Lab 04 Coalesced Memory Access

February 2022

1 Introduction

A coalesced memory transaction is one in which all of the threads in a warp access global memory at the same time in chunks of bytes. It basically means that that threads run simultaneously, try to access memory that is nearby.

Sequential threads in a warp access sequential memory that is not aligned will result in more transactions reducing the effective bandwidth.

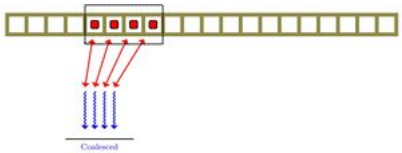

Figure 1: Coalesced memory access source

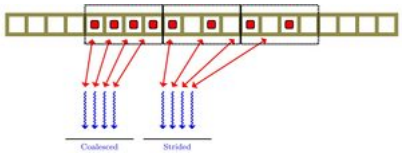

Figure 2: Uncoalesced memory access source

However, if we increase the "stride" of the access between the threads, it will require many more memory accesses.

2 Task

As shown during the lab perform vector addition using structure of arrays (SoA) and Array of Structures(AoS) and measure the following -

1. Warp level instructions for global loads
2. Executed Load/Store Instructions

3. Global Memory Load Efficiency
4. Global Memory Store Efficiency
5. Global Load Transactions
6. Global Store Transactions

The above metrics can be fetched using -

```
nvprof -m all ./<name of binary>
```

Submit the document reporting the values along with the code zipped as ⟨roll_no⟩_⟨name⟩.zip