**郑州大学毕业设计（论文）**

题 目： 同态加密硬件加速器模拟系统设计

指导教师： 范文兵 职称： 教授

指导教师(校外)： 赵路坦 职称： 博士后

学生姓名： 蔡明 学号： 201924110803

专 业： 电子信息工程

院（系）： 电气与信息工程学院

完成时间： 2022年5月16日

2022年 5 月 16

摘要

全同态加密（Fully Homomorphic Encryption, FHE）技术保护了数据安全，但是，FHE的计算开销是明文直接计算开销的上万倍。因此，为了弥补性能上的巨大差距，最近越来越多的基于GPU、ASIC和FPGA的硬件加速器开始出现。但是在硬件加速器设计过程中，如何选择合适的加速器架构参数是很困难的，因此加速器的设计过程中往往需要软件模拟器的配合。本文通过使用python语言进行模拟加速器架构，设计了一款基于同态加密FPGA硬件加速器的软件模拟器，该模拟器可以辅助架构设计人员调整加速器架构设计中的一些参数，比如外部存储内存带宽、加速器输入缓存和输出缓存的大小以及加速器片上带宽等等。该模拟器通过模拟执行一组同态操作，来得到参数与执行这组同态操作所需总周期数的关系。最后通过分析参数与执行时间之间的关系对如何选取参数进行了介绍。

**关键词** 全同态加密，加速器，模拟器，FPGA

**Abstract**

Though Full Homomorphic Encryption (FHE) technology protects the data security, the computing cost of FHE is tens of thousands of times that of plaintext direct computing. Therefore, in order to make up for the huge gap in performance, more and more hardware accelerators based on GPU, ASIC and FPGA have emerged recently. However, it is difficult to select appropriate architecture parameters in the process of accelerator design. So, the accelerator design process often needs the cooperation of software simulator. I design a software simulator based on FPGA hardware accelerator by using python language to simulate the accelerator architecture. This simulator can help architecture designers adjust some parameters in the accelerator architecture design, such as external storage memory bandwidth, the size of accelerator input cache and output cache, on-chip bandwidth. The simulator simulates the execution of a set of homomorphic operations to obtain the relationship between parameters and the total number of cycles required to perform this set of homomorphic operations. Finally, the selection of parameters was introduced by analyzing the relationship between parameters and execution time.

**Key Words** fully homomorphic encryption, accelerator, simulator, FPGA

目 录

**摘要…………………………………………………………..……….….……..…...….I**

**Abstrct…………………………………………………………………...……..………I**

**1**绪论**………………………………………………………………….….……………1**

[1.1 研究背景和意义 1](#_Toc105099934)

[1.2 国内外研究现状 1](#_Toc105099935)

[1.3 本文主要工作 3](#_Toc105099936)

[2 同态加密介绍和理论分析**… … … …. …. … … … … … … … … … … … … … … … … … … 5**](#_Toc105099938)

[2.1 同态加密算法概述 5](#_Toc105099939)

[2.2 CKKS算法原理 6](#_Toc105099943)

[2.3 本章小结 6](#_Toc105099943)

[3 全同态加密硬件加速器**… … … … … … … … … … … … … … … … … … … … … … … … 13**](#_Toc105099949)

[3.1 现有加速器架构分析 13](#_Toc105099950)

[3.2 本文所模拟的加速器架构介绍 19](#_Toc105099953)

[3.3 本章小节 19](#_Toc105099953)

[4 模拟器设计和实现**… … … … … … … … … … … … … … … … … … … … … … … … … … 25**](#_Toc105099957)

[4.1 模拟器系统设计框架 25](#_Toc105099958)

[4.2 模拟器子模块设计与实现 27](#_Toc105099962)

[4.3 本章小结 27](#_Toc105099962)

[5 模拟器的实验测试和结果分析**… … … … … … … … … … … … … … …… … … … … … 28**](#_Toc105099957)

[6 结论**… … … … … … … … … … … … … … … … … … … … … … … … ….. … … … … … … …** 34](#_Toc105099967)

[6.1 研究总结 34](#_Toc105099968)

[6.2 后续展望 34](#_Toc105099969)

**致谢 … … … … … … … … … … … … … … … … … … … … … … … … … … …. … … … … … …26**

**参考文献 … … … … … … … … … … … … … … … … … … … … … … … … … … … … … … … 26**

**附录1 外文资料翻译… … … … … … … … … … … … … … … … … … … … … … … … … … 27**

A1.1 译文：×××……×××…………………….……..…………………..………….…......….27

A1.2 原文：×××……×××…………………….……..………………………………….…….…31

1. 绪论

1.1研究背景和意义

同态加密可以在保持数据的加密的同时进行计算，这颠覆了传统加密方式无法加密计算的局面。同态加密可以广泛应用于云计算、金融科技、多方安全计算和健康医疗等领域，为保护数据安全提供了安全和高效的解决方案。在云计算领域，同态加密可以解决在云平台上进行计算时数据隐私泄露的问题。同时，同态加密可以减少多方参与计算时数据交换的问题，大大提高计算速度。在金融科技领域，同态加密可以实现保证隐私的交易处理，可以有效防止黑客入侵和数据窃取。在健康医疗领域，同态加密可以保护个人隐私，同时允许研究人员访问高度敏感的健康数据，以促进医学研究的进步。

但是同态加密，尤其是全同态加密的运算会导致计算成本、时间开销达到明文直接计算的上万倍。因此，为了提高FHE的效率，加速器的设计是一个不可避免的方向，近年来也出现了越来越多的基于FPGA、GPU和ASIC的硬件加速器。但是，为了设计出合理高效的硬件架构，在设计过程中往往需要进行繁琐的实验进行架构参数调整，而通过模拟器进行硬件加速器的设计调参可以大大降低研究时间和成本。因此，一种能够辅助研究者进行加速器设计调参的模拟器是非常必要的。

1.2国内外研究现状

1.2.1 同态加密加速器设计的研究现状

对于同态加密技术的应用场景，如云计算和物联网中，计算速度和保护隐私往往是一对矛盾。在这种情况下，研究人员已经开始探讨使用硬件加速器来提高同态加密算法的计算性能。目前主要的研究成果如下：

2018年，Axel Feldmann等人提出了一种名为F1的计算加速器[1]，用于支持全同态加密（FHE）技术下的安全高效计算。该加速器是使用Verilog HDL描述并实现，在Xilinx Virtex-7 FPGA平台上进行了测试。

2020年，Nikola Samardzic等人基于Axel Feldmann等人提出的F1加速器，结合FHE的密文大小通常非常大这一背景对加速器进行了改进，提出了CraterLake加速器[2]。

此外还有Yinghao Yang等人提出的Poseidon加速器[3]、Rashmi Agrawal[4]等人提出的FAB加速器以及Soran Ahmed[5]等人提出的一种针对RSU资源受限网络中全同态加密应用和安全性需求特点的加速器设计，实现了FHE算法的低功耗和高效计算。

1.2.2 硬件加速器模拟器设计的研究现状

硬件加速器模拟器是一个用于模拟硬件加速器的软件工具，它可以在不需要专用硬件平台的情况下，评估加速器的性能、功耗等指标。模拟器又分为系统级模拟器、功能级模拟器和行为级模拟器，本文主要考虑功能级模拟器。在针对功能级模拟器设计方面目前有:

