GPU Kernel Fusion for Scientific Computing

Mina Rao, UNC Charlotte
Tyler Allen, College of Computing and Informatics

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

Abstract

- Research project aims to optimize computational libraries for large-scale scientific apps using GPUs.
- Complex computational libraries can lose efficiency between routine calls, resulting in suboptimal performance.
- Long-term goal is to merge and unfold library calls to achieve optimization and performance gains.

Background

GPUs (Graphic Processing Unit)

- A type of processor designed to
- handle complicated mathematical computations

What can they do for us?

- GPUs are good for tasks that involve large amounts of data
- The way they are designed allows them to perform operations on multiple data points at once which can increase efficiency (especially compared to CPUs)

Motivations

GPUs have vast capabilities and can further scientific learning in various fields including:

- Machine Learning
- Data Analytics
- Medical Imaging
- Real-Time Graphics
- Finance Modeling
- Cyber-security

The vast potential of GPUs extends far beyond the examples listed above, and their ability to accelerate computations is opening up new avenues for scientific discovery and innovation.

Frontier, the world's fastest supercomputer, has 37,888 GPUs.

Method

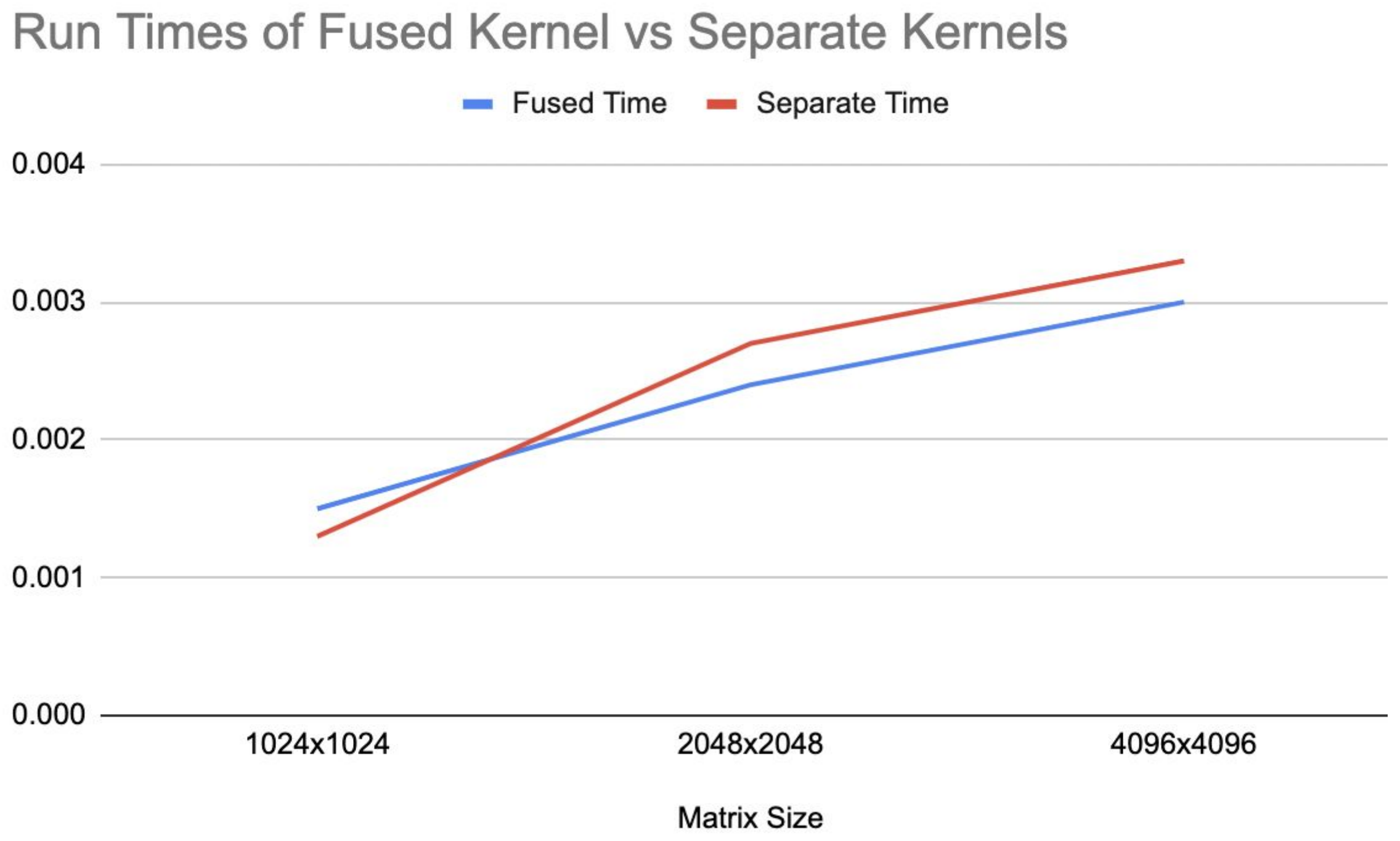
This study focuses on testing this approach using linear algebra:

1. Matrix addition
2. Matrix multiplication
3. Multiplication + addition

$$2 \cdot \begin{bmatrix} 10 & 6 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 2 \cdot 10 & 2 \cdot 6 \\ 2 \cdot 4 & 2 \cdot 3 \end{bmatrix}$$

$$\begin{bmatrix} 4 & 8 \\ 3 & 7 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 5 & 2 \end{bmatrix} = \begin{bmatrix} 4+1 & 8+0 \\ 3+5 & 7+2 \end{bmatrix}$$

After comparing the run times we got the following results:



As the size of the matrix increases, the run times of the fused matrix kernel gets progressively lower than that of the separate kernel.

Conclusions

The main issue is how to automate this idea. If we are able to implement this method on a large scale we could optimize many processes. The benefits of faster GPUs can be significant, including reduced processing time, increased productivity, improved accuracy and efficiency, and the ability to tackle more complex problems and applications. This can lead to faster innovation, better decision-making, and ultimately, improved outcomes for individuals, businesses, and society as a whole.