

# Lab 6: Advanced Topic in CUDA

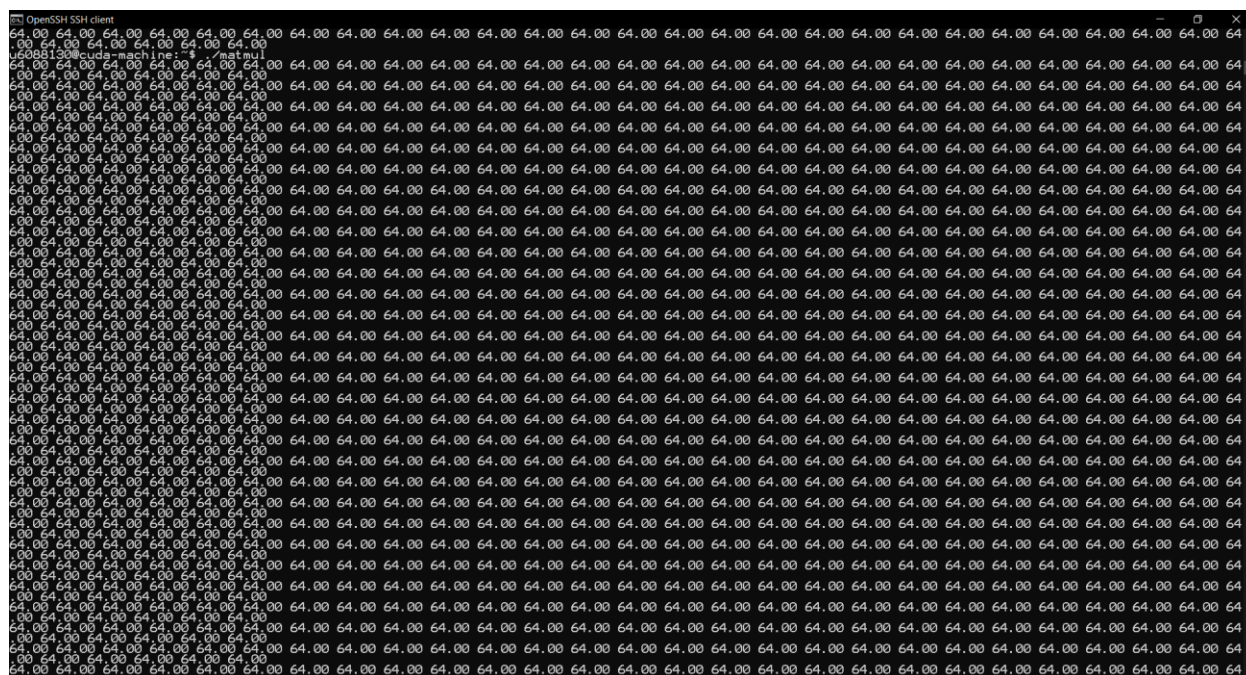
Name: Sunat Praphanwong

ID: 6088130

Sec: 1

Save your file to “lab6\_60xxxxxx.pdf”, ZIP the pdf file and source code files, and upload it to MyCourses website.

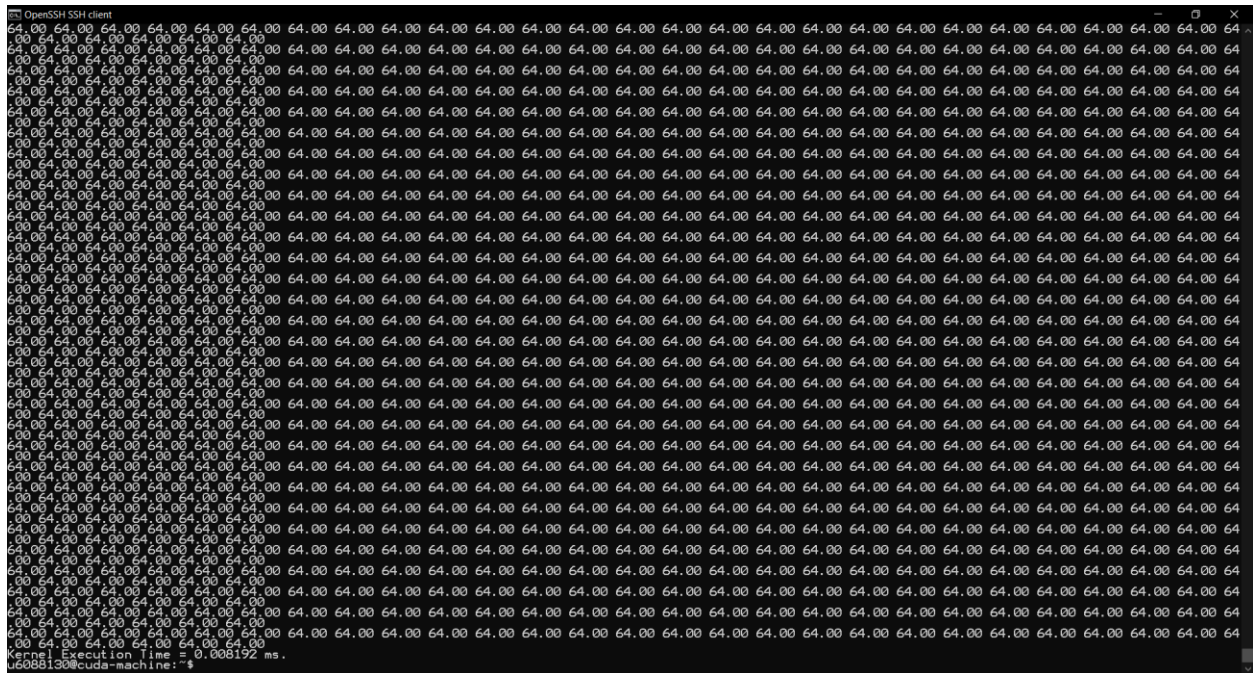
## Lab: 1. Result from “matmul.cu”



