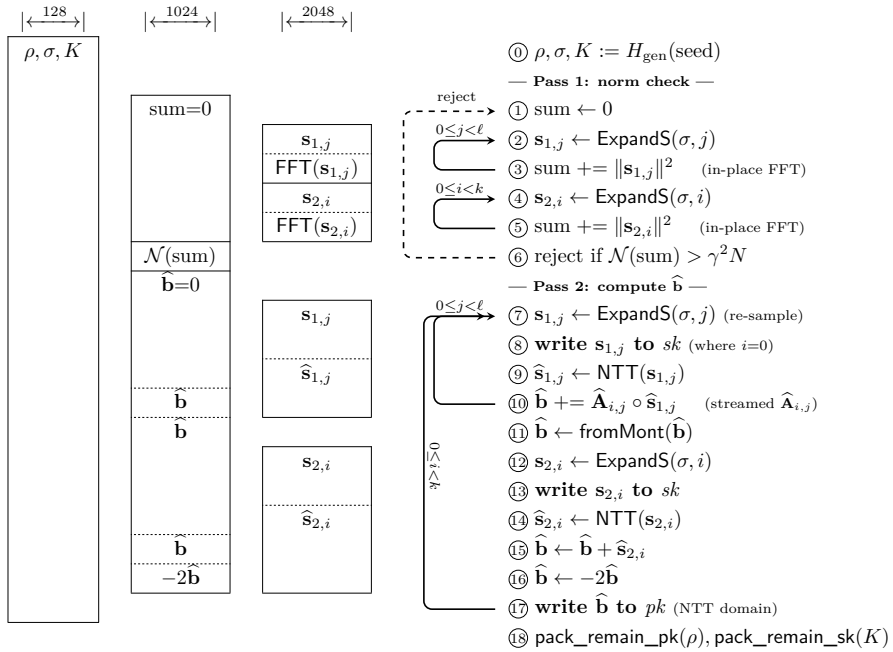

Figure 3: Streamed memory layout of HAETAE-5 key generation (d=0). Two caller-level unions (1 024 B and 2 048 B) share memory across passes. Pass 1 checks the norm; Pass 2 computes $\widehat{\mathbf{b}} = \mathsf{NTT}(-2\mathbf{b})$.

The entire key generation thus requires only two single-polynomial scratch slots (`poly b, s`) and one norm accumulator on the stack.

Additional stack reductions. SHAKE-based sampling routines (`poly_uniform` and `poly_uniform_eta`) are rewritten to squeeze one block at a time, replacing multi-block stack buffers with a single-block buffer without altering the output distribution. Together with the streaming matrix multiplication, these techniques remove full-vector temporaries and multi-block squeeze buffers, yielding a substantially smaller and more predictable stack profile. Note that streaming and re-computation alter data-access patterns; a thorough side-channel analysis is left for future work.

HAETAE-5 ($d = 0$) key generation. For HAETAE-5, the key generation follows the same high-level streaming approach as prior work [HCK+26]: on-the-fly matrix generation, row-by-row accumulation of $\widehat{\mathbf{b}} = \mathsf{NTT}(-2\mathbf{b})$, and direct storage in the NTT domain. The two-pass structure is illustrated in Figure 3: Pass 1 (norm check, steps ①–⑥) and Pass 2 (matrix–vector product, steps ⑦–⑰). Our implementation additionally introduces two caller-level `union`s that tightly pack all large temporaries into a fixed-size caller frame, eliminating the need for deep callee stack frames. The first union shares the 2 048-byte FFT scratch used for the singular-value norm in Pass 1 (③, ⑤) with the 1 024-byte sampling polynomial (`poly s`) used in both passes (②, ⑦); the FFT is invoked directly in the caller rather than delegated to a callee, so the scratch never appears as a separate stack frame. The second union shares the norm accumulator (`int32_t sum[N]`, 1,024 B) in Pass 1 (①) with the row accumulator (`poly b`, 1,024 B) in Pass 2 (⑩), exploiting the strict lifetime disjointness across the two passes. Together, the two unions compress the dominant temporaries to $2\,048 + 1\,024 = 3\,072$ bytes in the caller frame, with no additional callee overhead for the norm computation. By analyzing the stack frames of each callee

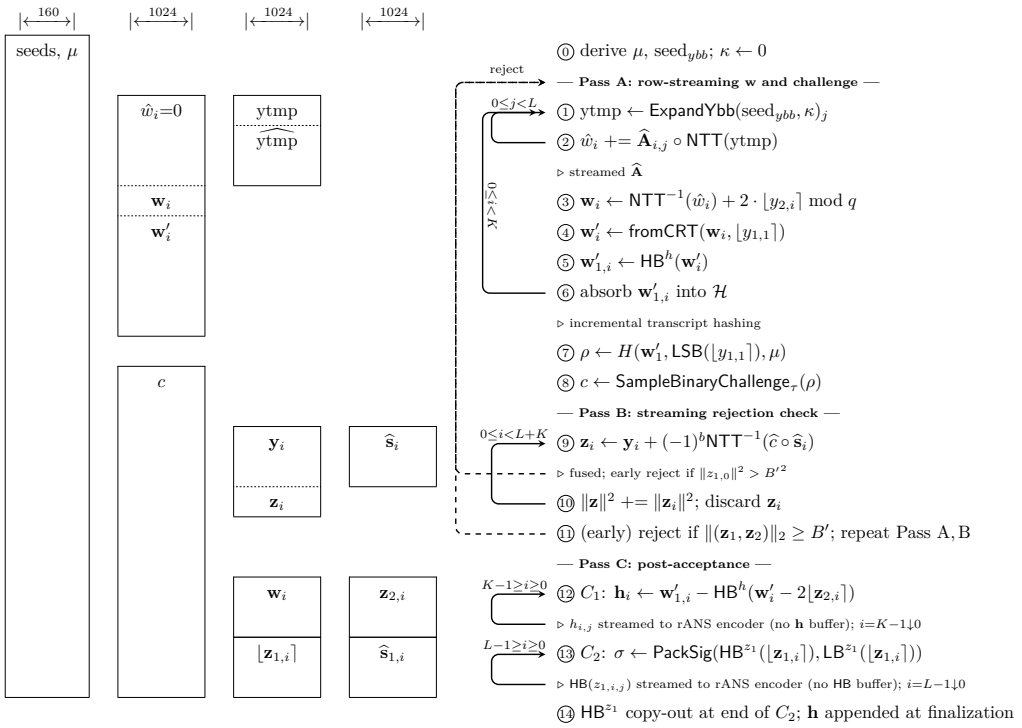

Figure 4: Streamed memory layout of pass-decomposed HAETAE signing. The driver keeps seeds and μ (160 B) plus three 1 024 B polynomial slots reused across passes. Dashed arrow: rejection loop (Pass A + B); Pass C runs once after acceptance. C_1/C_2 use reverse-order loops to stream directly to the rANS encoder.

and identifying lifetime-disjoint buffers across the two passes, this union-based layout reduces the key-generation stack by 436 B compared to [HCK+26] (4,776 B vs. 5,212 B on the same `pqm4` framework).

4.2 Low-Stack Signing via Pass Decomposition and Streaming

Recall that HAETAE signing (Algorithm 3) mainly consists of a rejection-sampling loop: each iteration draws ephemeral vectors $(\mathbf{y}_1, \mathbf{y}_2)$ from a hyperball distribution, computes the matrix–vector product $\mathbf{w} \leftarrow \mathsf{NTT}^{-1}(\widehat{\mathbf{A}} \circ \mathsf{NTT}(\lfloor \mathbf{y}_1 \rceil)) + 2\lfloor \mathbf{y}_2 \rceil \bmod q$, derives a challenge c, forms the response $\mathbf{z} = \mathbf{y} + (-1)^b c \star \mathbf{s}$, and repeats until the norm bounds are satisfied. The reference implementation [CCD+23] keeps $\mathbf{w}$, $\mathbf{z}$, and the hint $\mathbf{h}$ live simultaneously, leading to high peak stack usage.