Ananda Samajdar等人针对DNN卷积神经网络脉动阵列加速器设计的周期精准的SCALE-SIM模拟器[6]。SCALE-Sim是一个专为深度学习应用设计的模拟器，可以支持各种常见的DNN架构（如CNN、RNN、LSTM等）及其优化技术。

Daniel J.Sorin等人开发用于内存模拟的DRAM-Sim2[7]，它主要关注模拟DRAM中存储单元和数组的操作，以及内存控制器的操作。通过模拟内存的具体操作行为，DRAM-Sim2可以帮助开发者进行内存系统的性能分析和优化。相比于功能级模拟器，行为级模拟器更加逼真，能够更好地反映实际系统的行为。

Binkert M.I等人出的用于通用处理器的GEM5模拟器[8] 。通常被用于CPU架构相关的模拟器和仿真领域。它支持多种ISA（Instruction Set Architecture）CPU体系结构（如ARM、x86、MIPS等），可以用于研究微处理器和计算机系统的各种性能指标、缓存一致性、功耗以及体系结构调优等方面的研究。

目前还没有用于同态加密加速器设计的专用模拟器。

1.3本文主要工作

本文主要是基于现有的全同态加密加速器架构设计一种用于辅助调整同态加密加速器架构设计参数的行为级周期精准模拟器。

基于同态加密算法本身影响加速器架构，加速器架构影响模拟器设计的这样一个思路将本文主要工作分为以下内容：

（1）对同态加密算法的基本概念、应用场景以及各种同态算法进行了基本介绍，然后针对本文采用的CKKS（Cheon-Kim-Kim-Song）全同态算法进行了算法原理的分析以便后文的加速器设计和模拟器设计的阐述更加清晰。

（2）对现有的基于ASCI和FPGA的同态加密硬件加速器的架构进行了分析，根据其架构特点指出一种用于架构设计的模拟器的必要性和实用性，并详细介绍了本文所模拟的加速器架构。

（3）对基于python设计的模拟器的软件系统框图、各模块子框图以及所采用的的数据结构、数据流图等进行了介绍。

（4）介绍如何通过设计好的模拟器进行分析加速器架构参数和加速器性能关系，并对所使用的程序进行了阐述和分析。

1. 同态加密介绍和理论分析

2.1 同态加密算法概述

2.1.1 同态加密算法的介绍

在密码学领域中，使用同态作为一种加密类型。同态加密（HE）是一种加密方案，它允许云服务商这样的第三方对加密的数据执行某些可计算的功能，同时保留加密数据的功能和格式的特征。第三方在得到同态加密的数据之后，对密文进行特定的计算，得到的计算结果再进行同态解密就可以得到和使用明文直接进行计算相同的结果。

典型的同态操作有加法同态和乘法同态，如图2.1所示，假设有两条明文消息m1和m2，则有：

对于同态加法，可以满足等式（2.1）和（2.2），其中Dec是同态解密函数、Enc是同态加密函数[10]。

(2.1)

(2.2)

对于同态乘法，可以满足等式（2.3）和（2.4），其中Dec是同态解密函数、Enc是同态加密函数。 对于HE（Homomorphic Encryption）中的表示法，它只允许加法和乘法运算。对于任何布尔电路，只能通过XOR门执行加法和和And门执行乘法[9]。

(2.3)

(2.4)

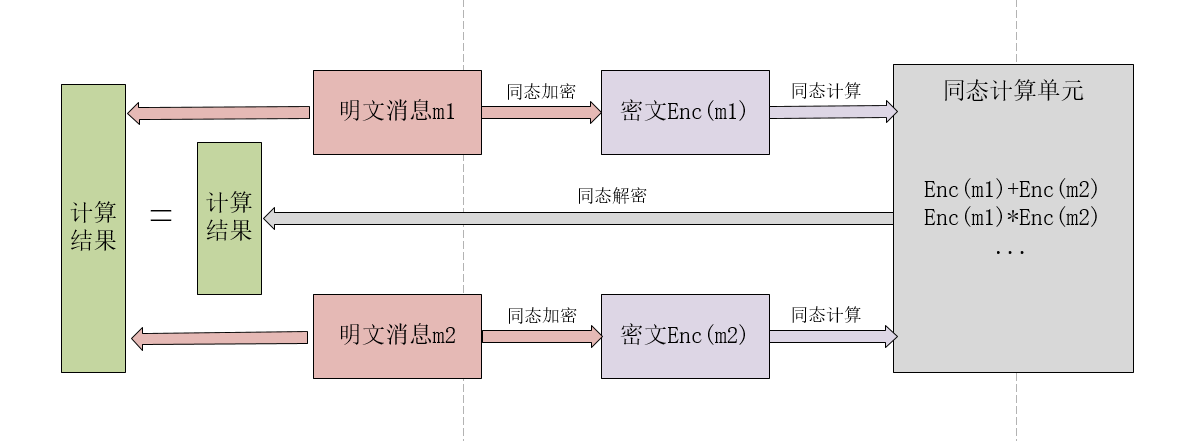


图2.1 同态加密计算示意图

2.1.1 同态加密算法的分类

在同态加密技术中，为了能够实现在加密状态下进行计算，需要将明文经过加密算法转换成密文，同时添加一个随机噪声。当使用同态加法或者同态乘法对密文进行运算时，这个随机噪声会随着运算而不断累加，导致密文中的噪声变得越来越大。

当密文中的噪声超过一定的阈值后，就会影响计算结果的准确性，从而导致解密结果的错误。一般情况下，同态加密算法会规定一个最大的可以容忍的噪声上限，如果噪声超过这个上限，则需要重新生成密文或者采用其他技术处理。因此，同态加密算法并不是都可以进行无数次的同态加法或者同态乘法的，同态加密算法可以根据对允许操作的数量的加密数据整齐地归类为三种类型的方案：

PHE（Partially Homomorphic Encryption）部分同态加密，如Paillier加密算法EIGamal加密算法和RSA（Rivest Shamir Adleman）加密算法[9]，由于其设计的简单性，可以运行在资源有限的设备上，并且不需要过多的密文噪声控制。因此，PHE通常拥有很快的运行速度和较低的计算成本，但其功能受到了一定程度的限制。下面对典型PHE算法进行简单的介绍：

Paillier加密算法，该算法的分析主要包括密钥生成、加密、解密和同态操作实现。

。

(2.5)

。其中lcm（a,b）是指获得a和b的最小公倍数。

(2.6)

(2.7)

步骤3：随机选择一个满足等式（2.8）的g，等式（2.8）中的函数T如公式（2.9）所示。

(2.8)

(2.9)

步骤4：通过公式（2.10）和（2.11）生成公钥pubk和私钥seck。

(2.10)

(2.11)

。

步骤2：通过公式（2.12）对一个明文pla进行加密得到密文cic。

(2.12)

* 通过公式（2.13）可以将密文cic转成明文pla。

(2.13)

* 公式（2.14）表示了Paillier加密算法是如何实现对密文和密文同态乘法的。

(2.14)

SWHE（Somewhat Homomorphic Encryption）有限同态加密，如Gentry-Halevi-Smart(TFHE)算法、Brakerski-Gentry-Vaikuntanathan(BGV)算法和Van Dijk-Landais Encrypt Scheme(VLLE)算法。其通常支持多个同态运算类型，如同态加法和少量同态乘法。相比于PHE，SWHE提供的更大的功能集，可以执行更复杂、更实用的计算任务。由于SWHE使用了更复杂的算法并且需要对噪声进行更精细的控制，在效率方面会稍逊于PHE。

