高能效格基后量子密码并行采样算法与硬件结构研究

别梦妮,李 伟*,付秋兴,陈 韬,杜怡然,南龙梅
(信息工程大学,河南郑州 450001)

摘 要: 在后量子密码高速演进的过程中,为兼顾灵活性与高效性的需求,本文面向多种格基后量子密码算法提出了一款并行可重构的采样加速器. 本文结合数学推导分别提出了7种采样的高效并行实现模型,并从中提炼了4种共同运算逻辑. 以这4种共同运算逻辑为核心,引入数据重排限制运算数据的有效位宽,提高了拒绝采样的接受率并简化了运算逻辑,提出了一种高能效的可重构并行采样算法. 为提升采样算法的硬件实现效能,本文采用蝴蝶变换网络在单个时钟周期内完成任意有效位宽数据的并行切分、归并与查找,高效实现了算法前后处理的并行化,构建了参数化的并行可重构采样加速器架构模型,结合实验探索,提出了一款数据带宽为1 024 bit的并行可重构采样加速器. 实验结果表明,使用40 nm CMOS工艺库,在ss、125 ℃工艺角条件下进行后仿,电路最高工作频率可达到667 MHz,平均功耗为0.54 W. 完成256点均匀采样需6 ns,完成256点拒绝值小于2^{16}的拒绝采样平均仅需22.5 ns,完成256点8 bit以内的二项采样需18 ns,完成509点简单三值采样需36 ns,完成701点非负相关三值采样需124.5 ns,完成509点固定权重三值采样需11.18 μs,完成一次Falcon算法中的离散高斯采样需3 ns. 与现有研究相比,本文提出的采样器完成一次均匀-拒绝采样的能耗值降低了约30.23%,完成一次二项采样的能耗值降低了约31.6%.

关键词: 后量子密码算法;格;采样器;高能效;可重构
基金项目: 河南省自然科学基金(No.232300421393)
中图分类号: TN402;TP309 **文献标识码:** A **文章编号:** 0372-2112(2025)02-0420-11
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20241036

Research on Energy-Efficient Parallel Sampling Algorithm and Hardware Architecture for Lattice-Based Post-Quantum Ciphers

BIE Meng-ni, LI Wei*, FU Qiu-xing, CHEN Tao, DU Yi-ran, NAN Long-mei
(*University of Information Engineering*, *Zhengzhou*, *Henan* 450001, *China*)

Abstract: During the rapid evolution of post-quantum cryptography, considering the needs for flexibility and efficiency, we proposed a parallel reconfigurable sampling accelerator for various lattice-based post-quantum cryptographic algorithms. We analyzed seven sampling processes involved in lattice-based post-quantum cryptography and proposed seven efficient parallel implementation models for these samplings, respectively, based on mathematical derivations. Then we extracted four common operational logics from these models. Using these four common operational logics as the core, we introduced data rearrangement to limit the effective bit width of operation data, which improved the acceptance rate of rejection sampling and eliminates the complex modular reduction operations in finite field operations. Then we proposed a high energy-efficient reconfigurable parallel sampling algorithm. To enhance the hardware implementation efficiency of the sampling algorithm, we adopted the butterfly transform network to complete the parallel splitting, merging, and lookup of data with any effective bit width within a single clock cycle, efficiently realizing the parallelization of the algorithm's pre- and post-processing, and constructed a parameterized parallel reconfigurable sampling accelerator architecture model. Aiming for high energy efficiency, combined with logic synthesis experimental results, we determined the optimal parallel degree parameters of the architecture model and proposed a parallel reconfigurable sampling accelerator with a data bandwidth of 1 024 bits. Experimental results showed that, using a 40 nm CMOS process library, and performing post-simulation under

收稿日期:2024-11-15;修回日期:2025-02-19;责任编辑:朱梅玉
*通讯作者:李伟

the ss, 125 ℃ process corner conditions, the circuit´s highest operating frequency can reach 667 MHz, with an average power consumption of 0.54W. Completing a 256-point uniform sampling requires 6 ns, completing a 256-point rejection sampling with a rejection value less than 2^{16} on average only takes 22.5 ns, completing a 256-point binary sampling within 8 bits requires 18 ns, completing a 509-point simple ternary sampling requires 36 ns, completing a 701-point non-negative correlated ternary sampling requires 124.5 ns, completing a 509-point fixed-weight ternary sampling requires 11.18μs, and completing a discrete Gaussian sampling in the Falcon algorithm once requires 3 ns. Compared with existing research, the sampler proposed in we reduce the energy consumption value for a uniform-rejection sampling by about 30.23%, and the energy consumption value for a binary sampling by about 31.6%.

Key words: post-quantum cryptographic algorithms; lattice; sampler; energy efficient; reconfigurable

Foundation Item(s): Natural Science Foundation of Henan Province (No.232300421393)

1 引言

随着量子计算技术的突破和NIST(National Institute of Standards and Technology)后量子加密/签名算法的标准化,实现后量子密码的应用和部署日渐紧迫,面向后量子密码的高效硬件实现已成为当前研究的热点. 后量子密码算法发展至今,基于格上困难问题的算法逐步展现出优势,是后量子密码发展和应用的主流选择[1],而采样过程在任意一种格基后量子密码算法中都扮演着重要角色. 当前各国家和组织中,仅NIST确定了加密/签名标准,尚未确定的密钥封装标准,且中国等国家和组织的后量子密码标准均未确立,现阶段构建的后量子密码系统需要保留一定的灵活更新机制[2]. 同时,后量子密码算法在小型移动设备上的实时通讯应用要求芯片具有很高的能量效率. 因此,兼顾格基后量子密码芯片的灵活性与高效性需求,研究一种适用于多算法的高效采样器具有重要意义.