Our design follows the [BRS22] principle of trading recomputation for reduced peak memory, and adopts the seed-based on-the-fly matrix streaming used by [HCK+26]: each entry of $\widehat{\mathbf{A}}$ is regenerated from the public seed rather than stored in RAM. Beyond these shared foundations, we introduce several techniques not described by [HCK+26]: (i) a *Rejection-aware pass decomposition* that confines encoding and hint computation to a post-acceptance path; (ii) *Component-level early rejection* that short-circuits the response computation in Pass B when a partial norm already exceeds the bound; and (iii) *Reverse-order streaming entropy coding* that eliminates full hint and high-bit staging buffers by aligning computation order with the rANS encoder's backward symbol consumption. Additionally, our signing path achieves zero mutable BSS (`.bss` = 0 B) through a streaming Gaussian backend (Section 4.4). Moreover, while [HCK+26] targets HAETAE-5 only, our

design supports all three security levels including HAETAE-2/3 (d>0). Figure 4 illustrates the resulting memory layout.

Rejection-aware pass decomposition. We decompose the signing loop into three passes whose large buffers have strictly non-overlapping lifetimes:

- **Pass A** (steps ①–⑧ in Figure 4) samples $(\mathbf{y}_1, \mathbf{y}_2)$ via the two-pass hyperball sampler (Section 4.4), computes $\mathbf{w}$ row by row using on-the-fly matrix streaming, then derives $\mathbf{w}'_1 = \mathsf{HB}^h(\mathsf{fromCRT}(\mathbf{w}, \lfloor y_{1,1} \rceil))$, and obtains the challenge $c \leftarrow \mathsf{SampleBinaryChallenge}_\tau\big(H(\mathbf{w}'_1, \mathsf{LSB}(\lfloor y_{1,1} \rceil), \mu)\big)$. The single-polynomial row accumulator for $\mathbf{w}$ reuses the output location of c (overwritten only at the end of the pass), eliminating a dedicated scratch buffer. For d>0 (HAETAE-2/3), the first column of $\widehat{\mathbf{A}}$ is derived on-the-fly from the packed public key via in-place CRT reconstruction, requiring no additional polynomial buffer.

- **Pass B** (⑨–⑪) regenerates $(\mathbf{y}_1, \mathbf{y}_2)$ from the stored seed and accepted nonce, computes $\mathbf{z} = \mathbf{y} + (-1)^b c \star \mathbf{s}$ one component at a time via fused accumulation (described below), and evaluates the rejection tests. No matrix product is needed, since the ℓ_2-norm and ℓ_∞ tests involve only $\mathbf{z}$, $\mathbf{y}$, and b'. Each component is discarded after its norm contribution is accumulated.

- **Pass C** (⑫–⑭) runs only once, after acceptance. It recomputes $\mathbf{z}$ and $\mathbf{w}$ to derive the hint $\mathbf{h} \leftarrow \mathbf{w}'_1 - \mathsf{HB}^h(\mathbf{w}' - 2\lfloor \mathbf{z}_2 \rceil) \bmod^+ \frac{2(q-1)}{\alpha_h}$ and packs the signature. The variable-length payloads $\mathbf{h}$ and $\mathsf{HB}^{z_1}(\mathbf{z}_1)$ are encoded via reverse-order streaming rANS (described below). Pass C is further split into two sub-phases (C_1, hint encoding, and C_2, $\mathbf{z}_1$ packing) with non-overlapping buffer lifetimes.

Because the rejection loop comprises only Pass A and Pass B, all encoding and hint computation are deferred to the post-acceptance path. The peak signing stack is therefore

$$S_{\text{sign}} \;=\; S_{\text{driver}} + \max\big(S_A,\; S_B,\; S_{C_1},\; S_{C_2}\big), \tag{1}$$

where S_{driver} is the persistent driver state (seeds, nonce counter, hash prefix) and the four terms are the transient allocations of each pass. Our decomposition guarantees that the peak follows Equation (1) by compiling the passes as a separate translation unit, so that no pass is inlined into the driver frame, confining encoding and hint computation to the single post-acceptance execution.

In contrast, the minimum-state estimate from Section 3 requires the full hint and high-bits arrays to be materialized before encoding:

$$S_{\text{sign}}^{\text{conv}} \;=\; (k{+}3) \cdot |\texttt{poly}| + \ell \cdot N.$$

Our pass-aware design replaces this with Equation (1), where for HAETAE-5:

$$S_{C_1} \approx 3 \cdot |\texttt{poly}| + B_h, \qquad S_{C_2} \approx 2 \cdot |\texttt{poly}| + B_{hb},$$

with B_h and B_{hb} denoting the compact streaming encoder buffers (bounded by the base entropy of $\mathbf{h}$ and $\mathsf{HB}^{z_1}(\mathbf{z}_1)$ respectively), which are substantially smaller than the full staging arrays they replace. This means we are even below the minimum-state estimate. Since S_{C_1} and S_{C_2} dominate the peak while Pass A requires only a single polynomial accumulator, the pass-aware structure leaves unused stack budget in Pass A. We exploit this by batching two matrix rows per iteration of the inner product, halving the number of ephemeral-vector regenerations from $k\ell$ to $\lceil k/2 \rceil \cdot \ell$ at the cost of one additional polynomial accumulator (+1 024 B in S_A), without increasing the overall peak S_{sign}.

Component-level early rejection. The reference implementation computes the full response vector $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$ before evaluating any rejection condition. Adapting the early-abort idea of [BRS22] to HAETAE, we add component-level early exits that reduce the average cost of rejected iterations in Pass B (steps ⑨–⑩ in Figure 4). The first component $z_{1,0} = y_{1,0} + (-1)^b \cdot c$ involves no secret-key material (the challenge c is derived from public data), so a data-dependent branch on its norm is constant-time safe. Pass B therefore computes $z_{1,0}$ before any NTT-based multiplication (⑨) and rejects immediately if $\|z_{1,0}\|^2 > B'^2$ (dashed arrow between ⑨ and ⑩), skipping the remaining $\ell{+}k{-}1$ fused multiply-accumulate operations. During the subsequent per-component norm accumulation (⑩) $\|\mathbf{z}\|^2 = \sum_i \|z_i\|^2$, we check the running partial sum after each polynomial and skip the remaining components once the bound is exceeded. Additionally, the hyperball scaling factors, which are deterministic functions of the seed and nonce counter, are computed once in the driver frame and passed by pointer to Passes B, C_1, and C_2, eliminating three redundant recomputations per accepted iteration.

Reverse-order streaming entropy coding. The signature includes two variable-length rANS-encoded payloads: the hint $\mathbf{h} \in \mathbb{Z}^{k \times N}$ and the high-bit decomposition $\mathsf{HB}(\mathbf{z}_1) \in \mathbb{Z}^{\ell \times N}$. Conventional implementations materialize these as full arrays before encoding; for HAETAE-5, a `polyveck` hint buffer ($k{\times}\,1\,024\,\mathrm{B} = 4\,096\,\mathrm{B}$) and an `int8_t` array ($\ell{\times}\,N = 1\,792\,\mathrm{B}$) for $\mathsf{HB}(\mathbf{z}_1)$.

rANS is a *backward* entropy coder: for a flat array $X[0], \ldots, X[m{-}1]$, the encoder consumes symbols in the order $X[m{-}1], X[m{-}2], \ldots, X[0]$. The conventional approach first materializes the entire array, then iterates backward:

$$\mathsf{RansEncPut}(X[m{-}1]),\ \mathsf{RansEncPut}(X[m{-}2]),\ \ldots,\ \mathsf{RansEncPut}(X[0]).$$

We show that for a row-major $R \times C$ array ($m = R \cdot C$), the identical reverse sequence can be produced without materializing the full array, by traversing rows from $i = R{-}1$ down to 0 and coefficients from $j = C{-}1$ down to 0 within each row:

$$\begin{aligned}&\text{for } i = R{-}1 \text{ downto } 0:\\ &\quad \text{for } j = C{-}1 \text{ downto } 0:\\ &\qquad \mathsf{RansEncPut}\big(X[i \cdot C + j]\big).\end{aligned}$$

This allows each coefficient to be fed to the encoder immediately upon computation, without materializing the full $R \times C$ array. In Pass C_1 (⑫ in Figure 4), each hint row $\mathbf{h}_i$ ($R{=}k$, $C{=}N$) is recomputed via row-streaming and immediately encoded; in Pass C_2 (⑬), $\mathsf{HB}^{z_1}(\mathbf{z}_{1,i})$ ($R{=}\ell$, $C{=}N$) is handled analogously. Both full staging buffers are eliminated entirely.