The screenshot shows a terminal window titled "OpenSSH client". The prompt is "u0088130@cuda-machine: ~\$". The user has entered the command "matmul". The output of the program is a dense grid of the number "64.00", filling the entire terminal area. The grid is approximately 30 columns wide and 30 rows high, with the numbers "64.00" repeated in a regular pattern.

## Lab: 2. Result from “matmul\_shared.cu”





#### Lab: 4. CUDA Occupancy calculation for “matmul.cu” and “matmul\_shared.cu”

Download CUDA Occupancy Calculator from [https://docs.nvidia.com/cuda/cuda-occupancy-calculator/CUDA\\_Occupancy\\_Calculator.xls](https://docs.nvidia.com/cuda/cuda-occupancy-calculator/CUDA_Occupancy_Calculator.xls)

##### 4.1 “matmul.cu”

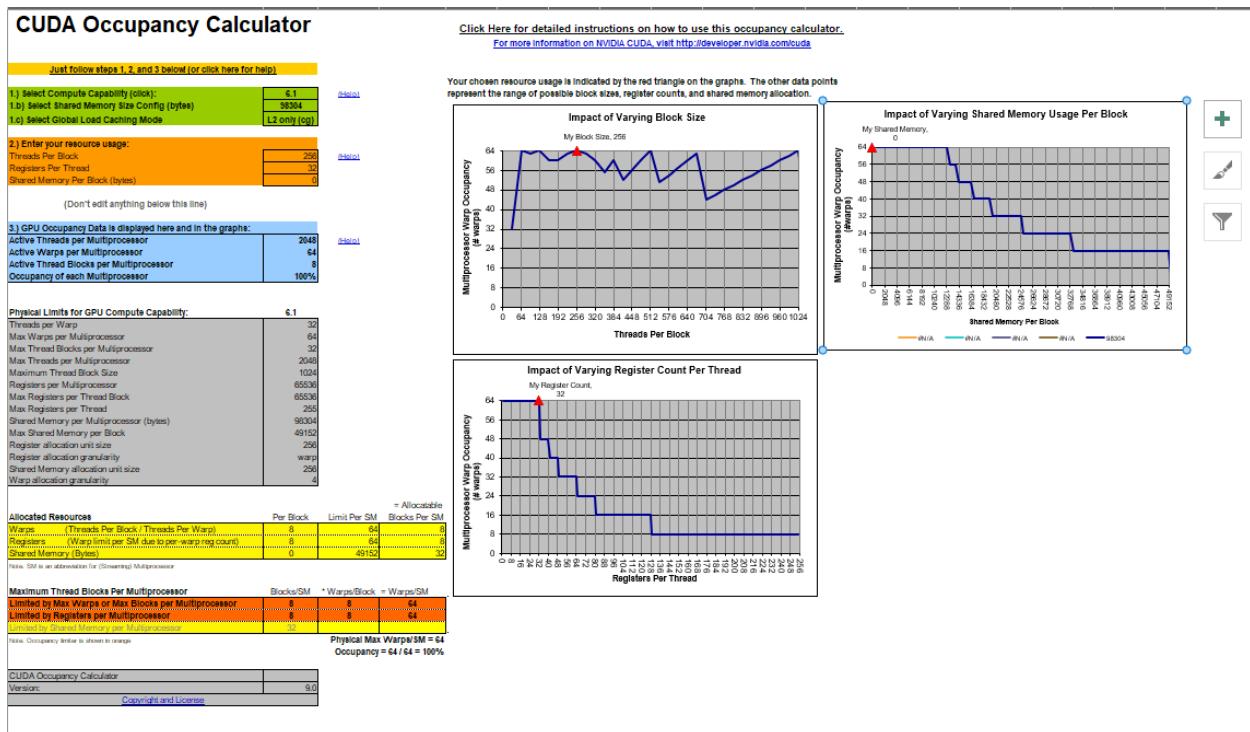
Threads per block = 256

Registers used per thread = 32

Shared memory used per thread block = 0

**Screenshot of the result from NVIDIA Occupancy Calculator:**





## 4.2 “matmul\_shared.cu”

Threads per block = 256

Registers used per thread = 32

Shared memory used per thread block = 0

Screenshot of the result from NVIDIA Occupancy Calculator:

# CUDA Occupancy Calculator

Just follow steps 1, 2, and 3 below (or click here for help)

1.) Select Compute Capability (click): **6.1**  
 1.b) Select Shared Memory Size Config (bytes): **38304**  
 1.c) Select Global Load Caching Mode: **L2 only (opt)**

2.) Enter your resource usage:  
 Threads Per Block: **256**  
 Registers Per Thread: **25**  
 Shared Memory Per Block (bytes): **2048**  
 (Don't edit anything below this line)

3.) GPU Occupancy Data is displayed here and in the graphs:  
 Active Threads per Multiprocessor: **2048**  
 Active Warps per Multiprocessor: **64**  
 Active Thread Blocks per Multiprocessor: **8**  