FHE（Fully Homomorphic Encryption）完全同态加密，如Gentry-Full Homomorphic EncryptionAlgorithm、Brakerski-Gentry-Vaikuntanathan-Smart(FV)算法和Cheon-Kim-Kim-Song(CKKS)算法，其是同态加密中最强的形式。FHE支持任何同态运算类型，包括任意多次的同态加法和同态乘法，即便密文也能直接执行相应操作。FHE相比于前两者具有更高的安全级别和更高的功能性，但由于其使用了更加复杂的算法和更精细的噪声控制，因此在效率方面通常比SWHE和PHE更慢。

2.2 CKKS算法的原理

由于FHE算法可以进行任意多次的同态加法和同态乘法，所以FHE的实用性更高，并且FHE的功能性和安全性也更高。此外，FHE算法中CKKS算法由于其采用近似计算，所以它的细节得到了很大的简化，并且计算效率也有很大的提升，所以现在的加速器主要针对的便是CKKS算法。本小节将分析本文模拟的加速器所采用的CKKS算法的原理。

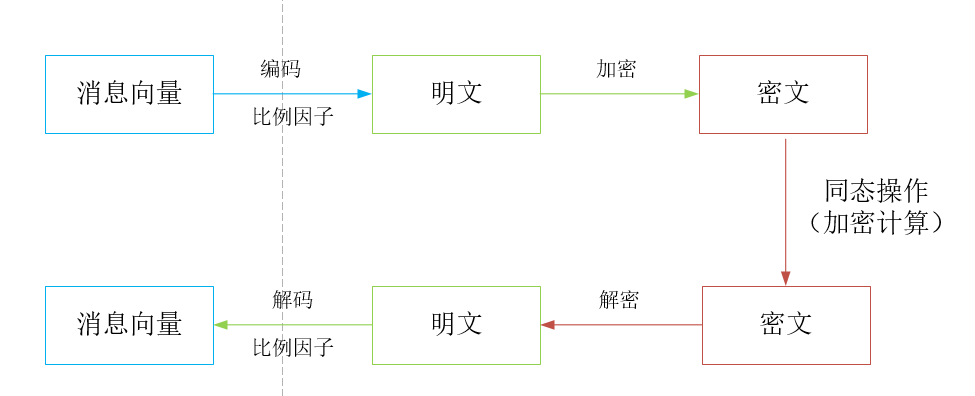


图2.2 CKKS计算示意图

如图2.2所示，是CKKS算法的工作流程，由于CKKS算法的处理对象是复数域，所以要先通过编码，将复数消息向量编码成整数多项式，也就是,然后在进行加密和同态操作，同态计算结果解密之后也需要相应的进行解码变成实际的复数域消息。接下来将逐一介绍CKKS算法的编码、解码、密钥生成、加密、解密、同态加法、同态乘法、重线性化、重缩放、旋转、密钥转换（keyswithing）、自举（bootstrapping）以及安全性。

* 编码：完成如（2.15）所示的将消息空间的信息 编码到明文空间。

(2.15)

步骤1：

由于的长度为,并且是复数，所以可以通过将个元素取共轭，然后连接到的后面得到长度为n的向量。这里之所以要取共轭得到一个长度为n的向量，是因为后面需要用到共轭的这个特性得到实数多项式。

步骤2：

现在我们得到了一个长度n的复数向量，其中前个元素和后个元素是共轭的。我们又知道在复数域内多项式有n个复数根，并且前个复数根和后个复数根是共轭的，这n个复数根组成一个向量。

那么便可以通过寻找一个插值多项式，使其满足公式（2.16），其中表示的便是图2.2中的比例因子，通过它可以将的数值放大。当的数值放大之后，再用插值法获得插值多项式，此时的插值多项式的系数相较于没有放大就会扩大倍，那么取出这个插值多项式的系数的近似整数组成一个整数向量。这个插值多项式可以通过牛顿插值、拉格朗日插值法等方法获得。

(2.16)

通过上述两个步骤，便可以将消息空间的复数向量编码成整数多项式。

* 解码：解码就是编码的一个逆过程，也就是完成（2.17）所示的转换，将明文空间的整数多项式转成消息空间的复数向量。

(2.17)

那么只需要按照编码的相反步骤进行就可以了，如（2.18）(2.19)所示。

(2.18)

(2.19)

* 加密：加密便是把明文加密成密文，现在我们已经有明文整数多项式plaintext，我们要做的事情便是完成公式（2.20）表示的转换。其中, Q是一个模数。

(2.20)

步骤1：

首先要生成加密使用的公钥pubk和私钥seck。为了生成密钥，我们要先取一个离散高斯分布、一个错误分布以及一个随机分布。

然后，通过公式（2.21）和（2.22）获得公钥和私钥。

(2.21)

(2.22)

步骤2：那么通过公钥便可以对明文进行加密得到密文，加密公式如（2.23）所示。

(2.23)

* 解密：对于密文,我们可以通过公式（2.24）解密得到明文plaintext。

(2.24)

……………………….待完成剩余CKKS介绍……………………………

2.3 本章小结

本章首先介绍了同态加密的基本原理，然后对同态加密的三种分类及其典型代表进行了分析，分别是部分同态、有限全同态和完全全同态。然后又对本文接下来将使用的完全同态算法CKKS的原理进行了详细的分析，包括CKKS算法中的编码、解码、加密、解密、同态加、同态乘、keyswitching、bootstrapping、重线性化、重缩放以及安全性等方面。

1. 全同态加密硬件加速器

3.1 现有加速器架构分析

3.1.1 概述

同态加密作为一种重要的保护数据隐私的技术，在安全计算和数据共享领域有着广泛的应用前景。然而，由于其运算复杂度高、效率低等特点，使得同态加密在许多实际场景中很难满足实时性和可扩展性的要求。这时候，使用硬件加速器对同态加密算法进行优化可以大幅提升加密效率和计算速度。

硬件加速器通常是指专门设计的芯片或模块，可以加速计算机任务中的某些繁琐的计算过程。相比于软件实现，硬件加速器更具有并行处理能力和基于自定义指令的加速计算来支持更快、更节能的计算。使用硬件加速器对同态加密进行加速的原理是将同态加密算法的某些计算步骤或核心操作加速到专用硬件上，提高同态加密算法的整体运行效率和计算速度，并显著地减少整个加密过程中需要的处理时间。例如硬件加速器可以通过预先计算表格里面的数值，以及在同态加密中使用的FFT变换、NTT变换、多项式乘除运算、线性代数运算等核心部件使用一些特定硬件模块，使得同态加密计算的速度提升很大程度上。另外，出于安全考虑，有些智能芯片可以内嵌物理随机数发生器、TRNG (真随机数发生器) 或 PUF （可靠性不区分）技术等安全模块来保护关键数据和实现高水平的密码保护。

因此，结合专用硬件的同态加密实现能够使机密数据的加密效率有了明显的提升，也为企业的数据安全提供了更强的保障。本小节将对基于FPGA和ASIC的全同态加密加速器进行分析。