格密码算法的安全性极大程度上依赖于采样器提供的不确定性,在早期提出的格基后量子密码方案中,离散高斯采样占据主流地位,NIST第一轮提交的算法版本基本都采用了离散高斯采样. 因此,早期的采样器设计均针对离散高斯采样进行. 随着数学理论上的深入研究和证明,算法设计者逐步以杂凑函数配合简单采样的方式取代了复杂的离散高斯采样方案,从算法源头上提高了格密码的性能表现,与此同时也增加了采样过程的多样性. 当前,针对格基后量子密码的采样器研究可以分为三类. 第一类是针对单算法的硬件实现研究. 由于传统密码学中不存在采样类操作,关于均匀、拒绝[3]、二项[4]和三元采样[5]的硬件实现研究多数伴随着后量子密码学的发展而发展,在当前的文献中针对这几类采样的硬件实现研究较少. 与这几种简单采样过程不同,离散高斯采样因其在通信等领域的广泛应用及其本身运算逻辑的复杂性得到了相对充分的研究,被广泛采用的方案主要集中在Knuth-Yao采样[6]和积累分布表采样[7]. 这一类研究针对单一功能采样器进行了特定的优化,取得了较好的能效,但灵活性较低,不适合面向多算法的芯片设计. 第二类文献是面向后量子密码处理器的整体实现方案,例如文献[8,9],这一类文献利用通用处理器架构的普适性,也实现了多种采样功能,但并未设计相应的单元级加速器,其灵活性更强,但性能较低,采样速度较第一类加速器方案至少低一个数量级. 第三类研究是面向多种后量子密码算法设计的采样加速器. 这一类文献采用的是前两者的折中方案,文献[10,11]实现了格基后量子密码中所有类型的采样,但仅仅是将这些采样加速器进行集成,致使电路主频低,面积大,效率低下;文献[12,13]是针对某一类算法实现进行的整体实现,以提高并行度、固定参数优化等方式实现了较高的效率,但支持的采样算法种类仍然较少,灵活性介于前两者之间,依然无法满足后量子密码多算法需求.

综上所述,当前针对格基后量子密码采样器设计的研究在灵活性与高效性之间存在严重制约关系. 针对可重构采样器设计进行的研究或仅对单一功能采样器进行粗略的集成,或直接利用通用环境满足灵活性需求,均忽视了采样的综合能效表现. 鉴于此,本文在保证足够灵活性的前提下,以取得与单采样器设计近似的能效表现为核心目标,面向可重构的格基后量子密码芯片展开研究,解决了多算法芯片的效率瓶颈问题.

本文分别针对7种采样过程,从数学理论的角度对其在并行实现条件下的性能进行优化,从而提出其各自的并行实现模型. 从模型中提炼其共同运算逻辑,并提出一个统一的并行度可调节的高效可重构并行采样算法. 构建了一个参数化的可重构并行采样加速器模型,并结合实验分析确定了一款高能效的可重构后量子密码采样加速器. 与以往研究相较,本文从算法层次上实现了对7种采样的高效统一,并面向并行实现环境,在单元级实现了7种采样类型多参数的高效重构.

本文的贡献总结如下:

(1)结合数学推导分别提出了7种采样的高效并行实现模型,并提炼了各类采样过程的共同运算逻辑.

(2)提出了一个高效可重构并行采样融合算法. 该

算法引入了数据重排限制运算数据的有效位宽,提高了拒绝采样的接受率并简化了运算逻辑,为多参数、多类型采样提供了一个高效并行化范式.

(3)采用蝴蝶变换网络高效实现了算法前后处理的并行化,构建了参数化的并行可重构采样加速器架构模型.结合实验探索,提出了一款数据带宽为1 024 bit的并行可重构采样加速器.

2 后量子密码算法中的采样过程研究与分析

不同的后量子密码算法设计者根据需求,采用了不同的采样过程与一些附加操作来生成所需的多项式系数.本文选取当前已提出的数个具有代表性的后量子密码方案,对其中的采样过程进行统计分析.如表1所示,当前后量子密码算法中涉及的采样共7类,分别为均匀采样、拒绝采样、二项采样、简单三值采样、非负相关三值采样、固定权重三值采样以及离散高斯采样.表1中附加操作为各算法中独立于常规采样过程的额外操作.对于离散高斯采样,生成采样点后,需要通过查找表取得实际采样值,这个查找表的内容根据后量子密码算法不同取值不同,最简单的情况下甚至无需进行这一步骤,故而不纳入常规离散高斯采样的范畴,归为附加操作.

表1 后量子密码算法中采样过程分析

算法	采样类型							附加操作
	均匀*	拒绝*	二项*	三值			离散高斯	
				简单	非负相关	固定权重	(σ/Depth/λ)	
Dilithium	64、8、24、18、20	8、23、18、20						τ次(τ<64)条件选择取补
kyber	12、4、6	12	2、3					
Newhope	16	16	8					n个16 bit数据反序
NTRU	8、30			模3(sign)	三值(–1、0、1)乘累加,取符号位与偶数项相乘	左移两位加1(或2、0)		n–2个32 bit数据排序
Saber	13、6、8、10		3、4、5					
LAC	2		1					
FrodoKEM							2.8/12/15 2.3/10/16 1.4/6/16	查找表
SEAL							3.19/42/128 3.19/52/192 3.19/60/256	查找表
qTESLA							8.5/80/64 8.5/112/125	查找表
LP							3.33/35/80 3.33/37/90 3.33/39/100	查找表
FALCON							2/18/53 $\sqrt{5}$/37/200	查找表

注:均匀、拒绝、二项采样统计的参数均为参与该过程的有效数据位宽.

基于表1的统计结果,可重构采样加速器共需要实现7种采样过程,下面本文分别针对这7种采样展开研究.提高算法的可并行化程度是实现高能效的一种常规思路,本文从各采样算法的原理出发,结合数学分析推导,提出了7个高效并行实现模型.其中根据各类采样的特性采用了限制数据有效位宽、比特串按位拆分、运算逻辑转换等方式提高了算法的并行化程度,且在不同程度上提高了各算法的执行能效.各算法的优化分析过程描述如下.

(1)均匀采样

均匀采样是最基础的采样模型,其实质是对一串比特流按照一定位长进行数据切分,在参数可配置的并行处理结构中,该过程不仅仅为硬连线逻辑,而是对多比特的分配,本文称之为数据切片.

(2)拒绝采样

拒绝采样的实质是比较与选择,假设拒绝值为q,输入的随机数为$\beta=\{\beta_0,\beta_1,\cdots,\beta_{n-1}\}$,则拒绝采样可表述为式(1):

$$f=\begin{cases}\beta, & \beta<q\\ \text{拒绝}, & \beta\geqslant q\end{cases} \tag{1}$$

式(1)的拒绝概率为$(\beta-q)/\beta$,其值与输入随机数β的大小息息相关,当β尽可能接近q时,拒绝概率低. 若设计参数固定化的拒绝采样器,通常以硬连线的方式限定β的位宽n满足$n=\lceil\log_2 q\rceil+1$. 此方法应用在参数可配置的拒绝采样器中,需要对输入的随机数进行数据切片,因此本文提出的拒绝采样器实质上可以称作均匀-拒绝采样器.

在并行处理框架下,每轮采样可生成一组多个采样值,但在拒绝采样过程中并非每个数值均为有效数值,如若仅接受每个采样值均有效的采样组数据,则随着并行度的提高拒绝概率也随之增大,采样性能随之降低. 为提高并行条件下的采样速度,本文以标志位标识有效元素,同时增加数据归并后处理过程,以生成规整的采样值,便于后续运算.