 Occupancy of each Multiprocessor: **100%**

Physical Limits for GPU Compute Capability: <b>6.1</b>	
Threads per Warp	32
Max Warps per Multiprocessor	64
Max Thread Blocks per Multiprocessor	32
Max Threads per Multiprocessor	2048
Maximum Thread Block Size	1024
Registers per Multiprocessor	65536
Max Registers per Thread Block	65536
Max Registers per Thread	256
Shared Memory per Multiprocessor (bytes)	98304
Max Shared Memory per Block	49152
Register allocation unit size	256
Register allocation granularity	warp
Shared Memory allocation unit size	256
Warp allocation granularity	4

Allocated Resources	Per Block	Limit Per SM	* Allocatable Blocks Per SM
Warp (Threads Per Block / Threads Per Warp)	8	64	8
Registers (Warp limit per SM due to per-warp reg count)	8	64	8
Shared Memory (Bytes)	2048	49152	24

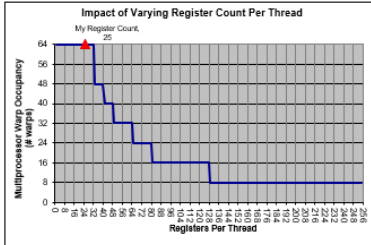
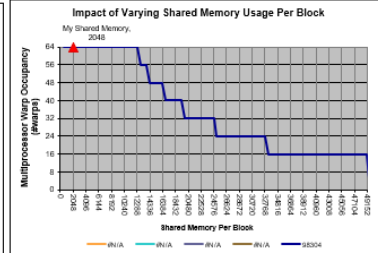
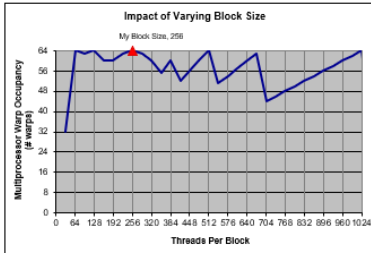
Maximum Thread Blocks Per Multiprocessor	Blocks/SM	* Warps/Block = Warps/SM
Limited by Max Warps or Max Blocks per Multiprocessor	8	8
Limited by Registers per Multiprocessor	8	8
Limited by Shared Memory per Multiprocessor	24	24

From: Occupancy limit is shown in orange. Physical Max Warps/SM = 64  
 Occupancy = 64 / 64 = 100%

CUDA Occupancy Calculator  
 Version: **9.0**  
[Copyright and License](#)

[Click Here for detailed instructions on how to use this occupancy calculator.](#)  
[For more information on NVIDIA CUDA, visit <http://developer.nvidia.com/cuda>](#)

Your chosen resource usage is indicated by the red triangle on the graphs. The other data points represent the range of possible block sizes, register counts, and shared memory allocation.



**Do not forget to include the source files into the zip file before submission.**