3.1.2 基于FPGA的加速器介绍

基于FPGA设计全同态加密加速器具有更高的灵活性，因为FPGA可以通过重新编程适应不同算法和应用需求，允许运行时配置，可以在开发过程中进行调试和修改。此外，FPGA加速器可以扩展到具有更多计算资源和更高性能的设备，以满足不断增长的计算需求。并且FPGA的设计成本和周期相对于ASIC来说要小的多。不过FPGA由于受到芯片内交叉连接架构的限制，它的性能可能低于专用ASIC设计。尽管大多数 FPGA 并没有完全利用所有可用硬件资源来支持 HE 计算。并且，由于FPGA通常使用很多密集的SRAM存储器，它们可能会消耗相对较高的功耗。本小节将介绍以下几个基于FPGA的全同态加密加速器：F1[1]、Poseidon[3] 和FAB[4]，并重点介绍F1加速器，因为下一小节的ASIC加速器便是基于F1加速器改进的，而我们的模拟器便是基于下一小节的ASIC加速器的架构进行设计的。

1. **F1加速器**
2. F1加速器概述：

F1是一种针对同态加密的高性能硬件加速器。它能够提供不同的计算性能和资源配置来支持流处理、批处理或自适应任务处理等。该加速器可以支持不同类型的同态加密算法，包括Gentry的基于理想格的全同态加密方案和Brakerski Gentry Vaikuntanathan (BGV) 的环同态加密方案。F1加速器主要由数字信号处理器（DSP）和固定逻辑门阵列（FPGA）构成，其中重点优化了在FPGAs中实现同态加密所需的操作 (如乘除运算、加减运算、NTT变换和多项式插值等)。其设计旨在实现尽可能高的时钟速度，并兼顾硬件面积和功耗参数，从而保障加速器的高性能和低延迟，并为用户提供灵活的编程接口。在F1加速器中，同态加密算法的原始输入数据被转化为一组抽象向量并通过专门的电路进行处理。这种转换处理可以有效降低加密过程中产生的数据冗余和计算量，使得同态加密算法的实现更快、更高效。同时，F1加速器还支持并行计算和异构计算、加密计算与明文计算的同时执行，使得加密速度更快，性能也更稳定。

② F1的加速主要是通过以下三个方面实现的：

* 对长向量的复杂操作

FHE使用达到几千个元素长的向量编码信息，并使用模算法对它们进行处理。F1采用矢量处理，并对FHE操作量身定制了宽功能单元，以实现大规模的加速。f1还采用了专门的NTT单元和自同构函数单元。因为数论变换（NTT）和自同构，不是元素级的，需要复杂的数据流，而这些数据很难实现为向量操作。

* 使用超长指令静态调度

FHE程序是对向量上的算术运算的数据流图。所有操作及其依赖关系都提前知道。F1通过采用静态调度来利用这一点：在超长指令（Very Long Instruction Word ，VLIW）处理器中，所有组件都有固定的延迟，编译器负责调度操作和组件之间的数据移动，没有硬件机制处理冲突的（即没有失速逻辑）。由于这种设计，F1可以每个周期发出许多操作，最小的控制开销；结合向量处理，F1可以每个周期发出成千上万个标量操作。

* 减小数据流移动

尽管加速器需要复杂的功能单元，但在目前的技术中，有限的片上存储和内存带宽是大多数FHE加速器的瓶颈。因此，F1主要工作的最小化数据移动。首先，F1具有显式管理的芯片上内存层次结构，内存结构有很多的bank和寄存器堆。第二方面，F1解耦数据移动，并通过在使用前加载数据来隐藏获取数据的延迟。第三方面，F1使用FHE定制的调度算法来最大化数据重用来充分的利用有限的带宽。第四方面， F1使用相对较少的具有极高吞吐量的功能单元，而不是像之前的工作那样使用低吞吐量的功能单元。这减少了必须同时驻留在芯片上的数据量，从而允许更高的重用性。

③ F1加速器所使用的两点已有研究成果：

* 通过NTT进行多项式乘法

两个多项式 相乘需要卷积它们的系数，这是一个昂贵的操作。就像用快速傅里叶变换可以更快地实现卷积一样，多项式乘法也可以用数论变换（NTT）实现得更快，这是模算法的离散傅里叶变换的一种变体。NTT以一个𝑁系数多项式作为输入，并返回一个表示在NTT域中的输入的𝑁元素向量。多项式乘法可以在NTT域内进行元素级乘法。具体来说，如等式（3.1），其中，示的是元素级乘法。

(3.1)

因为NTT只需要次模操作，所以把一个多项式乘法拆分成两个正向NTT，一个元素间乘法以及一个逆NTT就可以做到以的复杂度实现多项式乘法。

* 用残差多项式来表示很大的数。

FHE的密文元素大小很大，可以达到512位，但宽的数的算术很昂贵。在我们感兴趣的范围内，模乘数的代价随位宽度呈二次增长。此外，我们需要有效地支持广泛的宽度范围（例如，32位增量的64到512位），这既是因为程序需要不同的宽度，也因为模量切换逐步降低系数宽度。残差多项式将一个系数较宽的单个多项式表示为多个系数较窄的多项式。为了实现这一点，我们选择模量𝑄为𝐿较小的不同素数的产物。然后，中的一个多项式可以表示为𝐿个多项式，分别在中，其中第𝑖个多项式的系数是模的宽系数。

④ F1的架构图分析：

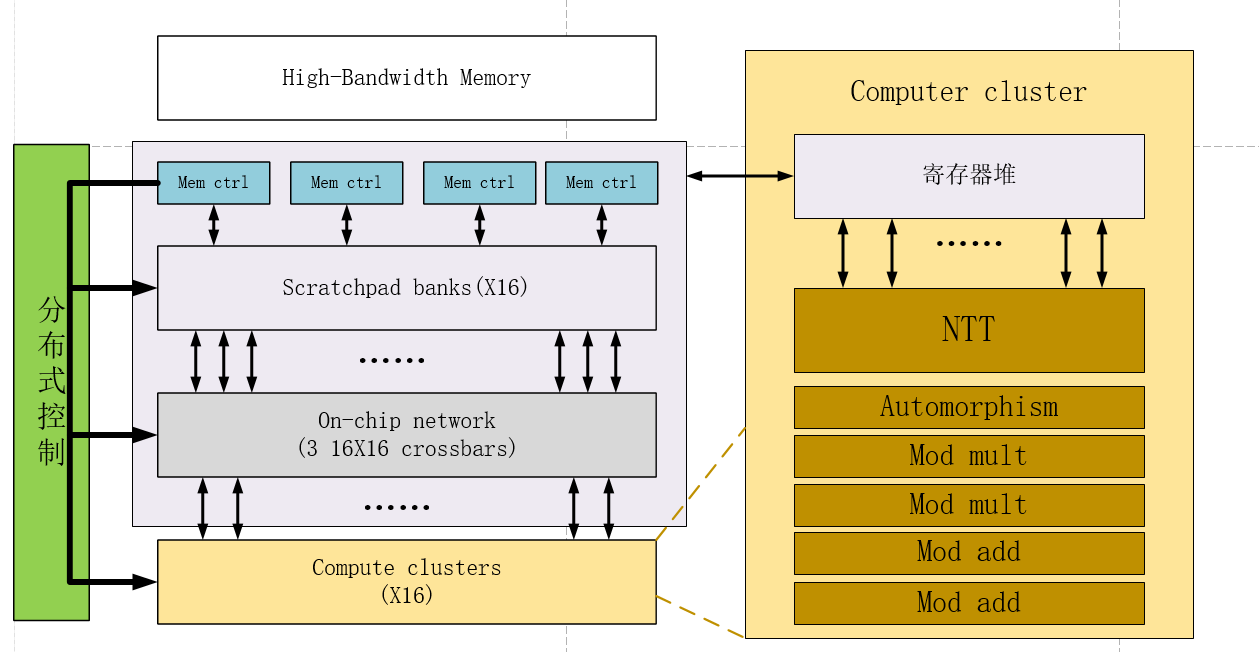


图3.1 F1加速器架构图

如图3.1所示 ,F1加速器的架构由基本向量处理功能单元、计算簇、内存系统、分布式控制以及寄存器堆五部分组成。下面一一对这五部分内容进行介绍：

* 基本向量处理功能单元

F1的特点是宽向量执行操作与功能单元（FUs）定制的原始FHE操作。具体来说，f1实现了用于模加法、模乘法、ntt和自同构的向量功能单元。因为我们利用RNS表示，这些功能单元使用固定的、较小的算术字大小（在F1实现中是32位），避免了大位宽的算术。