(3)二项与简单三值采样

二项与三值采样从操作解构上看极其相似,对于二项采样,取长度为n比特随机串$\{\beta_0,\beta_1,\cdots,\beta_{n-1}\}$,其中$n=\text{even}$,则二项采样值可以表述为式(2):

$$f=\sum_{j=0}^{n/2-1}\beta_j-\sum_{j=n/2}^{n-1}\beta_j \tag{2}$$

取长度为n比特随机串$\{\beta_0,\beta_1,\cdots,\beta_{n-1}\}$,简单三值采样值可以表述为式(3):

$$f=\sum_{j=0}^{n-1}2^j\cdot\beta_j \bmod 3 \tag{3}$$

简单三值采样的实质是对一个二进制数执行取模操作,若此处以普适性的取模算法实现需要引入昂贵的乘法操作,使得采样效率降低. 本文研究发现,在模数为3的情况下,该操作可以简化为与二项采样类似的比特级运算,描述如式(4)所示:

$$\begin{aligned}f&=\sum_{j=0}^{n-1}2^j\cdot\beta_j \bmod 3\\&=\begin{cases}\sum_{j=0}^{n/2-1}\left(\beta_{2j}+2\cdot\beta_{2j+1}\right) \bmod 3, & n=\text{even}\\ \sum_{j=0}^{(n-1)/2}\left(\beta_{2j}+2\cdot\beta_{2j+1}\right) \bmod 3, & n=\text{odd}\end{cases}\end{aligned} \tag{4}$$

观察可知,二项采样与简单三值采样具有相似的运算逻辑,均需要将输入随机比特按一定的规则拆分后求汉明权重,其中二项采样根据长度折半划分,简单三值采样按奇偶进行划分;均需要对两部分汉明权重进行算术运算,其中,二项采样为求差值,简单三值采样求移位后加.

(4)非负相关三值采样

非负相关三值采样是在完成简单三值采样的基础上额外增加三值乘累加操作(只保留符号位)以及偶数项符号乘,其中,三值乘累加操作由三值乘法和一级加法组成. 由于三值乘法的乘数只有有限种情况,可以简化为布尔函数表达式,如式(5)所示:

$$\begin{aligned}(s_1,s_0)&=(a_1,a_0)\times(b_1,b_0)\\&=\big(\big(\overline{a_1}a_0b_1\overline{b_0}+a_1\overline{a_0}\,\overline{b_1}b_0\big),\\&\quad\big(\overline{a_1}a_0\overline{b_1}b_0+a_1\overline{a_0}b_1\overline{b_0}\big)\big)\end{aligned} \tag{5}$$

简化后,三值乘累加中不再出现复杂的乘法. 三值乘法后缀的累加过程亦可归为并行算术运算.

(5)固定权重三值采样

固定权重三值采样是将采集数据左移2位并加上一个2 bit以内的权重调节数值,该过程显然是一种并行的算术运算. 在完成所有采样值后,进行固定权重三值采样特有的排序操作即可.

(6)离散高斯采样

离散高斯采样具有多种比较成熟的实现方式,其中基于CDT的采样实现效率高且易于扩展. 采用CDT方法进行采样时,只需要将随机数与CDT表中数据进行比较,以确定所在区间即可. 其核心操作可以解构为比较(即带借位减法)与统计1,所生成的即为采样点. 在并行处理模型中,与三值乘累加类似,比较操作可以实现完全并行,但统计1步骤在采样规模足够大时可以并行加法的形式实现部分并行,最终的各通道之间的累加操作则必须顺序执行.

如图1所示为本文提出的7种采样的高效实现模型图. 观察图1可以发现,这种采样在并行实现时具有极其类似的操作步骤. 例如拒绝采样算法实际上是数据拆分、比较选择与数据归集操作组成;二项采样算法也是由数据拆分、计算汉明权重以及减法三部分操作组成. 就以上2个子算法为例,他们共同包含将输入随机比特串按一定的位长进行切片的操作,进一步观察发现,这种数据切片存在于每一个采样算法中. 与之类似,计算汉明权重操作同时存在于3个算法中,并行比较同时应用于2个算法,移位加减同时应用于5个算法中.

综上所述,格基后量子密码中的采样均以数据切片、计算汉明权重、并行比较和移位加减这4类操作为核心,这4类操作可覆盖每种采样类型中60%以上的步骤. 故而,与独立实现这7种采样相比,复用这4种共同逻辑设

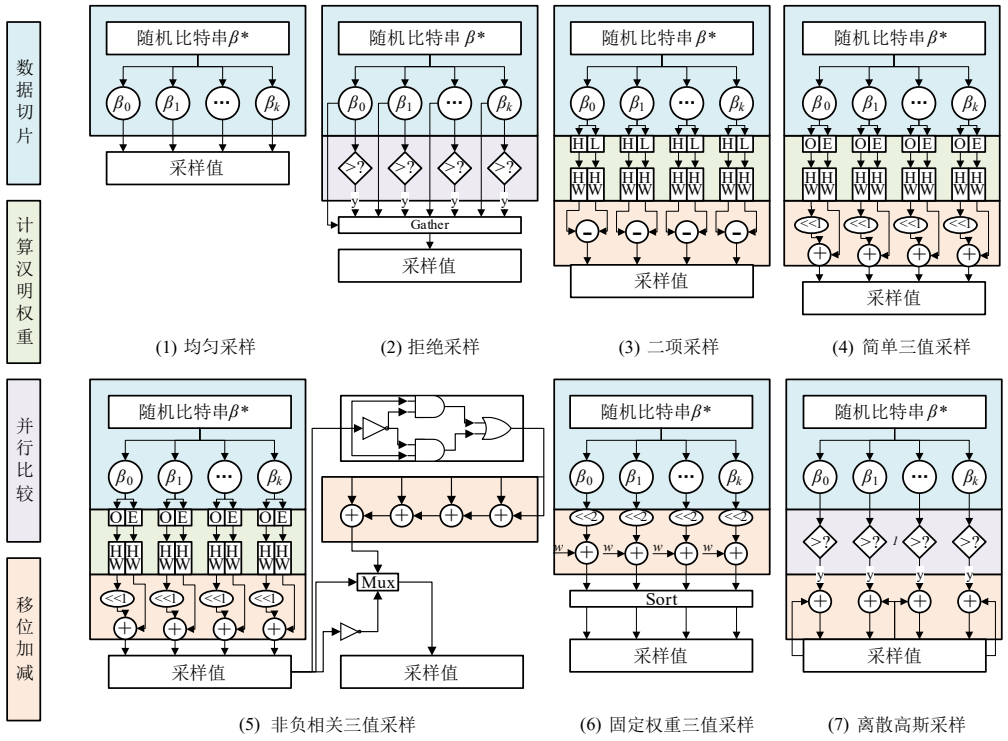


图1 7种采样的高效实现模型

计的采样加速器的硬件资源利用率及能效均更高.