The encoded output is accumulated in a compact stack-local encoder context and copied to the signature buffer only after encoding completes, following an *internal encoding with copy-out* boundary model.

Lemma 1. *For a fixed configuration, the reverse-order streaming encoder is byte-identical to the full-array encoder.*

Proof. For a fixed rANS configuration (tables, normalization, and flush/output convention), encoding is deterministic in the symbol sequence. The full-array and reverse row-major streaming encoders both feed entries in index order $m{-}1, m{-}2, \ldots, 0$, and each streamed entry $X[i \cdot C + j]$ is recomputed from the same accepted seed/nonce, signing state, and rounding/decomposition routines, so it equals the materialized array entry. The two encoders therefore receive identical symbol sequences and emit identical bytes. □

As a regression check, differential tests (HAETAE-2/3/5, 10 000 signatures each) found byte-identical streams.

Incremental transcript hashing. Both signing (Pass A, steps ⑤–⑥ in Figure 4) and verification (⑧–⑨ in Figure 5) derive the challenge polynomial by hashing a transcript that includes the packed high bits of the matrix–vector product. The reference implementation allocates a contiguous buffer of `POLYVECK_HIGHBITS_PACKEDBYTES + POLYC_PACKEDBYTES` bytes (608–1 184 B depending on the security level), packs all k rows into it, and then feeds it to SHAKE256 in one shot. Because SHAKE256 is a sponge and satisfies $\text{absorb}(A\|B) = \text{absorb}(A); \text{absorb}(B)$, we replace this monolithic buffer with *row-by-row incremental absorb*: each row's high bits are packed into a small single-row buffer (`POLY_HIGHBITS_PACKEDBYTES`), immediately absorbed into the SHAKE256 state, and discarded before the next row is processed. After the row loop, the remaining fields ($\mathbf{w}'$ and the message digest μ) are absorbed and the state is finalized. This eliminates the full transcript buffer entirely, yielding measurable stack savings (up to 656 B for HAETAE-5).

Sparse challenge multiplication with fused accumulation. In Pass B (⑨ in Figure 4), for HAETAE-2/3, where the challenge $c(X) = \sum_{t=1}^{\tau} X^{i_t}$ has Hamming weight τ=58 (HAETAE-2) or τ=80 (HAETAE-3), we replace NTT-based multiplication by a purely coefficient-domain *signed shift-and-add* rule in the negacyclic ring $R_q = \mathbb{Z}_q[X]/(X^N+1)$:

$$(c \star s)[j] \;=\; \sum_{t=1}^{\tau} \varepsilon_t \cdot s[(j - i_t) \bmod N], \qquad \varepsilon_t = \begin{cases} +1 & j \geq i_t, \\ -1 & j < i_t, \end{cases}$$

where ε_t arises from the negacyclic relation $X^N = -1$. For HAETAE-5 (τ=128, dense binary challenge), we retain NTT-based multiplication as the dense weight makes a coefficient-domain approach less efficient. In all cases we fuse the accumulation directly into the response: each signed shift (or partial product) is added in place to the corresponding component of $\mathbf{y}$, producing $\mathbf{z}$ without allocating a separate product polynomial (saving 1,024 B per multiplication).

Algorithm-level BSS elimination. Our complete signing path uses no mutable BSS (`.bss` = 0 B). The hyperball sampler's static temporary buffers (`tmp_samples[N+1]` and `tmp_signs[(N+7)/8]`, totaling 2,088 B in the reference implementation) are replaced by a fully streaming Gaussian backend that generates samples on-the-fly from the SHAKE sponge state, as detailed in Section 4.4. The pass decomposition requires no static workspace, so the signing path reported in Table 3 achieves `.data` = 0 and `.bss` = 0.

4.3 Low-Stack Verification via View-Style Decoding and Row Streaming

HAETAE verification (Algorithm 4) reconstructs $\tilde{\mathbf{w}}'_1 = \tilde{\mathbf{h}} + \mathsf{HB}^h(\tilde{\mathbf{w}}')$ from $\tilde{\mathbf{w}} = \widehat{\mathbf{A}} \circ \mathsf{NTT}(\tilde{\mathbf{z}}_1)$ (up to the CRT lift mod $2q$), derives the auxiliary component $\tilde{\mathbf{z}}_2$, and re-hashes the packed transcript $(\tilde{\mathbf{w}}'_1, w')$ to recompute the challenge. Rather than following the column-streamed approach of [HCK+26], our verification is built on the [BRS22] method of trading recomputation for transient memory, combining (i) *view-style decoding* with union overlays, (ii) *row-streamed* matrix–vector multiplication, and (iii) *incremental transcript hashing* that eliminates the full-transcript buffer (the same technique introduced for signing in Section 4.2). This design supports all three security levels, including HAETAE-2/3 (d>0), which is not addressed by [HCK+26], and achieves a verification stack of 4,824 B–4,848 B across all security levels — for HAETAE-5, a 22 % reduction (4,840 B vs. 6,220 B) compared to [HCK+26]. Figure 5 illustrates the resulting memory layout.

View-style decoding. We avoid materializing the decoded signature vectors (steps ⓪–② in Figure 5). Concretely, we keep $\mathsf{LB}^{z_1}(\lfloor \mathbf{z}_1 \rceil)$ as a pointer into the signature buffer, decode only $\mathsf{HB}^{z_1}(\lfloor \mathbf{z}_1 \rceil)$ into an `int8` array (①), and decode $\mathbf{h}$ into a compact `uint16` array (⑦). These

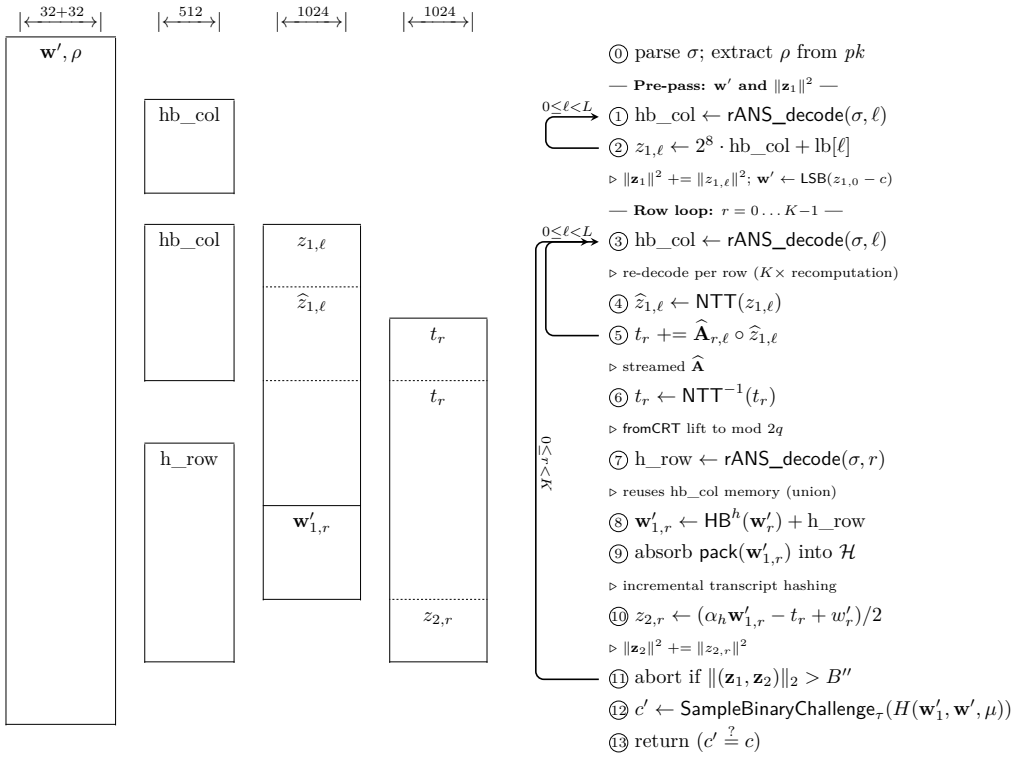

Figure 5: Streamed memory layout of row-streamed HAETAE verification. Four slots: persistent buffers (64 B), a 512 B union alternating high-bits columns and hint rows, and two 1 024 B polynomial slots. Each row re-decodes $\mathsf{HB}^{z_1}(\mathbf{z}_1)$ via streaming rANS, trading factor-k re-computation for eliminating a full `polyveck` buffer.

