周报-向嘉豪 (2025 年 2 月 24 日)

摘要：本周完成了 SHA256 算法的 CUDA 并行实现与性能分析实验，实现了二种不同的 GPU 并行化方案。在 RTX 4090 平台上，最优配置 (Grid=128, Block=256) 达到了 478,324 MB/s 的吞吐量，相比 CPU 实现获得 2076 倍加速比。通过详细的性能分析，发现消息大小与线程配置对计算性能有显著影响，并确定了最佳运行参数。同时完成了论文前置知识章节的撰写，主要包括 SPHINCS$^+$ 签名方案的结构组成和 GPU 计算模型的基础架构介绍。

下周计划：1) 将动态线程配置策略从 HASH 函数扩展到整个 SPHINCS$^+$ 签名过程，以进一步提升系统性能；2) 探索 GPU 多流处理技术，提高 GPU 计算资源的利用率。

1 SHA256 实验

本实验实现了使用 CUDA 进行 SHA256 哈希计算的并行加速。实验采用了三种不同的实现方式：批处理、数据并行处理和多流处理。实验在配备 RTX 4090 GPU（128 个 SM，16,384 个 CUDA 核心）的环境下进行。

1.1 最大吞吐量总结

实验测试了不同配置下 SHA256 哈希计算的性能表现，包括 CPU 单核、GPU 单核、GPU 并行以及 GPU 多流处理等多种实现方式。其中 GPU 并行和多流实现尝试了不同的线程配置组合，以下是各种实现方式的最佳性能数据：

表 1: 不同实现方式的最大吞吐量对比

实现方式	最大吞吐量 (MB/s)	消息大小 (B)	相对 CPU 加速比
CPU 单核 [WDC+25]	230.40	131,072	1×
GPU 单核 [WDC+25]	25.39	16,384	0.11×
GPU 并行 (128*256)	478,324.05	1,024	2,076×
GPU 多流 (128*1024*8)	22,923.71	16,384	99×

从最大吞吐量的对比可以看出，GPU 单核性能反而低于 CPU 单核，这是由于 GPU 单线程执行效率较低，且存在额外的数据传输开销。在所有配置中，Grid 大小为 128、Block 大小为 256 的 GPU 并行实现获得了最佳性能，在处理 1024B 消息时达到了 478,324.05 MB/s 的吞吐量，是CPU 性能的 2076 倍。

1.2 性能分析

1.2.1 GPU 并行实现性能对比

不同线程配置下的 GPU 并行实现测试结果如下：

表 2: 不同线程配置的 GPU 并行实现性能对比

线程配置	最大吞吐量 (MB/s)	消息大小 (B)	相对 CPU 加速比
GPU 并行 (82*512)	320,969.49	2,048	1,393×
GPU 并行 (128*256)	478,324.05	1,024	2,076×
GPU 并行 (128*1024)	316,416.92	512	1,374×
GPU 并行 (256*1024)	349,549.00	512	1,517×

从性能测试结果分析可以得出以下结论：128*256 的配置获得了最佳性能，在处理 1024B 消息时达到了 478,324.05 MB/s 的吞吐量。增加 Block 大小（从 256 到 1024）反而导致性能下降，表明更大的线程块可能引起更多的资源竞争。

网格大小为 128 时性能较优，这与 RTX 4090 的 128 个 SM 完美匹配。所有配置在较小消息大小（512B-2048B）时达到峰值性能，这表明此时计算资源和内存访问达到最佳平衡。相比 CPU 实现，GPU 并行实现获得了 1,374× 到 2,076× 不等的性能提升。

这些结果印证了之前的配置分析，即Grid 大小需要匹配 SM 数量，而 Block 大小需要在资源利用和竞争之间取得平衡。较大的 Block size 虽然理论上可以提供更多的并行性，但实际上可能因为资源争用而降低整体性能。

1.2.2 吞吐量变化分析

根据 GPU 并行 (128*256) 配置的实验数据，我们可以清晰地观察到吞吐量随消息大小变化的三个阶段：

表 3: 消息大小对 SHA256 哈希计算吞吐量的影响

阶段	消息大小范围	吞吐量范围 (MB/s)	典型性能
初始增长期	4B-512B	1,044-352,862	小消息大小下快速提升
峰值性能期	1024B-4096B	457,400-478,324	达到最佳平衡状态
性能下降期	>4096B	425,784-461,580	趋于稳定并小幅波动

初始增长期 (4B-512B)：在这个阶段，吞吐量从 1,044MB/s 快速增长至 352,862MB/s。这种快速增长主要是因为内核启动和内存传输等固定开销被更多数据分摊，同时硬件利用率随数据量增加而提高。**峰值性能期 (1024B-4096B)**：在此阶段达到性能最优，特别是在 1024B 时达到最高吞吐量 478,324MB/s。此时 GPU 内存带宽（1,008 GB/s）被充分利用，计算资源与内存访问达到最佳平衡。**性能下降期（>4096B）**：当消息大小继续增加时，吞吐量在 425,784-461,580MB/s 之间波动。这主要是因为内存带宽成为主要瓶颈，同时缓存失效率增加，大消息处理的内存延迟影响也随之加大。总的来说，相同的并发配置下，单个运算规模的大小（此处为消息长度），对整体性能有显著影响。因此动态的调整线程配置和消息分配具有研究价值。

2 前置知识写作

完成了论文前置知识写作，主要包含两部分：SPHINCS$^+$ 概述和 GPU 计算模型。

SPHINCS$^+$ 是无状态的基于哈希的后量子签名方案，由三个核心组件构成：WOTS+：一次性签名方案，用于认证路径, FORS：少次签名方案，使用 k 个 t 元素的伪随机子集, Hypertree：h 高度的 d 层结构，每层包含 h/d 高度的默克尔树

GPU 计算模型方面：硬件由多个流式多处理器 (SM) 组成，每个 SM 包含多个 CUDA 核心, 采用 SIMT 执行模式，线程组织为 warp 和 block 结构, 包含共享内存、寄存器和缓存等多级存储系统, CUDA 框架提供合并访存等优化策略。

参考文献

[WDC+25] Ziheng Wang, Xiaoshe Dong, Heng Chen, Yan Kang, and Qiang Wang. Cuspx: Efficient gpu implementations of post-quantum signature sphincs⁺. *IEEE Transactions on Computers*, 74(1):15–28, 2025.