3 高效可重构并行采样融合算法研究与设计

以第2节抽取的4个共同逻辑为核心,本文提出并行度可配的高效可重构并行采样融合算法,如算法1所示. 算法1中,取定长随机数(Stage1)和数据分配(Stage2. 1)这2个过程实现了4个核心操作中的数据切片功能. 在大多数常规重构设计方案中,该过程通常被省略,替代为按照2的幂次进行冗余设计. 相较而言,本文的方案兼容了均匀采样的不同参数情况,减少了随机比特消耗量,同时限制了输入数据位宽,简化了有限域上的运算过程,提高了采样速度.

4个核心操作中的计算汉明权重和并行比较2个操作均由算法1的中间计算阶段(Stage2.2)实现,此部分的核心功能即预先计算并保留中间结果,可减少重复计算的时间和相应的资源消耗,从而提高可重构算法的执行能效.

移位加减核心操作主要在生成各类采样值时使用,实际上是对中间计算结果的整合加工,对应于算法1的采样值生成阶段(Stage2.3). 由于此阶段执行的操作均遵循先移位后加减的范式,故而对应在硬件结构中只是一套可选择输入数据的加减法器.

除4类核心操作外,还需要循环控制机制以实现多点采样,此过程即算法1中的统计判断采样是否完成阶段(Stage2.4). 该阶段对于离散采样,考虑到概率分布的样本数量可能超过并行度,需要记录已经进行过的比较与统计;对于拒绝采样,为提高拒绝采样的接受率,本文直接记录接受的数据,而不是整体拒绝,需要记录每轮接受的采样点数;对于余下其他类型的采样,则只需要累加并行度即可.

上述步骤仅可以实现多点常规的采样过程,由于格基算法中的采样还融合了表1中体现的各类附加操作,所以算法1也对应设计了后处理阶段(Stage3)来实现完整的格基后量子密码算法中的采样. 该阶段主要用于三值采样与离散高斯采样,涉及所有采样数据之间的运算,故而只能在并行采样过程完成后顺序进行. 但在硬件实现过程中,诸如乘累加、排序比较等相互之间无直接数据冲突的步骤,也可并行进行.

综上所述,算法1以4种共同运算逻辑为核心,采用数据重排、拒绝标记等方式实现了7种采样过程的高效融合统一,且以参数化的方式适配了不同的并行度环境.

4 高效的并行可重构采样加速器

4.1 并行可重构采样加速器架构

并行处理架构每轮次可完成多个数据的加载/存储或运算操作,大大减少了完成算法所需的时钟周期数,是提高性能的有效选择. 同时,在并行处理架构中,数据通路通常具有高带宽,可以同时读写多个数据,与后量子密码应用需求高度吻合,也契合算法1的高效实现. 故而,本文基于算法1,提出了一种高效的可重构采样加速器架构,如图2所示.

算法 1　高效可重构并行采样融合算法

输入: 无限随机比特串 β^*;样本量/采样数量 num (离散采样为样本数量,否则为采样个数);并行度 k;有效位宽 w;模值 q;加权样本 imm;CDT 表 ram[num − 1:0];模式选择控制 mode[2:0](0:二项;1:简单三值;2:非负相关三值;3:固定权重三值采样;4:均匀;5:拒绝;6:离散);

输出: 采样值 f;

```
cnt = 0; f = 0;
WHILE cnt < num DO
    ////////Stage 1 取定长随机数/////////
    IF mode = 0 OR mode = 1 OR mode = 2 THEN  β^{2kw} ← β*; //取 2kw bit 数据
    ELSE  β^{kw} ← β*; //取 kw bit 数据
    ///////Stage 2 按通道运算,此处可并行//////////
    S = 0; j = 0;
    FOR i = 0 TO k − 1
        ///////Stage 2.1 数据分配//////////
        IF mode = 0 THEN  a[i] = {β_{2iw}, β_{2iw+1}, ⋯, β_{2iw+w−1}}; b[i] = {β_{2iw+w}, β_{2iw+w+1}, ⋯, β_{2iw+2w−1}};
        ELSE IF mode = 1 OR mode = 2 THEN  a[i] = {β_{2iw}, β_{2iw+2}, ⋯, β_{2iw+2w−2}}; b[i] = {β_{2iw+1}, β_{2iw+3}, ⋯, β_{2iw+2w−1}};
        ELSE  a[i] = {β_{iw}, β_{iw+1}, ⋯, β_{iw+w−1}}; b[i] = mode[2]?(mode[1]?ram[cnt]:q):imm;
        ///////Stage 2.2 中间计算//////////
        Ma[i] = (a[i] > b[i])?1:0; S = S + Ma[i];
        Ha[i] = HW(a[i]); Hb[i] = HW(b[i]);
        ///////Stage 2.3 采样值生成//////////
        IF mode = 0 THEN  f[cnt + i] = Ha[i] − Hb[i];
        ELSE IF mode = 1 OR mode = 2 THEN  f[cnt + i] = Ha[i] + (Hb[i] << 1);
        ELSE IF mode = 3 THEN  f[cnt + i] = (a[i] << 2) + b[i];
        ELSE IF mode = 4 THEN  f[cnt + i] = a[i];
        ELSE IF mode = 5 THEN
            IF M[i] = 0 THEN  f[cnt + j] = a[i]; j = j + 1;
        ELSE IF mode = 6 THEN f[i] = f[i] + Ma[i];
    ///////Stage 2.4 统计判断采样是否完成//////////
    cnt = cnt + mode[2]?(mode[1]?1:(mode[0]?S:k)):k;
///////Stage 3 后处理//////////
IF mode = 2 THEN  sum = Σ_{i=0}^{num−2} (f[i] × f[i+1] mod± 3)
    IF sum < 0 THEN
     FOR i = 0 TO (k − 1)/2
       f[2i] = −f[2i];
ELSE IF mode = 3 THEN  sort(f[0]~f[num − 1]) mod 4;
ELSE IF mode = 6 THEN  f[i] = LUT(f[i]);
```

图 2 所示的结构可以与算法 1 的执行过程一一对应,从图中看,虚线框中部分为从输入的比特流中截取需要的部分,故而在本节中将之统称为前处理模块;Stage2.2 与 Stage2.3 均是采样的核心运算过程,故统称为核心处理模块;Stage3 包含了拒绝采样的有效数据归并、非负相关三值采样的符号处理、固定权重三值采样的数据重排与离散高斯采样的查找表操作,统称为后处理模块. 3 个模块中,前、后处理模块在多参数条件下较为复杂,本文分别在 4.2 节和 4.3 节展开.