two decoded arrays share a `union` overlay (`int8_t hb_col[N]` / `uint16_t h_row[N]`), exploiting lifetime disjointness: the high-bits columns are consumed during the inner column loop, after which the same memory is reused for the decoded hint row. We also represent $w' \in \mathcal{R}_2$ as a packed bitstring of $N/8$ bytes. A pre-pass (①–②) over coefficients recomposes $\tilde{z}_{1,\ell} = \mathsf{HB}^{z_1} \cdot 256 + \mathsf{LB}^{z_1}$ on-the-fly, accumulates $\|\tilde{\mathbf{z}}_1\|_2^2$, and derives $w' = \mathsf{LSB}(\tilde{z}_{1,1} - c)$. Together, the working set for decoded data is a single 512-byte union plus a 32-byte packed bitstring, independent of the module dimensions k and ℓ. Signatures that fail format checks (length, zero-padding, or rANS decoder errors) are rejected before any transcript recomputation.

Row-streamed multiplication. To minimize peak stack, we compute $\tilde{\mathbf{w}}$ one row at a time rather than column-by-column (steps ③–⑪ in Figure 5). For each output row r, we recompute an NTT-domain accumulator $\tilde{w}_r \leftarrow \sum_{\ell=0}^{L-1} \widehat{\mathbf{A}}[r, \ell] \circ \mathsf{NTT}(\tilde{z}_{1,\ell})$ (③–⑤) by (re)composing and transforming one column $\tilde{z}_{1,\ell}$ at a time and streaming $\widehat{\mathbf{A}}[r, \ell]$ from the public seed ρ. After an inverse NTT and CRT lift via $\mathsf{fromCRT}$ (⑥), we decode the hint row $\tilde{\mathbf{h}}_r$ (⑦) and form $\tilde{w}'_{1,r} = \tilde{\mathbf{h}}_r + \mathsf{HB}^h(\tilde{w}'_r)$ (⑧), then immediately absorb the packed high bits into the incremental SHAKE256 transcript state (⑨). Within the same loop we compute $\tilde{z}_{2,r}$ (⑩), accumulate $\|\tilde{z}_{2,r}\|_2^2$, and abort early as soon as the bound test fails (⑪). This reduces the working set to a constant number of polynomials plus small decoded arrays, eliminating concurrent `polyveck` temporaries. The tradeoff is a factor-k increase in re-computation: each row re-decodes the ℓ columns of $\mathsf{HB}^{z_1}(\lfloor \mathbf{z}_1 \rceil)$ via streaming rANS and re-applies ℓ forward NTTs.

Algorithm 6 Streamed hyperball sampler

Require: Gaussian dimension N, integers L, K; seed $\mathsf{seed} \in \{0,1\}^{\mathsf{CRHBYTES}}$; nonce $\mathsf{nonce} \in \{0,1\}^{16}$; hyperball bound B_0 and scaling constant Λ (e.g. $\Lambda = L + K$)
Ensure: $(\mathbf{y}_1, \mathbf{y}_2) \in \mathbb{Z}^{L\times N} \times \mathbb{Z}^{K\times N}$ with $\|(\mathbf{y}_1, \mathbf{y}_2)\|_2 \le B_0\Lambda$

repeat
  ▷ Pass 1: compute the (unnormalized) squared norm S
  `InitStream(st, seed, nonce)`
  $S \leftarrow 0$ ▷ S is a scalar accumulator (e.g. 64-bit integer)
  for $i \leftarrow 0$ to $L + K - 1$ **do**
    for $j \leftarrow 0$ to $N - 1$ **do**
      $x_{i,j} \leftarrow \mathtt{SampleGauss}(\mathsf{st})$
      $S \leftarrow S + x_{i,j}^2$
    end for
  end for
  ▷ Compute scaling factor $\alpha \approx \dfrac{B_0\Lambda}{\sqrt{S}}$
  $\alpha \leftarrow \mathtt{InvSqrt}(S, B_0, \Lambda)$
  ▷ Pass 2: regenerate samples and scale immediately
  `InitStream(st, seed, nonce)`
  for $i \leftarrow 0$ to $L - 1$ **do**
    for $j \leftarrow 0$ to $N - 1$ **do**
      $x_{i,j} \leftarrow \mathtt{SampleGauss}(\mathsf{st})$
      $y_{1,i,j} \leftarrow \mathtt{Round}(\alpha \cdot x_{i,j})$
    end for
  end for
  for $i \leftarrow 0$ to $K - 1$ **do**
    for $j \leftarrow 0$ to $N - 1$ **do**
      $x_{L+i,j} \leftarrow \mathtt{SampleGauss}(\mathsf{st})$
      $y_{2,i,j} \leftarrow \mathtt{Round}(\alpha \cdot x_{L+i,j})$
    end for
  end for
until $\|(\mathbf{y}_1, \mathbf{y}_2)\|_2^2 \le B_0^2\Lambda^2$

In contrast, [HCK+26] uses a *column-streamed* approach that accumulates $\tilde{\mathbf{w}}$ into a full `polyveck` buffer ($k \times 1\,024\,\mathrm{B}$) while sweeping the columns of $\tilde{\mathbf{z}}_1$ once. Our row-streamed design trades this k-fold re-decode cost for a $(k-1) \times 1\,024\,\mathrm{B}$ reduction in peak live memory. Combined with the incremental transcript hashing described in Section 4.2, which absorbs each row's high bits directly into the SHAKE256 state, the full `POLYVECK_HIGHBITS_PACKEDBYTES` transcript buffer (up to $1\,184\,\mathrm{B}$) is eliminated entirely, an optimization not applied by either the reference implementation or [HCK+26].

For $d>0$ (HAETAE-2/3), the first column of $\widehat{\mathbf{A}}$ is derived on-the-fly from the public seed and the packed public key, adding one polynomial expansion per row but no additional persistent storage.

In the second pass, the sampler restarts the pseudorandom stream from the same $(\mathsf{seed}, \mathsf{nonce})$ pair, so that $\mathtt{SampleGauss}(\mathsf{st})$ reproduces the identical sequence $(x_{i,j})$. This time, each sample is immediately rescaled and rounded,

$$y_{1,i,j} = \mathtt{Round}(\alpha x_{i,j}) \quad (0 \le i < \ell), \qquad y_{2,i,j} = \mathtt{Round}(\alpha x_{\ell+i,j}) \quad (0 \le i < k),$$

and written to the output polynomials $\mathbf{y}_1$ and $\mathbf{y}_2$. The algorithm then evaluates the squared norm $\left\|(\mathbf{y}_1, \mathbf{y}_2)\right\|_2^2$ and rejects if it exceeds $B_0^2\Lambda^2$, repeating the entire two-pass procedure until acceptance.

