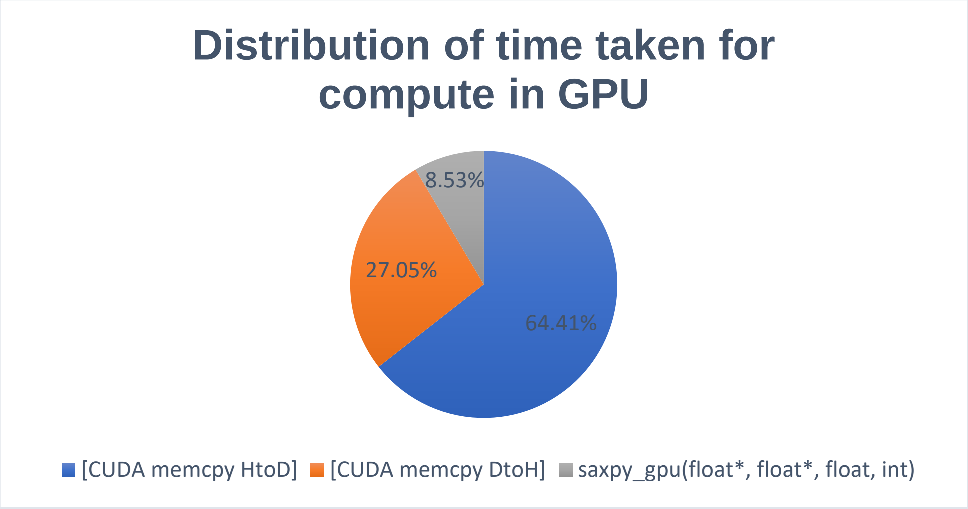
SHREYA GAUR

ECE60827: PROGRAMMABLE ACCELERATOR ARCHITECTURE

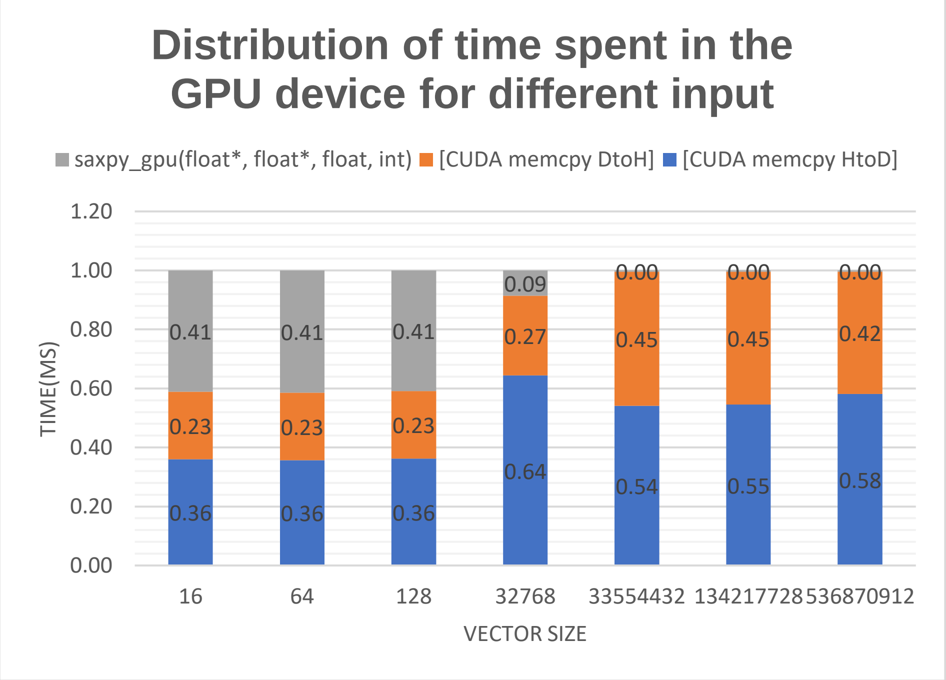
CUDA PROGRAMMING ASSIGNMENT

SAXPY

For vectorSize equal to 326768, the pie chart below shows the distribution of time spent on the GPU for different kernels and transfer between device and host memory. The maximum time is spent in the transfer of data to the GPU device memory. This is understandable since both matrices need to be transferred from host to device.