图 2 中的核心处理模块由并行比较、求解汉明权重和移位加减三部分组成. 并行比较即 Stage2.2 的减法器,产生 1 bit 的比较结果输出. HW 模块主要执行求解汉明权重的任务,由 3 组并行加法链组成,分别求解 Ha[i]、Hb[i]和 Ma[i],由于这组加法链仅需要统计 1 的个数,加法器的位宽较小,电路结构简单,故而此处按最大参数情况设计. 移位加减即 Stage2.3 部分,考虑到该过程的移位只有 2 种情况,左移 1 位或左移 2 位,故该部分电路以硬连线配合数选逻辑实现.

图 2 中标注了 2 种采样过程的数据流通路,其中,红色通路代表简单三值采样过程,蓝色通路代表离散高斯采样过程. 执行简单三值采样时,前处理过程从存储器中读取 $2kw$ 长度的数据,分别分配到各路 a、b 寄存器中,在 Stage2.2 阶段直接进行汉明权重求解,在 Stage2.3 阶段选通 Ha 和 Hb 寄存器结果执行移位加减后直接输出. 执行离散高斯采样时,前处理过程分别从存储器读取 kw 长度的数据分配到各路 a 寄存器中,读取一行概率表中数据 ram[i]分配到各路 b 寄存器中,在 Stage2.2 阶段执行并行比较过程并统计并行比较结果暂存在 b 寄存器中,在 Stage2.3 阶段读取上一轮暂存的比较结果并与 b 寄存器的值累加,循环上述过程直到生成一组采样值.

4.2　高效的并行前处理模块结构

前处理模块是根据参数将连续的比特流分块,然后按顺序分配至每个处理通道中,事实上均匀采样的实质就是一种数据切分. 若切分长度为固定值,该过程通过简单的硬连线即可实现,若切分长度取值范围小,亦可通过数选逻辑实现. 但仅在本文统计的各类算法中,数据切分的长度就有十余种不同取值情况,采用数选逻辑实现不仅面积大、关键延迟长,且不能对本文未列出的其他算法实现有效扩展,灵活性偏低.

由于比特流切分操作的一种特性即有效位顺序保持不变,该操作可以直接理解为在一个全 0 的数据串中将低位数据按序插入指定位置,文献[14]研究并提出了应用 butterfly 网络高效实现比特插入操作的方案,给出了基于统计 1 过程的选路算法. 本文应用该算法,提出高效前处理模块结构,如图 3 所示.

4.3　高效的并行后处理模块结构

后处理模块主要实现有效数据归并、三值符号处

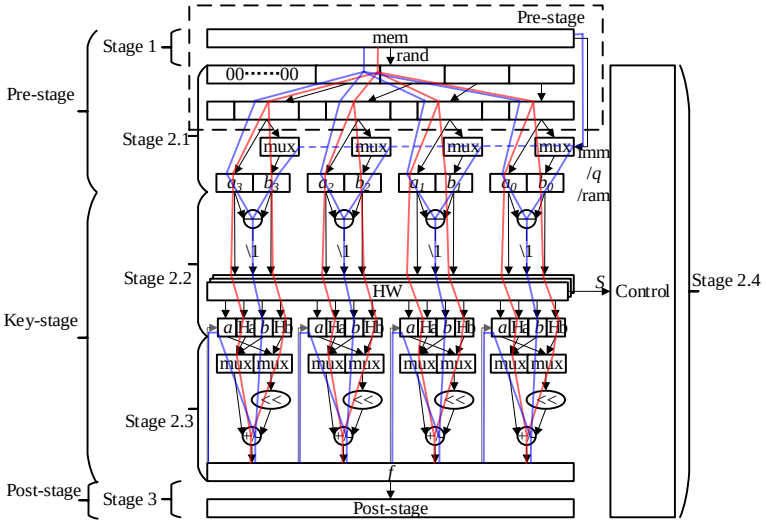


图2 资源高效的并行可重构采样加速器架构

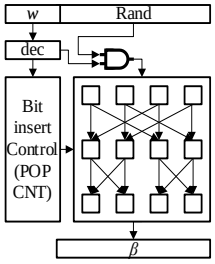


图3 高效前处理模块结构

理、数据排序与查找表. 其中,三值符号处理为乘累加与条件取反的结合操作,限定在模3域上时,乘法操作又可进一步简化为逻辑运算.

有效数据归并、查找表均是以元素为单位进行位置变换后输出,均可描述为将一系列元素中的有效元素收集到一侧,其余元素按序排列在另一侧. 有效数据归并如图4所示,假设将元素组$\{a_1,a_2,\cdots,a_n\}$与元素组$\{b_1,b_2,\cdots,b_n\}$中的有效元素进行拼接,则等价于将元素组$\{a_1,a_2,\cdots,a_n,b_1,b_2,\cdots,b_n\}$中的有效元素归集到新元素组右侧,其余元素置0.

对于查找表,并行查找的延迟和面积开销过大,且在采样算法中,每个周期也只能产生一个元素的位置信息,故本文采用逐个查找方式,将新查找到的元素按序插入已查找到的元素列表中,如图5所示. 假设从元素组$\{a_1,a_2,\cdots,a_i,\cdots,a_n\}$中查找第$i$个数据,并写入到已查找的元素列表$\{b_1,b_2,\cdots,b_n\}$中,则等价于将元素组$\{a_1,a_2,\cdots,a_i,\cdots,a_n,b_1,b_2,\cdots,b_n\}$中的$\{a_i,b_1,b_2,\cdots,b_n\}$归集到新元素组的右侧,其余元素置0.