4.4 A Memory-Friendly Sampler

The hyperball sampler is invoked in Pass A of signing (① in Figure 4) to draw the ephemeral vectors $(\mathbf{y}_1, \mathbf{y}_2)$ used in each rejection-sampling iteration. We require a sampler that outputs a random vector

$$(\mathbf{y}_1, \mathbf{y}_2) \in \mathbb{Z}^{\ell \times N} \times \mathbb{Z}^{k \times N},$$

whose distribution is (approximately) Gaussian, conditioned on the norm constraint

$$\left\|(\mathbf{y}_1, \mathbf{y}_2)\right\|_2 \leq B_0 \Lambda,$$

for some bound $B_0 > 0$ and scaling factor Λ (in our case $\Lambda = \ell + k$). A naïve implementation samples all $(\ell + k)N$ Gaussian coefficients, stores them in memory, computes their squared norm, rescales the entire vector, and finally applies a rejection test. This requires $\Theta((\ell+k)N)$ words of transient storage. The two-pass approach that avoids this large buffer was first described by [LWKP25] for BLISS and is also adopted by [HCK+26] for HAETAE-5. We apply the same high-level structure but further eliminate all static (BSS) scratch buffers through a fully streaming Gaussian backend.

Algorithm 6 implements what we call the *two-pass* sampler with *streaming* Gaussian sampling and it only uses $O(1)$ additional memory. The sampler is parameterized by a seed $\mathsf{seed} \in \{0,1\}^{\mathsf{CRHBYTES}}$ and a nonce nonce, which are used to initialize a pseudorandom stream st (e.g. based on SHAKE or $\mathsf{stream256}$).

In the first pass, the algorithm draws $(\ell + k)N$ independent discrete Gaussian samples $(x_{i,j})_{0 \leq i < \ell+k,\ 0 \leq j < N}$ from $\texttt{SampleGauss}(\mathsf{st})$, and maintains only the scalar accumulator

$$S \;=\; \sum_{i=0}^{\ell+k-1} \sum_{j=0}^{N-1} x_{i,j}^2.$$

No sample is stored beyond its contribution to S, so the memory footprint of this pass is independent of N, ℓ, and k. After the loop, the algorithm computes a scaling factor

$$\alpha \approx \frac{B_0 \Lambda}{\sqrt{S}},$$

using a fixed-point approximation of the reciprocal square root.

Remark 1. The two-pass sampler reproduces the one-pass reference exactly. For a fixed $(\mathsf{seed}, \mathsf{nonce})$, the deterministic `InitStream` and `SampleGauss` regenerate the same Gaussian sequence $(x_{i,j})$ in both passes, so with the same fixed-point `InvSqrt` and rounding rule they compute the same S, α, rounded $(\mathbf{y}_1, \mathbf{y}_2)$, and acceptance predicate. Each attempt therefore maps $(\mathsf{seed}, \mathsf{nonce})$ identically to an output or a rejection, and nonce retries preserve this equality, so the accepted-output distribution is identical.

Therefore the only difference is implementation strategy: the two-pass sampler never stores more than a bounded number of Gaussian samples at any one time. A further quantitative analysis of the fixed-point approximation error is left to future work.

Streaming Gaussian backend. The reference HAETAE sampler allocates temporary arrays `samples[N·(ℓ+k)]` and `signs[N·(ℓ+k)/8]` in stack or BSS, requiring $\Theta((\ell + k)N)$ words of transient storage. Our implementation replaces this with a fully streaming backend: each discrete Gaussian sample is generated from the SHAKE sponge state, consumed, and discarded immediately. Since squeezing one sample at a time produces the identical output as batch squeezing, this has no effect on the output distribution. This eliminates all static sampler buffers, reducing the scheme-wide `.bss` to zero bytes.

Table 3: Performance comparison on Nucleo-L4R5ZI (pqm4). Cycles: `-O3`, median of 100 runs ($K = 10^3$). Stack/size: `-O3`. Ratio is relative to each scheme's speed baseline—HAETAE: **m4fspeed**, ML-DSA: **m4f** (<1: faster). (C) = pure C; (asm) = assembly. **m4fspeed**/**m4fstack** = reference M4 ports [HCK+26] (speed- and stack-optimized); [HCK+26] (C): published numbers (HAETAE-5 only), measured with `-O3` for cycles but `-Os` for stack and size; every other row uses `-O3` for all metrics. **ours (C)**/**ours (asm)**: this work. ML-DSA from `pqm4`; **m4fstack** follows the strategy of [BRS22].

	Key generation			Signing			Verification				
Impl.	Cycles	Ratio	Stack	Cycles	Ratio	Stack	Cycles	Ratio	Stack	.text	.total
HAETAE-2											
ref (C)	9,253 K	1.30×	23,844	43,710 K	2.71×	73,112	1,732 K	1.89×	33,448	26,490	29,098
m4fspeed (asm)	7,109 K	1.00×	19,764	16,147 K	1.00×	55,552	916 K	1.00×	23,288	30,104	30,624
m4fstack (asm)	6,465 K	0.91×	17,204	30,472 K	1.89×	41,608	916 K	1.00×	23,288	30,104	30,624
ours (C)	11,268 K	1.59×	5,816	92,024 K	5.70×	5,836	1,098 K	1.20×	4,824	30,416	30,416
ours (asm)	9,881 K	1.39×	5,816	63,603 K	3.94×	5,836	852 K	0.93×	4,824	38,100	38,100
HAETAE-3											
ref (C)	12,007 K	0.89×	41,236	53,546 K	2.17×	113,328	3,170 K	1.81×	54,232	26,436	29,204
m4fspeed (asm)	13,524 K	1.00×	29,500	24,622 K	1.00×	83,296	1,756 K	1.00×	31,768	30,078	30,758
m4fstack (asm)	15,396 K	1.14×	22,340	35,388 K	1.44×	58,008	1,756 K	1.00×	31,768	30,078	30,758
ours (C)	22,437 K	1.66×	5,816	115,568 K	4.69×	6,020	2,267 K	1.29×	4,848	30,780	30,780
ours (asm)	15,768 K	1.17×	5,816	120,087 K	4.88×	6,020	1,715 K	0.98×	4,848	38,464	38,464
HAETAE-5											
ref (C)	16,491 K	0.95×	52,500	65,896 K	1.76×	144,408	3,968 K	1.71×	68,856	25,658	28,786
m4fspeed (asm)	17,308 K	1.00×	34,108	37,381 K	1.00×	103,856	2,317 K	1.00×	37,184	29,756	30,796
m4fstack (asm)	13,359 K	0.77×	22,332	50,813 K	1.36×	69,280	2,317 K	1.00×	37,184	29,756	30,796
[HCK+26] (C)	15,328 K	0.89×	5,212	300,881 K	8.05×	8,092	4,187 K	1.81×	6,220	26,494	27,534
ours (C)	10,804 K	0.62×	4,776	172,765 K	4.62×	6,004	2,758 K	1.19×	4,840	30,150	30,150
ours (asm)	12,190 K	0.70×	4,776	166,236 K	4.45×	6,004	2,205 K	0.95×	4,840	37,834	37,834
ML-DSA-44											
m4f (asm)	1,418 K	1.00×	38,312	3,017 K	1.00×	44,832	1,496 K	1.00×	8,912	19,592	19,592
m4fstack (asm)	1,797 K	1.27×	4,408	10,330 K	3.42×	5,064	3,816 K	2.55×	2,720	24,844	24,844
ML-DSA-65											
m4f (asm)	2,526 K	1.00×	60,840	5,963 K	1.00×	68,896	2,533 K	1.00×	9,888	19,328	19,328
m4fstack (asm)	3,422 K	1.35×	4,408	21,668 K	3.63×	6,608	6,771 K	2.67×	2,720	24,120	24,120
ML-DSA-87											
m4f (asm)	4,267 K	1.00×	97,704	7,145 K	1.00×	107,912	4,381 K	1.00×	12,064	19,500	19,500
m4fstack (asm)	5,770 K	1.35×	4,408	27,150 K	3.80×	8,144	11,713 K	2.67×	2,720	24,516	24,516

5 Results on Mid-range Microcontrollers

This section evaluates our low-stack HAETAE implementation. We report primary results on the `pqm4` benchmarking framework [KPR+] (Nucleo-L4R5ZI, ARM Cortex-M4), which enables direct comparison with the prior work of [HCK+26] and with ML-DSA. We additionally validate portability under RIOT-OS on two further targets (nRF52840 and ESP32-C6).