* 计算簇

功能单元分组到计算簇中，每个簇都有几个FUs（1个NTT、1个自同构、 2个加法器和2个加法器）和一个存储寄存器堆，它可以为每个周期提供足够的操作数，使所有FUs保持繁忙。该芯片有多个簇（在F1的实现中有16个）。

* 内存系统

F1具有一个显式管理的内存层次结构。F1具有一个大的、大量存储的暂存器。暂存器接口具有高带宽芯片外内存，并通过芯片上网络使用计算簇。F1使用解耦的数据编排来隐藏主内存延迟。暂存器从主存中获取数据，并且快于它的使用速度。连接暂存器和计算簇的片上网络必须具有非常高的带宽，因为寄存器堆很小，重用能力有限。

* 分布式控制

虽然静态调度是VLIW的标志，但F1的实现是完全不同的：在F1中，并不是每个操作有一个单独的指令流，而是每个组件有一个单独的指令流。虽然FHE程序可能有循环，但可以展开它们以避免分支情况，并将程序编译成线性的指令序列。这种方法可能会显得代价高昂。但是向量很长，所以每个指令都编码了大量的工作。此外，还支持一种紧凑的指令格式，表明了相邻两个操作之间的周期数，这种编码提高了VLIW指令的低利用率。每个FU、寄存器堆、网络交换机、暂存器和存储器控制器都有自己的指令流，控制单元从指令bank中获取并分配给组件。

* 寄存器堆

F1中的每个计算簇需要10个读端口和6个写端口来使所有FUs保持工作状态。为了以较低的成本实现这一点，F1使用了一个8-bank分区寄存器文件设计。每个向量分开存储于bank之中，每个FU随着时间的推移通过所有bank，每个周期使用一个bank。通过启动每个向量操作的开始，fu在每个周期访问不同的bank。

（2）Poseidon加速器

（3）FAB加速器·

3.1.3 基于ASIC的加速器介绍

基于ASIC设计的全同态加密加速器具有更高的性能以及更高的功率效率，并且由于ASIC可以严格控制各个计算阶段的物理状态，使其难以收到侧信道攻击等攻击手段，因此其还具有更高的安全性。本小节主要介绍一下Craterlake加速器[2]。

Craterlake是第一个使FHE程序具有无界大小（即无界乘法深度）能力的FHE加速器。这样的计算需要非常大的密文（每个达到几十个兆的大小）和不同的算法，这些以前的工作都是不支持。为了解决这一挑战，CraterLake引入了一种新的硬件架构，它可以有效地扩展到非常大的密文，新的功能单元来加速关键内核，以及新的算法和编译器技术来减少数据移动。CraterLake在深度FHE程序上评估FHE，包括像ResNet和lstm这样的深度神经网络，之前的工作在CPU上的每个推理需要几分钟到几小时。

先前的FHE加速器不能有效地处理无边界深度计算，因为它们支持的是有限大小的向量，并且它们使用的算法在深度程序中很难扩展到大型密文。因此，它们只能运行较小的FHE计算，并且它们不支持足够的深度来运行完整的引导过程。Craterlake是第一个支持无界深度FHE计算的FHE加速器。为了实现这一点，Craterlake提供了新的算法、专门的功能单元、硬件架构和编译器技术，以克服深度FHE计算的关键挑战：其极端的数据移动需求。

FHE方案编码信息是一个非常长且每一个元素的位宽都很大的向量。具体地说，支持无界深度计算需要64K个元素的向量，每个元素有1600位。这使得每个密文需要25 MB，比之前的FHE加速器的目标大12×。此外，之前的工作采用的FHE算法，需要大量的辅助数据。例如，在F1 中增加2 MB的密文需要32 MB的辅助数据，而将算法扩展到25 MB的密文需要超过1.4 GB的辅助数据，这显然不适合存储到芯片上。为了解决这一挑战，Craterlake采用一种称为增强关键的FHE算法，这消除了大部分辅助数据，将开销从1.4 GB减少到50 MB。增强关键化也降低了计算成本。

Craterlake主要做了以下几点贡献：

* 建立了一种新的宽达2048通道的矢量单处理器结构，它将每个向量操作扩展到芯片上。这种矢量单处理器方法减少了并发操作的数量，从而最小化了内存占用空间，减少了芯片外流量，并简化了编译器。
* 为了实现上述的结构，尤其是实现NTT和自同构这两个功能，Craterlake在物理通道的分布组之间建立了一个固定的转置网络。与以前的方法相比，这减少了片上数据的移动和互连成本。
* 一种新的功能单元，可以在运行中生成所需的一半辅助数据（将开销从50 MB减少到25 MB），节省片上存储和内存带宽。
* 一种向量链接技术用来构建长FU流水从而用很少的寄存器端口实现许多并发操作。