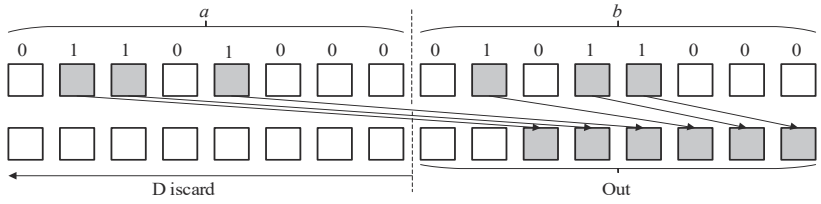


图4 有效数据归并过程示意图

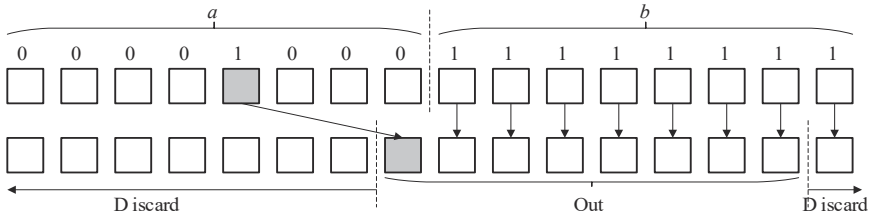


图5 单元素查找表过程示意图

可以发现,上述2种操作都是有效元素的归集操作.文献[14]证明了inverse butterfly网络可以高效完成比特级的抽取归集变换,并提出了相应的选路算法.本文将该思路应用在元素级的抽取归集操作中,只需要一个网络即可完成上述2种类型的操作,该网络结构与图3类似.

元素排序虽然也可视作是位置变换的一种,但其控制信息的产生更为复杂,每一个交叉开关的选路信号均需由一个减法器产生,当参与运算的数据量较大时,其关键路径延迟也相当大,通常采用迭代或流水的方式进行设计.故而,本文为元素排序单独设计一个加速单元,该单元只有在固定权重三值采样情形下被使用(即后量子密码中的NTRU算法),可提供给用户选择空间.在各类排序算法中,堆排序、归并排序、双调排序等均为性能表现较好的并行算法,其中,以双调排序更适合并行性的硬件实现,其在资源和性能上均占据明显优势,且为固定时间算法.如图6所示,以8输入为例的双调排序网络,其中交互单元1为2输入的升序排序,交互单元0为2输入的降序排序.对于n个数据的输入,经$\sum_{i=1}^{\log_2^n} i$级比较后输出结果.由于排序在采样算法中的使用频次较低,且只在固定权重三值采样中使用,为降低其对电路主频的影响,本文以时钟周期换取电路主频,采用流水方式实现该网络.

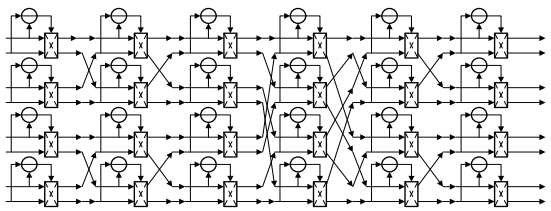


图6　8输入双调排序网络结构

5　实验与比较

在本节中,使用Verilog HDL对第4节提出的硬件结构进行描述,使用40 nm CMOS工艺库进行逻辑综合,在不同的并行度情况下,采样器的主频与面积变化趋势如图7所示.

图2所示结构的主频受并行执行部分(即Stage2部分)和串行执行部分(包括控制电路等外围逻辑)共同控制.假设串行执行部分的关键路径延迟为T_s,并行执行部分的关键路径延迟为T_p,则采样器的主频F可以表示为式(6):

$$F=\frac{1}{\max\left\{T_s,T_p\right\}} \tag{6}$$

又由于并行执行部分存在诸如求解汉明权重以及乘累加的操作,致使并行执行部分的关键路径延迟T_p随并行度k(即数据带宽)的增加而增加,可以进一步表示为式(7):

$$T_p=T_{p0}+k\cdot T_{p1} \tag{7}$$

其中,T_{p0}为一个基本并行通道的关键路径延迟,T_{p1}为增加一个并行通道所增加的延迟.

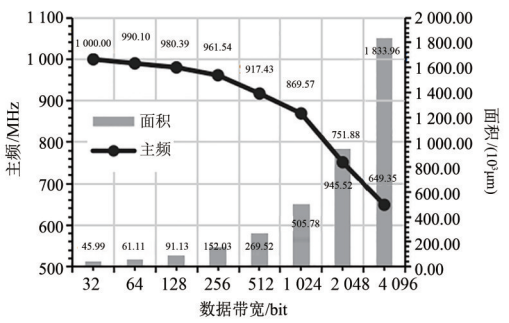


图7　采样器主频、面积随输入数据带宽变化趋势图

如图7所示,当数据带宽为32 bit时,实验表明电路的关键路即存在于加法链部分,即$T_s<T_{p0}<T_p$,故图7中主频随并行度变化的关系遵循式(8)所示关系,即类线性关系.

$$F=\frac{1}{T_{p0}+k\cdot T_{p1}} \tag{8}$$

同理,采样器的面积A为串行部分面积A_s和并行部分面积A_p之和.假设A_{p0}为一个基本并行通道的面积,则采样器的面积A如式(9)所示:

$$A=A_s+A_p=A_s+k\cdot A_{p0} \tag{9}$$

即电路面积随并行度增加线性增加,与图7中面积增长趋势相符.

为更直观地研究采样器的资源效率,在采样操作确定的情况下,时钟周期的消耗是一个定值,故而,本文使用主频与数据带宽的乘积代表采样器的性能,以性能与面积的比值代表资源效率,绘制可重构格基后量子密码采样器性能及单位面积性能变化趋势图,如图8所示.

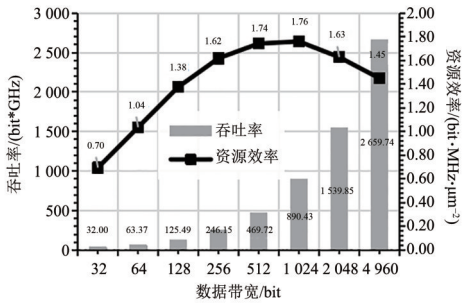


图8　采样器性能及单位面积性能变化趋势图

图8中,采样器吞吐率P可以表示为式(10):

$$P=F\cdot k=\frac{k}{T_{p0}+k\cdot T_{p1}} \tag{10}$$

求其一阶导数如式(11)所示,即恒为正值,故吞吐率恒定随并行度k增加而增长,其增长幅度随并行度k增加而降低. 其变化趋势与图8所示一致.

$$P'=\left(\frac{k}{T_{p0}+k\cdot T_{p1}}\right)'=\frac{T_{p0}}{\left(T_{p0}+k\cdot T_{p1}\right)^2} \tag{11}$$

同理,采样器资源效率E及其一阶导数E'可以分别表示为式(12)和式(13):

$$E=\frac{P}{A}=\frac{k}{\left(T_{p0}+k\cdot T_{p1}\right)\cdot\left(A_s+k\cdot A_{p0}\right)} \tag{12}$$

$$E'=\frac{A_sT_{p0}-A_{p0}T_{p1}k^2}{\left(A_sT_{p0}+kA_sT_{p1}+kA_{p0}T_{p0}+k^2A_{p0}T_{p1}\right)^2} \tag{13}$$

即采样器资源效率E随并行度k的增加先增后减,当$k=\sqrt{A_sT_{p0}/A_{p0}T_{p1}}$时达到峰值,其变化趋势与图8所示一致.