Our implementation is controlled by three compile-time knobs:

```
SAMPLER=2  STREAM_MATRIX=1  FROZEN_A=2
```

`SAMPLER=2` selects the two-pass streaming hyperball sampler (Section 4.4); `STREAM_MATRIX=1` regenerates matrix entries on demand from the public seed; and `FROZEN_A=2` fuses rejection sampling with pointwise multiplication to eliminate temporary polynomial buffers. When all three are enabled, the full-streaming signing path (`FULL_STREAM_SIGN`) is activated automatically, applying the *Rejection-aware pass decomposition* and *Reverse-order streaming entropy coding* techniques described in Section 4.2.

5.1 Results on `pqm4` framework (Nucleo-L4R5ZI)

All measurements are performed on the **Nucleo-L4R5ZI** board (STM32L4R5ZI, ARM Cortex-M4F, 2 MB Flash, 640 kB RAM), the default `pqm4` target and the platform used by [HCK+26]. Speed measurements use `CLOCK_BENCHMARK` (20 MHz, 0 WS flash) for cycle-accurate results; stack measurements use `CLOCK_FAST` (120 MHz). Stack measurements follow the `pqm4` framework's built-in stack-benchmarking infrastructure. All Cortex-M4 measurements—cycles, stack, and code size—use `-O3` (the `pqm4 -o speed` setting), so every row of Table 3 is compiled with identical flags. We use `-O3` on AVR and MSP430 as well; only the Cortex-M0+ falls back to `-Os`, as the full `-O3` image overflows its 32 kB flash. There, the cycle and stack builds use plain `-Os` and `-flto` is added only for the combined flash-size figure (Section 6). We report the median over 100 executions. All reported stack peaks are empirical measurements obtained with `arm-none-eabi-gcc` 14.2 under the `pqm4` build configuration (`-O3`, no link-time optimization). The signing passes are compiled as a separate translation unit, so without link-time optimization the compiler cannot inline them into the driver frame.

We benchmark five HAETAE configurations: **ref (C)** (latest C reference code [Kpq26]), **m4fspeed/m4fstack** (asm) (the reference `pqm4` M4 ports of [HCK+26], speed- and stack-optimized respectively), **ours (C)** (this work, pure C), and **ours (asm)** (this work with assembly NTT and sampler from [CCD+24]). Since the implementation of [HCK+26] is not publicly available, we report their published numbers directly. Table 3 presents the results.

Stack usage. Compared to the pure-C reference, **ours (C)** reduces peak signing stack by 92.0–95.8 % across all levels, bringing it to 5,836 B (HAETAE-2), 6,020 B (HAETAE-3), and 6,004 B (HAETAE-5). Verification stack ranges from 4,824 B to 4,848 B, an 86–93 % reduction. Key generation ranges from 4,776 B (HAETAE-5) to 5,816 B (HAETAE-2/3), a 75.6–90.9 % reduction vs. ref(C). **ours (asm)** uses the same stack as **ours (C)** across all operations, since the assembly only replaces the NTT and pointwise routines without changing the allocation. Note that the **m4fspeed/m4fstack** reference ports include ARM assembly optimizations and target speed and stack respectively, so a direct comparison with our implementation is not entirely fair; we include them for completeness. For HAETAE-5, the only level reported by [HCK+26], our implementation achieves lower stack across all three operations:

- Key generation: 4,776 B vs. 5,212 B (−8.4 %). Caller-level union analysis (Section 4.1) packs the FFT workspace and sampling scratch into shared memory slots.
- Signing: 6,004 B vs. 8,092 B (−26 %). The pass decomposition with separate-translation-unit pass boundaries (Section 4.2) ensures the peak is bounded by $S_{\text{driver}} + \max(S_A, S_B, S_{C_1}, S_{C_2})$, and the *Reverse-order streaming entropy coding* eliminates the full hint and HB^{z_1} staging buffers.
- Verification: 4,840 B vs. 6,220 B (−22 %). Our row-streamed design (Section 4.3) replaces the column-streamed `polyveck` accumulator with a single polynomial, combined with view-style decoding and incremental transcript hashing.

Both variants have `.data` = 0 and `.bss` = 0, whereas [HCK+26] reports a 1,040 B gap between `.text` and `.total`, whose breakdown is not disclosed.

Cycle counts. Key generation for HAETAE-5 is faster than **m4fspeed** for both **ours (C)** (0.62×) and **ours (asm)** (0.70×). Since d=0 key generation (Algorithm 2) performs the singular-value norm rejection before expanding the matrix $\mathbf{A}$, with an acceptance rate of roughly 10 %, most iterations only execute the lightweight norm check, and the

matrix–vector product runs only once after acceptance. For HAETAE-2/3 (d>0), the norm check is interleaved with the matrix–vector product, so each rejected sample also pays the streaming matrix cost, resulting in 1.17–1.66× slower key generation.

Signature generation requires 3.94–5.70× the cycles of **m4fspeed** (**ours (C)**: 4.62–5.70×; **ours (asm)**: 3.94–4.88×), reflecting the cost of seed-based recomputation in every pass. For HAETAE-5, **ours (C)** achieves 172,765 K cycles, 0.57× the cycles of [HCK+26] (300,881 K), with 26 % lower stack; *Component-level early rejection* (Section 4.2) reduces the average cost of rejected iterations by skipping unnecessary multiply-accumulate operations, contributing to this speedup. **Ours (asm)** further reduces this to 166,236 K (0.55× vs. [HCK+26]).

Against **m4fspeed**, verification with **ours (asm)** is faster at every level (0.93–0.98×), as it keeps the assembly NTT, whereas the pure-C **ours (C)** is 1.19–1.29×. For HAETAE-5, [HCK+26] is 1.81×, **ours (C)** 1.19×, and **ours (asm)** 0.95×, all at far lower verification stack than m4fspeed (37,184 B): [HCK+26] uses 6,220 B (83.3 % lower) and both ours variants 4,840 B (87.0 % lower).

Code size. **Ours (C)** shows a modest `.text` increase compared to the reference (30,150 vs. 25,658 for HAETAE-5, +17.5 %), reflecting the streaming rANS encoder, fused multiplication, and pass decomposition logic. **Ours (asm)** is larger (37,834 for HAETAE-5, +47 % vs. ref) due to the inlined NTT and CDT sampler routines. Both variants place all constant tables in read-only memory, resulting in `.data` = 0 and `.bss` = 0; the reported `.text` and `.total` are therefore identical. ML-DSA scheme-only code sizes (19 kB to 24 kB) are smaller than HAETAE's, reflecting the simpler structure of ML-DSA.

5.2 Results on RIOT-OS (nRF52840 and ESP32-C6)

RIOT [RIO] is an open-source operating system targeting resource-constrained IoT devices. Originally introduced in [BHG+13], it is written in C and designed with portability in mind, which allows it to support a broad range of boards and architectures. To validate portability beyond the `pqm4` bare-metal environment, we ran the same low-stack build under RIOT-OS on two targets: the **Nordic nRF52840** (ARM Cortex-M4F, 64 MHz, 256 kB RAM) and the **Espressif ESP32-C6** (RISC-V RV32IMAC, 80 MHz, 512 kB RAM). Stack usage is measured via a high-watermark technique: the unused stack region below the current stack pointer is painted with a canary pattern before each operation, and the deepest overwrite is recorded after the operation returns. Timing uses RIOT's `xtimer` layer converted to cycles via `coreclock_hz`. Results are averaged over 100 iterations.

Table 4: Low-stack evaluation on RIOT-OS (**this work**). nRF52840: ARM Cortex-M4F @ 64 MHz, 256 kB RAM. ESP32-C6: RISC-V RV32IMAC @ 80 MHz, 512 kB RAM. Cycles measured via `xtimer` function from RIOT-OS.