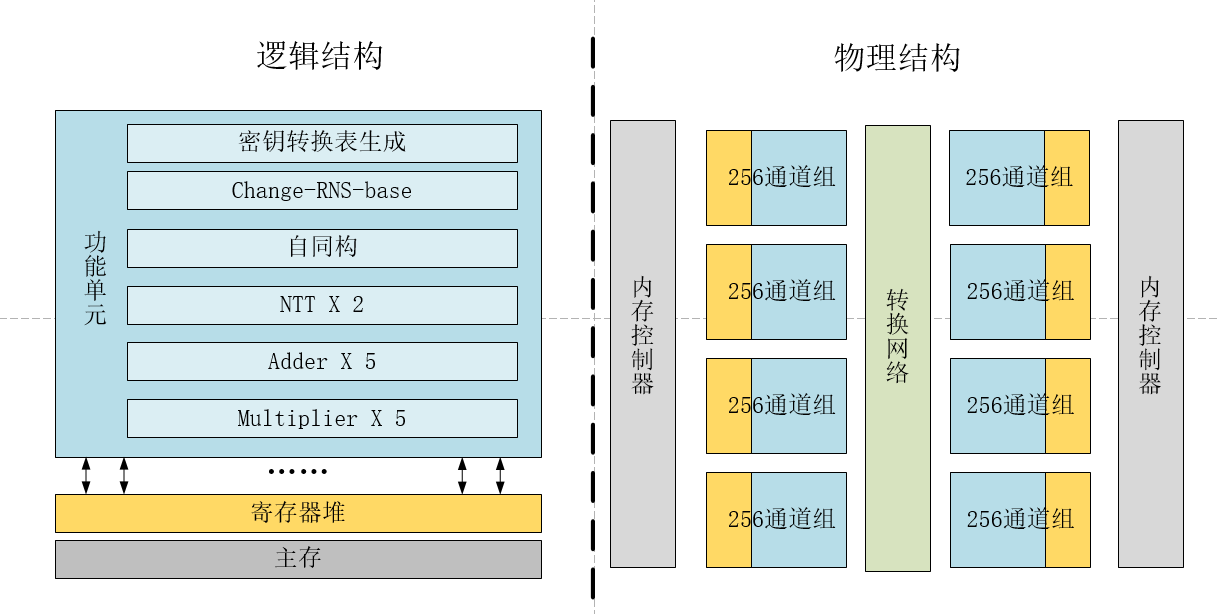


图3.2 Craterlake加速器架构

Craterlake的架构设计如图3.2所示，分为逻辑结构和物理结构两部分。

对于逻辑结构组成，主要是功能单元和内存系统。Craterlake具有专门针对FHE操作的功能单元（FUs）。包括从F1加速器中引用和改进的模块加法、模乘法、ntt和快速向量自同构模块。此外，Craterlake还提供了两个新 的功能单元：一个Change-RNS-Base单元（CRB），可以加速大量增强的密钥转换，和一个密钥转换提示生成器，可以在运行中生成每个密钥转换提示的一半，减少了内存流量和片上存储。Craterlake的片上存储被组织成一个256 MB的单级寄存器堆，由所有FUs共享。较小的寄存器堆会严重限制性能，该寄存器堆使用一个元素分区的设计来有效地模拟12个读和写端口。并且通过允许fu被链接起来形成执行更复杂操作的操作。

对于物理结构，实现一个𝐸=2048通道的向量处理器会导致片上流量过大。Craterlake通过将其车道分成𝐺=8车道组来解决这个问题。每个通道组宽为=256，占据芯片物理上不同的区域。由于通道组同时包含FU通道和寄存器堆，所以大多数数据移动可以在每个组内本地执行。

3.2 本文所模拟的加速器架构介绍

根据上一小节介绍的F1加速器架构和Craterlake架构图可以总结出完全同态加密加速器架构的一些要点，首先，加速器是采用静态结构，使用超长向量处理器的形式进行设计的。第二，完全同态加密的每一个密文都是一个向量，而且这个向量很大，在Craterlake的场景中甚至能达到几十兆一个密文大小，所以肯定要有一个内存系统，包含大容量高带宽外部存储、以及片上缓存器，我这里将把Craterlake中的内存系统进行简化，去掉内存交换网络，并添加一个FIFO输出缓存。第三，计算核的每一个功能单元都是全流水的，有固定延迟的，并且没有也不需要控制信号，所以计算核部分可以不用考虑每一个功能单元的具体实现，之余更复杂的操作，也会被分解成原操作，所以在模拟架构设计中也不用考虑。

所以参考Craterlake架构图和F1加速器架构图以及模拟器的复杂程度，我设计了如图3.3所示的这样一个架构图用于后面的模拟器设计。下面将对该架构进行详细的阐述。

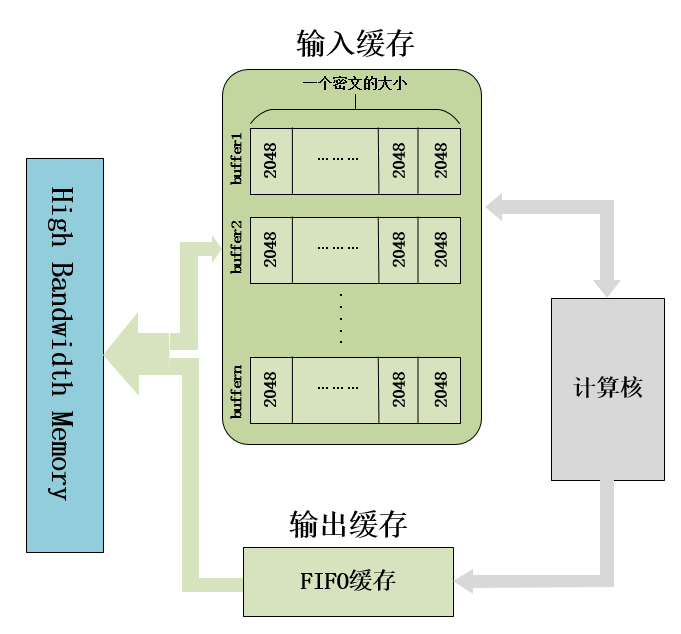


图3.3 模拟的加速器架构图

如图3.3所示，该架构由四部分组成，第一部分是最左边的高带宽内存（High Bandwidth Memory, HBM）,与传统的DDR内存相比，HBM内存具有更高的数据传输速度和更低的能耗。其设计与传统DDR结构有所不同，将多个片上存储器封装到一个多层堆叠的3D结构中，在同样大小的芯片上可以实现更高的存储密度和更短的访问延迟。此外，HBM内存还采用了基于片上互连技术的高速互联，从而进一步提升数据吞吐量，并支持与高性能处理器进行更紧密的集成。在需要大量高速内存的领域，如图形/视觉计算、人工智能、数据中心等，HBM内存已经成为越来越流行的选择。

第二部分是输入缓存，如3.3图中中间的上部分所示，整个输入缓存首先被分为多个输入子buffer，每一个输入子buffer的大小是一个密文的最大的大小，也就是32\*65536\*60bit的大小。然后，每一个buffer被分为一块一块的2048bit的大小的小存储块。这样的输入缓存结构可以减少资源占用，由于每个缓冲区只处理部分数据，因此可以有效减少系统资源占用，从而提高系统效率。采用多buffer结构的主要目的是两个，一个是提高数据的片上重用，第二个是对数据和操作进行解耦，在操作运行之前先把数据存储到buffer中，从而隐藏数据读取延迟。

第三部分是右边的计算核部分，计算核中是有着固定计算延迟和输入输出大小的各种计算功能单元，包括元素级向量加法、元素级向量乘法、快速数论变换（NTT）以及CRB等模块。

第四部分是中间的下半部分，这一部分是输出缓存，输出缓存用FIFO缓存。因为FIFO缓存非常适用于这个场景，由于FIFO缓存采用了先进先出的原则，每次离开队列的元素都是最早进入的元素，因此相比其它的缓存结构，他更好管理，在使用大量数据的时候，使用FIFO缓存会使结构更加高效。FIFO缓存非常简单易懂，没有复杂的逻辑关系和运算，只需要按照队列的顺序来读写数据即可，这使得它易于实现和调试，以便于人们理解他的工作原理。

3.3 本章小结

本章主要进行完全同态加密加速器的介绍，第一小节先是对为什么要用硬件加速器进行了介绍，然后对基于FPGA的完全同态加密加速器F1加速器的架构以及优化和加速器的方法技巧进行了描述，接着对基于ASIC的完全同态加密加速器Craterlake加速器的架构以及优化和加速器的方法技巧进行了描述。第二小节根据基于FPGA的完全同态加密加速器F1加速器和基于ASIC的完全同态加密加速器Craterlake加速器的架构进行分析和整合，提出了接下来本文将要模拟的加速器，并描述了加速器的各个部分的组成和作用。

1. 模拟器设计和实现

4.1 模拟器系统设计框架

4.1.1 模拟器概述

硬件加速器模拟器是一种软件工具，用于模拟硬件加速器的行为和性能。它们可用于开发、部署和测试许多类型的加速器，在不同级别上提供各种功能和复杂度。硬件加速器模拟器在硬件设计过程中发挥着至关重要的作用，不仅可以提高开发效率、缩短上市时间，同时还可以减少硬件故障和成本，并且为下一代加速器的开发奠定了基础。