如图8所示,资源效率在输入数据带宽为1 024时达到最高. 基于此趋势,本文选取输入数据带宽为1 024的情况,提出资源高效的可重构后量子密码采样器架构.

依据以上分析过程,本文实现了输入数据带宽为1 024的采样器. 基于Virtex-7 FPGA芯片进行了仿真综合,最高工作频率可达到192 MHz,表2给出了7种采样过程在FPGA上的性能表现. 基于40 nm CMOS(ss-125 ℃)工艺库完成了布局布线,最高工作频率可达667 MHz,表3给出了7种采样过程的后仿性能表现.

表2 7种不同采样过程性能表现(FPGA,采样n个值)

算法	参数	输入数据带宽	并行度	频率/MHz	时钟周期数	时延/ns	LUT	FF
均匀	n=256, q=3 329	1 024	64	192	4	20.8	39 150	13 438
拒绝	n=256, q=3 329	1 024	64	192	15(均值)	78		
二项	n=256, nr<8	1 024	64	192	12	62.4		
简单三值	n=509, q=2 048	1 024	64	192	24	124.8		
非负相关三值	n=701, q=8 192	1 024	64	192	83	431.6		
固定权重三值	n=509, q=2 048	1 024	32	192	7 456	38 771.2		
离散高斯采样	n=1, σ=2, λ=53, Depth=18	1 024	16	192	2	10.4		

表3 7种不同采样过程性能表现(后仿,采样n个值)

算法	参数	输入数据带宽	并行度	频率/MHz	时钟周期数	时延/ns	面积/um^2
均匀	n=256, q=3 329	1 024	64	667	4	6	652 807
拒绝	n=256, q=3 329	1 024	64	667	15(均值)	22.5	
二项	n=256, nr<8	1 024	64	667	12	18	
简单三值	n=509, q=2 048	1 024	64	667	24	36	
非负相关三值	n=701, q=8 192	1 024	64	667	83	124.5	
固定权重三值	n=509, q=2 048	1 024	32	667	7 456	11 184	
离散高斯采样	n=1, σ=2, λ=53, Depth=18	1 024	16	667	2	3	

为更全面地对本文提出的可重构采样器进行评估,利用PTPX工具进行功耗测试,平均功耗为0.54 W. 表4列出了近年来较为典型的采样器实现研究.

文献[11]与文献[15]均针对包含均匀-拒绝、二项、三值、离散在内的多种采样过程进行结构设计,但2篇文献中只展示了其中1种或2种采样过程的部分性能数据. 与文献[11]相比,本文提出的采样器的ATP值与能耗约为文献[11]的30.23%. 与文献[15]相比,本文提出的采样器能耗较之降低了约31.6%. 一方面是因为采样器并行度较低,完成一次采样的时钟数较多,累计能耗较高. 另一方面,2篇文献均采用分离的方式设计了多种采样过程,并未充分利用硬件资源,从而引入了较高的冗余面积与能耗.

文献[16]采用的技术路线是基于通用处理器平台进行指令扩展,针对二项采样取得了较好的效果,进行一次二项采样消耗的能量比本文低约23倍,但其未针对拒绝采样等其他采样类型进行优化加速,故而进行一次拒绝采样的能耗反比本文高出约4倍. 文中未给出三值和离散采样的数据,但由于未对三值和离散采样设计加速单元,仅依靠RISC-V原始指令集,虽然可预见其能实现这2种采样,但其性能与能效应当较低.

文献[17]针对均匀、拒绝、二项和离散高斯采样设计了一个较为灵活的采样单元,但不能实现三值相关采样. 与文献[17]相比,本文完成一次拒绝采样操作的能耗约为其的9.9%,完成一次二项采样的能耗约为其的50.97%. 由于文献[17]是一个完整的PQC芯片,仅

表4　后量子密码算法中采样器性能对比

文献	工艺	算法	参数	频率/MHz	时钟周期数	时延/ns	面积③	ATP/(10^{-3}×s·gates)	能耗/(nJ/op)
Sapphire[11]	40 nm	均匀-拒绝 二项	n=256,q=7 681 n=256, nr<4	72	461 505	6 402.77 7 013.88	8.48② k gates	54.295 59.477	50.90 58.20
ESSCIRC[15]	28 nm	二项	n=256, nr<4	160	191	1 193.75	—	—	14.22
CHES[16]	28 nm	均匀-拒绝 二项	n=256, q=3 329	500	4 657 32①	9 314 64	166.6 k gates	1 551.7 10.66	60.541 0.416
JSTS[17]	28 nm	拒绝 二项	n=256, q=3 329	83	808 99	9 734.94 1 191.96	—	—	155.76 19.07
本文	40 nm	均匀-拒绝 二项	n=256, q=3 329 n=256, nr<8	667	19 12	28.5 18	514 k gates	14.65 9.25	15.39 9.72
文献[18]	Artix-7	离散	(σ/λ/Depth) 8.5/125/110	119	49	411.76	(LUT/FF) 820/837	(ms·LUT / ms·FF) 0.34/0.35	—
TC[19]	Virtex-7	离散	(σ/λ/Depth) 8.5/125/110 3.3/64/33	162 218	3 2	18.52 9.17	(LUT/FF) 1 049/566 1 278/306	(ms·LUT / ms·FF) 0.019/0.010 0.012/0.002	—
本文	Virtex-7	离散	(σ/λ/Depth) 8.5/125/110 3.3/64/33	192	14 3	72.8 15.6	(LUT/FF) 39 150/13 438	(ms·LUT / ms·FF) 2.85/0.978 0.61/0.21	—

注:①原论文中给出的数据包含随机数生成过程,此处引用数据依据原文描述1个时钟生成的采样数据个数换算得到;②原文给出的是整块芯片面积,此处引用数据依据采样模块的占比与整体芯片面积乘积换算得到;③本文提出的架构均同时支持7种采样方案,面积均为整体架构的数据.

给出整块芯片的面积而未给出采样单元的面积数据,故而此处不讨论其ATP值.