		Key generation		Signing		Verification	
	Platform	Cycles	Stack	Cycles	Stack	Cycles	Stack
HAETAE-2	nRF52840	29,039 K	5,988	262,721 K	6,096	2,023 K	3,984
	ESP32-C6	49,371 K	6,100	476,223 K	6,068	3,619 K	4,004
HAETAE-3	nRF52840	46,520 K	6,208	829,928 K	6,288	3,838 K	3,984
	ESP32-C6	80,811 K	6,580	1,519,566 K	6,260	7,146 K	4,004
HAETAE-5	nRF52840	13,612 K	5,096	193,913 K	6,406	4,970 K	3,976
	ESP32-C6	21,513 K	5,254	526,047 K	6,244	9,536 K	4,004

As Table 4 shows, peak stack figures are close between the two RIOT-OS targets. For HAETAE-2, signature generation peaks at 6,096 B (nRF52840) and 6,068 B (ESP32-C6), suggesting that the memory footprint is dominated by the algorithm and build knobs, with only small platform/ABI effects.

Cycle-count gap between Cortex-M4F and RV32IMAC. Despite a 25 % higher clock frequency, the ESP32-C6 requires substantially more cycles for every operation. The largest gap is HAETAE-5 signing (2.71× the cycles on ESP32-C6), while the other operations range from 1.58 to 1.92×; the main reason is the absence of DSP/SIMD hardware on the in-order RV32IMAC core: the Cortex-M4F benefits from a single-cycle 32-bit multiplier suited to the NTT-heavy polynomial arithmetic in HAETAE.

6 Results on Low-end Microcontrollers

Assembly arithmetic on low-end targets. On the Cortex-M0+ and AVR, the `(asm)` variant replaces the NTT and inverse NTT with Barrett-based routines, applying the shape-independent Barrett multiplication of [HKS25], which is tailored to targets without a fast long-multiply instruction; the pointwise multiplication keeps Montgomery reduction. On the Cortex-M0+ (ARMv6-M, no `umull` or DSP multiply), the NTT butterflies compute the Barrett high-word terms with `muls`-based limb operations, avoiding the per-butterfly helper call emitted by the C path; on AVR the NTT/iNTT assembly inlines these steps instead of the compiler-emitted `__mulsi3`/`__mulsidi3` helpers. The assembly does not change the high-level stack allocation.

Table 5: Verify-only results (**ours (C)/(asm)**) on **Arduino Mega 2560 R3** (ATmega2560, 8-bit AVR, 16 MHz, 256 kB Flash, 8 kB SRAM). Flash/Stack/Cycles: `-O3` with `avr-gcc 15.2.0`. Stack: peak. Cycles: median of 10 runs ($K = 10^3$). (C) = pure C; (asm) = assembly NTT. Headroom is relative to the 7,986 B of SRAM available to the signature and stack after the measurement harness's 206 B of static data.

	Impl.	Flash	Stack*	Headroom	Cycles	ms
HAETAE-2	(C)	45,694	6,458	1,528 B	21,232 K	1,327
	(asm)	48,822	6,458	1,528 B	17,424 K	1,089
HAETAE-3	(C)	48,154	7,355	631 B	42,946 K	2,684
	(asm)	51,282	7,355	631 B	34,462 K	2,154
HAETAE-5	(C)	48,554	7,961	25 B	53,615 K	3,351
	(asm)	51,930	7,956	30 B	44,394 K	2,775

∗: $|sig|$ + verify stack. $|pk|$ in flash memory

On resource-constrained devices, the public key is stored in flash and the signature is received over a communication channel into SRAM [BZB+22, BDKR23], so the peak SRAM demand is $|sig|$ + verification stack. To validate the 8 kB feasibility on actual 8-bit hardware, we ported the HAETAE verification (**ours (C)**) to an **Arduino Mega 2560 R3** board, which features the ATmega2560 8-bit AVR microcontroller running at 16 MHz with 256 kB flash and 8 kB SRAM. The firmware was compiled with `avr-gcc 15.2.0` (`-O3`) and flashed via `avrdude 6.3` on Ubuntu 22.04. All constant tables and the public key reside in flash via `PROGMEM`, achieving `.data` = `.bss` = 0 for the HAETAE code; after the measurement harness's 206 B of static data, 7,986 B of SRAM remain for the signature and stack. As shown in Table 5, all three security levels complete verification within 8 kB, with HAETAE-5 leaving only 25 B of headroom. The assembly variant runs verification at 0.80–0.83× the cycles of pure C without increasing the stack. The concurrent HAETAE implementation [HCK+26] requires 9,168 B ($|sig|$=2,948 B + stack=6,220 B) for HAETAE-5 verification, exceeding the 8 kB boundary by over 1 kB and making such a demonstration infeasible. While [HKS25] (CHES'25) presents an ML-DSA implementation targeting 8-bit AVR, their results are obtained through the `IAR AVR simulator` with a memory budget exceeding 16 kB, as stated in their artifact, rather than on physical hardware with 8 kB SRAM.

Table 6: **ours (C)/(asm)** on **NUCLEO-C031C6** (STM32C031C6, 32-bit Cortex-M0+, 48 MHz, 32 kB Flash, 12 kB SRAM). Flash: `-Os -flto`. Stack: `-Os`, peak. Cycles: `-Os`, median of 100 runs ($K = 10^3$). (C) = pure C; (asm) = assembly NTT/pointwise.

			Key generation			Signing			Verification		
	Impl.	Flash	Stack†	Cycles	ms	Stack‡	Cycles	ms	Stack§	Cycles	ms
HAETAE-2	(C)	27,840	8,512	66,191 K	1,379	9,064	318,888 K	6,643	7,560	5,034 K	105
	(asm)	29,176	8,512	53,093 K	1,106	9,056	387,619 K	8,075	7,560	3,707 K	77
HAETAE-3	(C)	28,288	9,696	102,646 K	2,138	10,832	851,868 K	17,747	8,928	10,366 K	216
	(asm)	29,624	9,696	85,287 K	1,777	10,824	835,638 K	17,409	8,928	7,444 K	155
HAETAE-5	(C)	27,320	9,912	74,318 K	1,548	12,056	728,614 K	15,179	10,128	12,489 K	260
	(asm)	28,560	9,912	65,700 K	1,369	12,048	374,542 K	7,803	10,128	9,336 K	195

†: $|sk| + |pk|$ + keygen stack. ‡: $|sk| + |sig|$ + sign stack. §: $|pk| + |sig|$ + verify stack.

Table 7: **ours (C)** on **MSP-TS430PEU128** (MSP430F67791, 16-bit MSP430, 8 MHz, 512 kB Flash, 32 kB SRAM). Flash/Stack/Cycles: `-O3` with `msp430-elf-gcc 9.3.1`, `-ffunction-sections -fdata-sections -gc-sections`; median of 10 runs ($K = 10^3$). Stack is flag-insensitive (`-O3` = `-Os`). Stack notation follows Table 6.