硬件加速器模拟器的作用具体包括以下几点：开发，硬件加速器模拟器可以用于加速器设计过程中对加速器的全面验证，包括性能、功耗和可靠性等。部署，硬件加速器部署前，使用模拟器进行功能和性能验证可以减少错误和排除缺陷的时间和成本。测试，使用模拟器来测试硬件加速器在不同应用场景下的行为和性能，从而精确评估硬件加速器的性能特征，对后续优化工作具有指导作用。

硬件加速器模拟器主要有以下几种类型：

系统级模拟器：系统级模拟器旨在对整个计算机系统进行全面的模拟。它可以模拟操作系统、应用程序和硬件设备之间的交互过程，从而实现系统功能和性能的评估、验证和优化。系统级模拟器的主要作用有性能评估、可靠性评估。系统验证和计算机架构研究。性能评估，可以对计算机系统的性能进行定量分析，例如测量其处理器和存储介质的速度、带宽和延迟等等。这些数据对于基准测试、设计优化以及性能瓶颈分析非常有用。可靠性评估，可以对系统的可靠性进行分析、确定潜在问题，并查找和排除故障。此外，还可以对系统的容错特性进行测试，确保系统能够成功地恢复到预期状态。系统验证，可以通过模拟不同场景下的运行情况，对系统的完整性和正确性进行验证。这也有助于提前发现分析系统中的错误或问题，并及时纠正。计算机架构研究，可以模拟各种计算机系统，用于研究计算机体系结构的功能、优化以及新技术的设计等方面。

RTL 级别模拟器：RTL级模拟器(RTL Simulator)是一种硬件描述语言(HDL)仿真工具，通常用于验证集成电路(IC)设计的正确性。它可以将设计描述为使用逻辑电路的门级元素，并在时钟周期级别模拟其行为。RTL级模拟器有两个主要的应用场景：1.在集成电路的设计和验证阶段中，验证单个功能模块和整个系统的正确性和性能；2.在驱动开发和测试过程中，执行各种信号分析以识别问题或故障点。RTL级模拟器的主要原理是语法分析和抽象模型生成。简单来说，它通过读取并解释HDL代码生成模型，然后对输入注入刺激信号并对输出进行分析。由于RTL级模拟器对门级元素的模拟，它们运行速度较快，因此也称为“快速仿真器”。常见的RTL级模拟器包括ModelSim、VCS、Verilator等等。这些工具通常支持多种硬件描述语言，如Verilog和VHDL，以及C/C++等高级编程语言。RTL级模拟器可模拟从复杂的处理芯片到嵌入式微控制器的各种电路，同时还可以直接与电路板（PCB）中的设备集成。它可验证电路的各种功能，检测和处理时序问题、信号噪声、功耗等特征。RTL级模拟器在实际工程中发挥着非常重要的作用。它们可以提供一种可靠的方法来开发并进行IC设计的验证。在代码编写过程中，使用仿真器可以帮助排除潜在的硬件漏洞，从而节省设计时间，并且加速硬件产品上市速度。同时，这些工具还可以探究CPU驱动程序的行为，以便找到性能瓶颈并进行优化。

行为级别模拟器：行为级模拟器（Behavioral Simulator），主要用于逻辑电路的行为级仿真并验证。与RTL模拟器不同，行为级模拟器主要关注实现设计功能、处理器性能，并对目标体系结构的管线程序进行高层次的仿真评估。行为级模拟器可以处理高级编程语言描述的设计，并将其转换为一组硬件操作。在真实物理设备中，这些操作会直接影响硬件状态寄存器的值和通过管线流动的其他状态位。相比较 RTL 模拟器， 行为级模拟器更加抽象，它可能使用C, C++, SystemC或SystemVerilog等高级语言框架来获取行为级别的有效模型。行为级模拟器被应用于 ASIC （集成电路）开发过程的早期阶段. 它们可帮助设计人员快速建立系统构造，评估方案并验证正交功能。无需依赖硬件，只需要仿真验证就能快速确定不同设计选项对 CPU 性能的影响。值得注意的是，行为级模拟器通常也会在ASIC（集成电路）的拓扑结构变化之后进行大规模系统级的仿真和验证。此时它们在业界被广泛运用，以验证硬件设计的正确性，并判定如何才能发现可能的工程问题。行为级模拟器是在系统设计、集成电路开发与测试中广泛应用的工具。其抽象模型将硬件描述以可读形式呈现，并提供了一种快速有效的方法来验证CPU性能、功能和互操作性。在日益增长的IC设计领域中，行为级仿真已成为设计开发及验证过程不可缺少的环节。

4.1.2 模拟器系统设计

本文所设计的模拟器是行为级模拟器，主要通过模拟一组同态运算操作在经过从片外高带宽内存读取数据到片上输入buffer，同时计算核从片上输入buffer读取数据就行预算，然后计算核输出运算结果到输入buffer留重用或者直接输出到输出buffer留返回到片外高带宽存储器中的这一过程来获得加速器架构的参数与加速器的性能指标的关系。

模拟器的实现开发环境如表4.1所示：

|  |
| --- |
| 开发环境配置信息 |
| 操作系统：Windows10  版本控制与维护：Git分布式版本控制系统 |
| 开发语言及版本： python3.7 |
| 编译环境：Pycharm Community Edition 2021 |
| 所用到的第三方库：pandas、csv、matplotlib、tqdm、argparse等等 |

表4.1 模拟器设计开发环境

软件系统框图是一种表示整个软件系统的高层次设计视图，它将整个软件系统以及它们之间的关系、接口和交互等抽象概念以一种可视化的方式呈现出来。通常，软件系统框图以块状图或流程图的形式展现，并封装了一个完整的子系统，指导开发人员进行程序结构设计。如图4.1所示，是本文模拟器的软件系统框图，一部分是左半部分的下一章将进行详细介绍的用于实际实现和分析参数的实验部分，另一部分是模拟器的主体部分，包括用户接口部分、用户配置解析与数据结构定义部分、模拟器的总执行调度部分以及最重要的内存仿真和计算仿真部分。其中内存仿真又包括片外高带宽内存的读和写仿真、片上输入多buffer的管理、输入多buffer中的单buffer的管理以及输出缓存FIFO缓存的维护。计算仿真的内容主要是根据要执行的同态操作生成对应的输入输出数据流以及该操作所需的计算延迟用来对作为内存仿真的输入。接下来将对模拟器的五部分的各自的功能以及接口定义进行描述，具体的细节以及实现将放到下一小节。然后再对如何利用这样一个架构获得加速器的性能进行描述。