文献[18]与文献[19]为单独针对离散高斯采样过程进行的研究,与文献[18]相比,本文提出的采样器速度提高了约6倍,但由于本文的设计还包含了均匀、拒绝、二项、三值等采样过程的实现,且并行度较高,故而资源消耗显著. 而与文献[19]相比,速度降低了约2~4倍. 这是因为文献[18]与文献[19]均为硬件参数固化的单独离散采样设计方案,可以针对固定参数进行优化. 而本文采用的是算法与参数均可重构的设计路线,灵活性更强.

6　结论

本文以格基后量子密码算法中的采样过程为对象进行研究,针对涉及的7种采样类型分别提出了高效的并行实现模型,提炼了各类采样过程的4种共同运算逻辑. 以4种运算逻辑为核心,提出了一个高效可重构并行采样融合算法,并以参数化方式提供了一种适应任意并行度的并行采样解决方案. 以算法为牵引,提出了一个参数化的可重构并行采样加速器架构模型,并结合逻辑综合实验结果确定了架构模型的最佳并行度参数. 本文提出的解决方案可配置实现二项分布采样、拒绝采样、均匀采样、三值采样以及离散高斯采样,同时可适配任意并行度的向量处理架构,在提高运算加速单元灵活性的同时取得了较高的资源和能量效率. 与杂凑单元、多项式乘法单元相结合,可进一步展开高能效可重构后量子密码芯片的研究设计,为后量子密码的实际部署提供一种灵活高效的解决方案.

参考文献

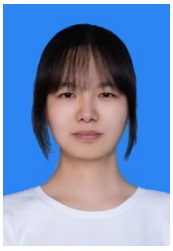
[1] 何诗洋, 李晖, 李凤华. 面向格基密码体制的高效硬件实现研究综述[J]. 密码学报, 2021, 8(6): 1019-1038.
HE S Y, LI H, LI F H. A survey on high-efficiency hardware implementation for lattice-based cryptosystem[J]. Journal of Cryptologic Research, 2021, 8(6): 1019-1038. (in Chinese)

[2] 王良成, 石元兵, 张舒黎, 等. 后量子密码迁移研究[J]. 通信技术, 2023, 56(8): 999-1006.
WANG L C, SHI Y B, ZHANG S L, et al. Research on post-quantum cipher migration[J]. Communications Technology, 2023, 56(8): 999-1006. (in Chinese)

[3] GÖTTERT N, FELLER T, SCHNEIDER M, et al. On the design of hardware building blocks for modern lattice-based encryption schemes[C]//Cryptographic Hardware and Embedded Systems - CHES 2012. Berlin: Springer, 2012: 512-529.

[4] ODER T, GÜNEYSU T. Implementing the NewHope-simple key exchange on low-cost FPGAs[C]//Progress in Cryptology - LATINCRYPT 2017. Cham: Springer International Publishing, 2019: 128-142.
[5] CHEN C, DANBA O, HOSTEIN J, et. al. NTRU algorithm specications and supporting documentation[EB/OL]. (2020-09-30) [2024-11-15]. https://www. ntru. org/resources.shtml.
[6] ROY S, REPARAZ O, VERCAUTEREN F, et al. Compact and side channel secure discrete Gaussian sampling[EB/OL]. (2014-07-31)[2024-11-15]. https://eprint.iacr.org/2014/591.
[7] DU C H, BAI G Q. Towards efficient discrete Gaussian sampling for lattice-based cryptography[C]//2015 25th International Conference on Field Programmable Logic and Applications (FPL). Piscataway: IEEE, 2015: 1-6.
[8] KARL P, SCHUPP J, FRITZMANN T, et al. Post-quantum signatures on RISC-V with hardware acceleration[J]. ACM Transactions on Embedded Computing Systems, 2024, 23(2): 1-23.
[9] FRITZMANN T, SIGL G, SEPÚLVEDA J. RISQ-V: Tightly coupled RISC-V accelerators for post-quantum cryptography[J]. IACR Transactions on Cryptographic Hardware and Embedded Systems, 2020(4): 239-280.
[10] ZHU Y H, ZHU W P, ZHU M, et al. A 28nm 48KOPS 3.4μJ/op agile crypto-processor for post-quantum cryptography on multi-mathematical problems[C]//2022 IEEE International Solid-State Circuits Conference (ISSCC). Piscataway: IEEE, 2022: 514-516.
[11] BANERJEE U, UKYAB T S, CHANDRAKASAN A P. Sapphire: A configurable crypto-processor for post-quantum lattice-based protocols[J]. IACR Transactions on Cryptographic Hardware and Embedded Systems, 2019(4): 17-61.
[12] AIKATA A, MERT A C, IMRAN M, et al. KaLi: A crystal for post-quantum security using kyber and dilithium[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2023, 70(2): 747-758.
[13] XIN G Z, HAN J, YIN T Y, et al. VPQC: A domain-specific vector processor for post-quantum cryptography based on RISC-V architecture[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2020, 67(8): 2672-2684.
[14] HILEWITZ Y, LEE R B. Fast bit gather, bit scatter and bit permutation instructions for commodity microprocessors[J]. Journal of Signal Processing Systems, 2008, 53(1): 145-169.
[15] KIM B, PARK J, MOON S, et al. Configurable energy-efficient lattice-based post-quantum cryptography processor for IoT devices[C]//ESSCIRC 2022- IEEE 48th European Solid State Circuits Conference (ESSCIRC). Piscataway: IEEE, 2022: 525-528.
[16] YE Z W, SONG R B, ZHANG H, et al. A highly-efficient lattice-based post-quantum cryptography processor for IoT applications[J]. IACR Transactions on Cryptographic Hardware and Embedded Systems, 2024, 24(2): 130-153.
[17] KIM B, MUN H G, KIM S, et al. A 1.03MOPS/W lattice-based post-quantum cryptography processor for IoT devices[J]. Journal of Semiconductor Technology and Science, 2024, 24(1): 55-61.
[18] TIAN S Q, WANG W, SZEFER J. Merge-exchange sort based discrete Gaussian sampler with fixed memory access pattern[C]//2019 International Conference on Field-Programmable Technology (ICFPT). Piscataway: IEEE, 2019: 126-134.
[19] KARABULUT E, ALKIM E, AYSU A. Efficient, flexible, and constant-time Gaussian sampling hardware for lattice cryptography[J]. IEEE Transactions on Computers, 2022, 71(8): 1810-1823.

作者简介



别梦妮 女,1997年2月出生于湖北省荆州市. 现为信息工程大学计算机科学与技术专业博士研究生. 主要研究方向为后量子密码处理器设计.
E-mail: raspberry0213@126.com

李 伟 男,1983年11月出生于天津市. 现为信息工程大学教授. 主要研究方向为体系结构、安全芯片设计、集成电路技术.
E-mail: try_1118@163.com