		Key generation			Signing			Verification		
	Flash	Stack†	Cycles	ms	Stack‡	Cycles	ms	Stack§	Cycles	ms
HAETAE-2	78,438	7,744	110,733 K	13,842	8,960	1,920,657 K	240,082	7,558	33,189 K	4,149
HAETAE-3	76,978	8,928	461,501 K	57,688	10,726	2,969,350 K	371,169	8,928	67,031 K	8,379
HAETAE-5	77,230	9,132	217,367 K	27,171	11,934	4,619,759 K	577,470	10,126	85,899 K	10,737

Full stream on 12 kB devices. We ported our pure-C implementation (**ours (C)**) to a **NUCLEO-C031C6** board (STM32C031C6, Cortex-M0+, 32 kB Flash, 12 kB SRAM, 48 MHz). The firmware was compiled with `arm-none-eabi-gcc` 14.2 (`-Os`; cycles measured at `-Os` as the full `-O3` image overflows the 32 kB Flash) and flashed via `st-flash` 1.8.0 on Ubuntu 22.04; all constant tables reside in Flash (`.data` = `.bss` = 0). As Table 6 shows, all three HAETAE processes complete within 12 kB at all security levels. HAETAE-5 signing is the tightest scenario (12,056 B peak, 232 B headroom). The combined code size (key generation, signing, verification) exceeds 32 kB of Flash under `-Os` alone (41,723 B for HAETAE-3); enabling link-time optimization (`-flto`) reduces it to 28,288 B (86 % of 32 kB). Verification completes in 105 ms to 260 ms across all levels (assembly: 77 ms to 195 ms). For comparison, ML-DSA-65 (m4fstack, `pqm4`) requires $|sk|+|sig|$+stack = 4,032+3,309+6,608 = 13,949 B for signing (Table 3), exceeding 12 kB by nearly 2 kB; ML-DSA-87 requires 17,667 B. The state-of-the-art work [HCK+26] also exceeds this bound for HAETAE-5 signing (2,752+2,948+8,092 = 13,792 B). We note that the recent ML-DSA benchmark on Cortex-M0+ [Chh26] targets the RP2040 (Raspberry Pi Pico H), which has 264 kB SRAM, 2 MB QSPI Flash, and a single-cycle 32×32 multiplier at 133 MHz; they report 50.6 kB of signing stack for ML-DSA-44. As the RP2040 is a mid-range device with substantially more resources, a direct comparison is outside the scope of this work. Finally, Table 7 reports HAETAE performance on a 16-bit **MSP-TS430PEU128** board, completing our cross-architecture evaluation across 8-bit (AVR), 16-bit (MSP430), and 32-bit (Cortex-M0+, Cortex-M4, and RISC-V) physical hardware.

7 Conclusion

We presented a low-stack implementation of HAETAE covering all three security levels. It reduces the signing stack from 71 kB to 141 kB in the reference to 5.7 kB to 5.9 kB. Verification therefore runs within 8 kB at every level, which we showed on an 8-bit AVR, and the full signing path runs within 12 kB, which we showed on a 32-bit Cortex-M0+. These results come from four techniques: rejection-aware pass decomposition, component-

level early rejection, reverse-order streaming entropy coding, and row-streamed verification, each of which trades extra recomputation for a lower peak. We designed the techniques for HAETAE, but they also apply to other Fiat–Shamir-with-aborts lattice signatures, which share a reject-and-retry signing loop, seed-expandable matrix products, and packable per-row signature data. Pass decomposition and reverse-order streaming carry over to the hint and bit-packing of ML-DSA, and row-streamed verification applies whenever an **Az** product can be generated and consumed one row at a time. This trade-off favors memory-bound devices and helps little on compute-bound ones. Our implementation is constant-time except for the rejection loop that the paradigm requires, and it does not address power or fault attacks. For future work, we plan to analyze the error of the sampler's fixed-point reciprocal square root, to instantiate the techniques for ML-DSA, and to add masking and fault countermeasures.

Acknowledgements

We thank the editor, the shepherd, and the anonymous reviewers for their helpful comments during the review process. This work was partly supported by the HYPERFORM consortium, funded by France through Bpifrance, and by the France 2030 program under grant agreement ANR-22-PETQ-0008 PQ-TLS; by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2026-25519543, Development of PQC migration technology for Quantum-Safe IC chip, 50%); and by the National Research Foundation of Korea (NRF) grant funded by MSIT (No. 2022R1C1C1013368, 50%).

References

[BBC+23] Joppe W Bos, Olivier Bronchain, Frank Custers, Joost Renes, Denise Verbakel, and Christine van Vredendaal. Enabling FrodoKEM on embedded devices. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2023(3):74–96, 2023.

[BDKR23] Joppe W. Bos, Alexander Dima, Alexander Kiening, and Joost Renes. Post-quantum secure over-the-air update of automotive systems. Cryptology ePrint Archive, Paper 2023/965, 2023.

[BHG+13] Emmanuel Baccelli, Oliver Hahm, Mesut Günes, Matthias Wählisch, and Thomas C. Schmidt. RIOT OS: Towards an OS for the Internet of Things. In *2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 79–80, 2013.

[BRS22] Joppe W. Bos, Joost Renes, and Amber Sprenkels. Dilithium for memory constrained devices. In Lejla Batina and Joan Daemen, editors, *Progress in Cryptology - AFRICACRYPT 2022: 13th International Conference on Cryptology in Africa, AFRICACRYPT 2022, Fes, Morocco, July 18-20, 2022, Proceedings*, volume 13503 of *Lecture Notes in Computer Science*, pages 217–235. Springer Nature Switzerland, 2022.

[BZB+22] Gustavo Banegas, Koen Zandberg, Emmanuel Baccelli, Adrian Herrmann, and Benjamin Smith. Quantum-resistant software update security on low-power networked embedded devices. In Giuseppe Ateniese and Daniele Venturi, editors, *Applied Cryptography and Network Security - 20th International Conference, ACNS 2022, Rome, Italy, June 20-23, 2022, Proceedings*, volume 13269 of *Lecture Notes in Computer Science*, pages 872–891. Springer, 2022.

[CCD+23] Jung Hee Cheon, Hyeongmin Choe, Julien Devevey, Tim Güneysu, Dongyeon Hong, Markus Krausz, Georg Land, Junbum Shin, Damien Stehlé, and MinJune Yi. HAETAE. Technical report, National Institute of Standards and Technology, 2023. available at https://csrc.nist.gov/Projects/pqc-dig-sig/round-1-additional-signatures.

[CCD+24] Jung Hee Cheon, Hyeongmin Choe, Julien Devevey, Tim Güneysu, Dongyeon Hong, Markus Krausz, Georg Land, Marc Möller, Damien Stehlé, and MinJune Yi. HAETAE: Shorter lattice-based Fiat-Shamir signatures. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2024(3):25–75, 2024.

[Chh26] Rojin Chhetri. Benchmarking nist-standardised ml-kem and ml-dsa on arm cortex-m0+: Performance, memory, and energy on the rp2040, 2026.

[HCK+26] Yulim Hyoung, Subeen Cho, Uijae Kim, Minwoo Lee, Hwajeong Seo, and Minjoo Sim. Memory-efficient implementation of SMAUG-T and HAETAE. Cryptology ePrint Archive, Paper 2026/442, 2026.

[HKS25] Vincent Hwang, YoungBeom Kim, and Seog Chung Seo. Multiplying polynomials without powerful multiplication instructions. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2025(1):160–202, 2025.

[Kor] Korean Post-Quantum Cryptography Research Group. KpqC Algorithms final specification documents. Accessed: 2026-03-30.

[Kpq26] Korean Post-Quantum Cryptography Standardization Committee. *HAETAE: Shorter Lattice-Based Fiat-Shamir Signatures*, 2026. Final specification. available at https://www.kpqc.or.kr/images/pdf2/HAETAE.pdf.

[KPR+] Matthias J. Kannwischer, Richard Petri, Joost Rijneveld, Peter Schwabe, and Ko Stoffelen. pqm4: Post-quantum crypto library for the ARM Cortex-M4. https://github.com/mupq/pqm4.

[LWKP25] Seungwoo Lee, Joo Woo, Jonghyun Kim, and Jong Hwan Park. Generalized centered binomial distribution for bimodal lattice signatures. *IEEE Access*, 13:2203–2214, 2025.

[Lyu09] Vadim Lyubashevsky. Fiat-Shamir with aborts: Applications to lattice and factoring-based signatures. In Mitsuru Matsui, editor, *Advances in Cryptology - ASIACRYPT 2009, 15th International Conference on the Theory and Application of Cryptology and Information Security, Tokyo, Japan, December 6-10, 2009. Proceedings*, volume 5912 of *Lecture Notes in Computer Science*, pages 598–616. Springer, 2009.

[NIS24] National Institute of Standards and Technology. *Module-Lattice-Based Digital Signature Standard (ML-DSA)*, 2024. Federal Information Processing Standards Publication 204, https://doi.org/10.6028/NIST.FIPS.204.

[RIO] RIOT Community. Riot – the friendly operating system for the internet of things. Accessed: March 31, 2026.