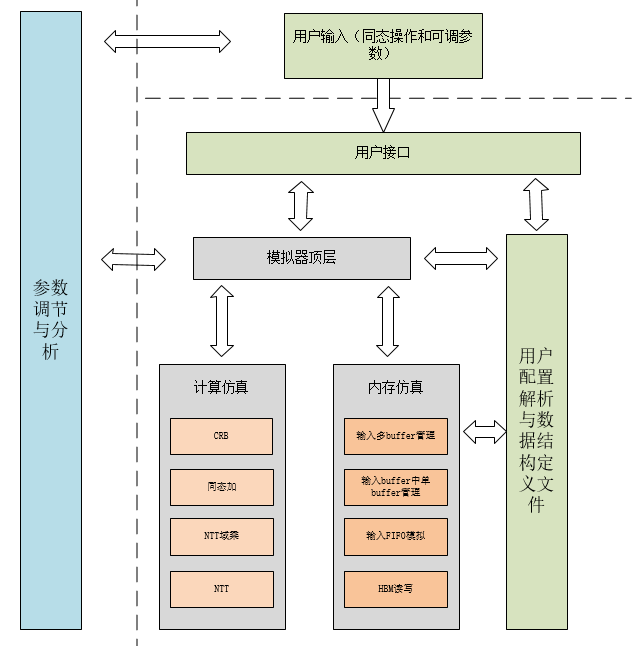


图4.1 模拟器软件系统框图

1. 用户接口

用户接口是软件系统和用户之间的一个交互层，它决定着用户如何与软件进行交互并操作软件。用户接口的作用在于让用户能够方便地使用软件功能，以满足其需求和完成任务。用户接口还可以增强软件的易用性、可理解性和用户满意度。

图4.2展示了模拟器的输入数据和输出数据，其中输入数据包括加速器的参数配置和加速器输入同态操作流。模拟器通过输入的加速器参数对加速器进行初始化，然后开始根据输入的同态操作数据流顺序执行同态操作，先把数据从片外高带宽内存中取到片上输入缓存汇总，然后计算核开始从片上输入缓存中读取数据然后执行对应操作，并在执行过程中跟踪记录输入缓存中的数据以及重用队列中的数据以及输出缓存的数据信息以便用来分析从片上缓存中读取数据的读取延迟以及将计算后的数据写回的写回延迟。同时在记录执行总周期时对不同成分的延迟进行分类统计，具体有哪几类以及如何计算将在后面描述。最后将内存中的数据的跟踪信息、各个操作的执行周期数以及总执行周期数的组成输出为三个结果文件。

其中参数信息包括同态加密数据的密文的长度、同态加密密文数据的残差多项式的长度、CKKS算法所支持的最大乘法深度、算法的安全系数、加速器的片上输出缓存大小、加速器的片上输入缓存的总大小、加速器的片上输入缓存的子buffer个数、加速器运行前从片外高带宽内存中预取到输入缓存的密文个数、计算核中各个计算功能单元的计算延迟大小、片外高带宽内存和片上输入缓存之间的输入带宽大小以及输出带宽大小。

输出数据包括三部分内容。第一个是每个同态操作在加速器中所需的执行周期数，也就是图中右边的的加速器性能分析结果1。第二个是加速器执行所有同态操作的所需的总的执行周期数，以及总周期数的组成情况，也就是左边的加速器性能分析结果2。该文件的每一条信息有加速器执行所有同态操作总的周期数、加速器理论最小所需执行周期数、加速器运行前从片外高带宽内存中预取数据到片上输入buffer中所需的预取周期数、计算核执行输入的同态操作流的所有操作的总的计算核执行周期数、计算核从片上输入缓存中读取数据的时候的读取等待周期数、将计算核的输出数据结果输出缓存FIFO写回到片外高带宽内存的写回等待周期数以及所有的输入同态操作都执行玩成之后输出缓存FIFO buffer中剩余的数据写回到片外高带宽内存中所需的最终完成周期数七个部分组成。 第三个输出就是在加速器运行过程中对每一个操作执行完成之后片上输入缓存中的数据、片上重用数据的维护信息以及片上输出缓存中的数据的记录。

 图4.2 模拟器的输入和输出数据

1. 用户配置解析与数据结构定义文件

在对模拟过程中，就像上面用户接口部分介绍的那样，我们需要输入两个输入文件,这里我们把这两个文件叫做用户配置文件，那么当这个用户配置文件输入到模拟器中之后，我们肯定是需要对其进行解析获取所需信息然后对加速器进行初始化的。

就像表4.2所显示的用户配置解析模块的python类的描述那样，该模块由四部分组成，第一个是向模拟器顶层调度模块开放的接口，也就是一些参数和同态操作流。第二个是读取解析参数配置文件功能函数，用来解析参数配置文件。第三个是读取解析同态操作流文件功能函数，用来解析同态操作流文件。第四个是对所有功能模块开放的接口，用于分享一些共有参数（第四部分并没用定义成表4.2类中的方法，而是单独定义为了全局共享函数）。

|  |
| --- |
| Class **Param\_Config** ( builtins.object ) #声明Param\_Config类 |
| #类内数据定义，通过\_\_init\_\_()函数，更多具体数据定义不再描述，更具体信息可以参考附录代码  **\_\_ParamData\_\_**  加速器参数  **\_\_OptDate\_\_**  同态操作流数据 |
| #类内方法声明  **Read\_param\_config\_file**(self, config\_file\_path)  该方法用于读取解析用户参数配置文件  ：config file path ：参数配置文件路径  **Read\_opt\_config\_file**(self, config\_file\_path)  该方法用于读取解析同态操作流文件  ：config file path：操作流文件路径 |

表4.2 用户配置解析模块类声明

在配置文件的输入形式上，除了直接定义一个文件以外，我们也支持在运行主程序时通过指定参数的形式进行参数配置信息的指定。除了参数配置信息的读取和同态操作流的读取之外，用户配置文件解析对加速器模拟的模拟器顶层调度开放一些接口，以便模拟器顶层模块可以方便的获取想知道的加速器信息以及同态操作流。

具体如何从配置文件和同态操作流文件中获取信息以及对模拟器顶层模块开放哪些信息都将在下一小节模拟器的具体实现中进行详细介绍。

1. 模拟器顶层调度

这个部分起到整体协调作用，向上可以从用户配置解析模块获得想要的用户配置信息以及同态操作流信息，向下可以综合调用计算仿真模块和内存仿真模块。主要流程就是先从用户配置解析文件获得配置信息进行初始化然后依次循环执行同态操作流中的操作，先根据操作类别从计算核模块获取输入数据流信息和输出数据流信息，然后根据输入数据流信息调用内存仿真模块模拟从片上输入buffer中获取该操作对应的数据流，最后再根据输出数据流信息调用内存仿真模块模拟将输出数据通过输出buffer写回片外高带宽内存，如果该输出数据需要重用，则同时写回到片上输入buffer中留到接下来重用。

如表4.3所示

|  |
| --- |
| Class **Simulator**() #声明Simulator类，这个模块用于总体调度 |
| #类内数据定义，通过\_\_init\_\_()函数，更多具体数据定义不再描述，更具体信息可以参考附录代码  **\_\_UpData\_\_**  从上层获取的参数  **\_\_Cycles\_\_**  各类型周期数定义 |
| #类内方法声明  **ArgumentParser** ( )  用于获取并解析命令行中指定参数  **GetInformation** ( Upclass )  获得用户配置文件解析模块中需要传递下来的数据  ：Upclass ：在本模块中定义的用户配置模块的类  **Run** ( )  整体运行框架，根据得到的操作流依次调用执行函数执行，然后更新输入输出buffer，并获得所需周期以及内存跟踪结果。  **Run\_Opt** () |

1. 计算仿真

计算仿真模块主要是模拟加速器的计算核的功能单元，通过输入对应操作的类型，可以得到

1. 内存仿真

参考文献：

[5]An Energy-Efficient Implementation of Fully Homomorphic Encryption for Resource-Constrained Networks

[7] 《DRAMSim2: A Cycle Accurate Memory System Simulator》

[8] 《The gem5 Simulator》

[9] Survey on Homomorphic Encryption and Address of New Trend

[10] A Survey on Homomorphic Encryption Schemes: Theory and Implementation