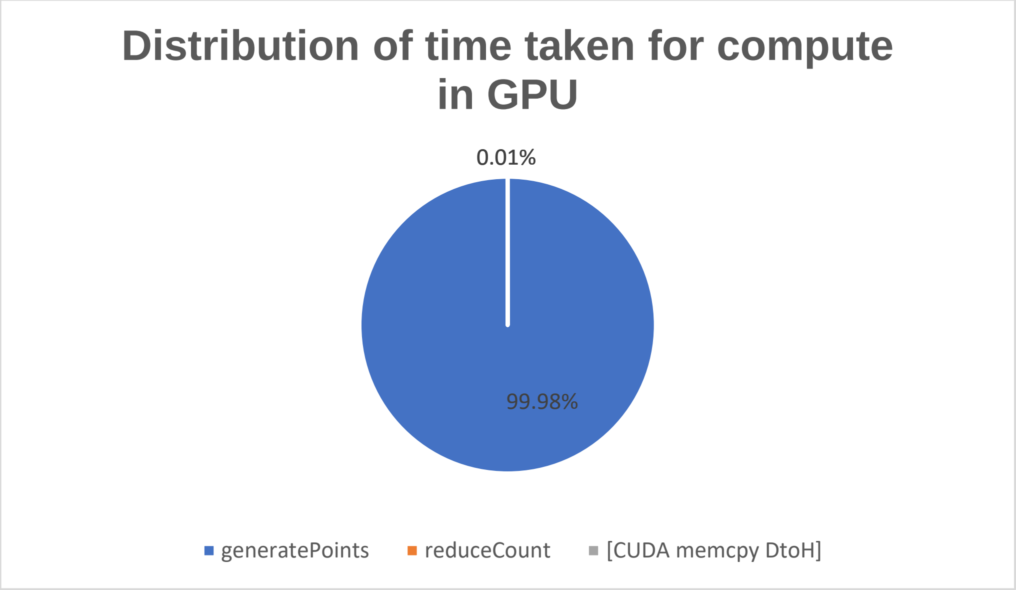


Following stacked bar graph displays the change in distribution of time spent communicating with the GPU Device for different vectorSize values. It can be observed that as the size of matrices increase, the time spent on computation shrinks tremendously as compared to the time spent in transfer.

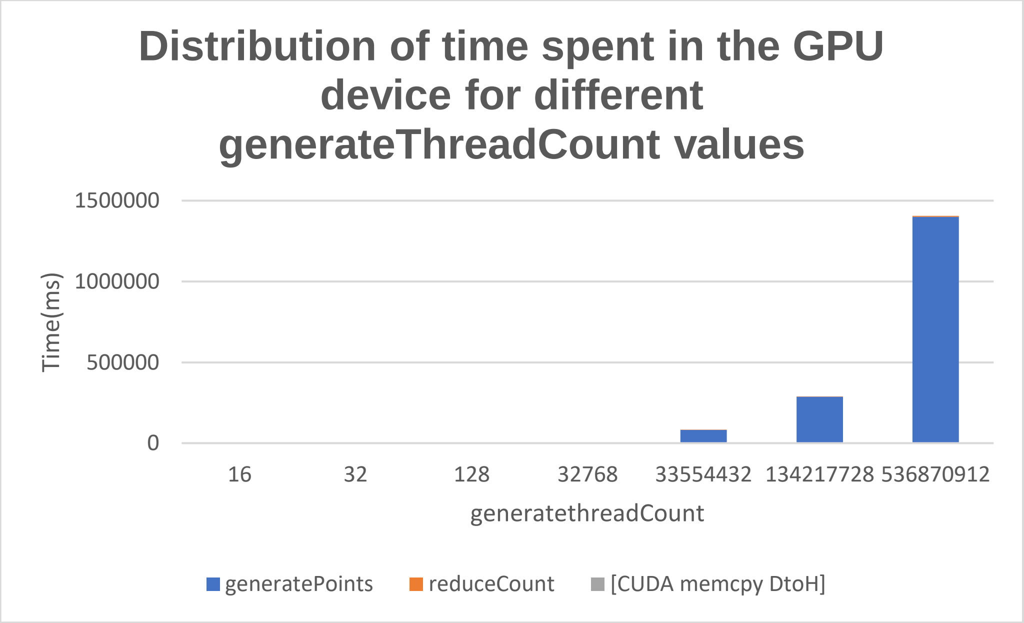


MONTE CARLO ESTIMATION OF PI

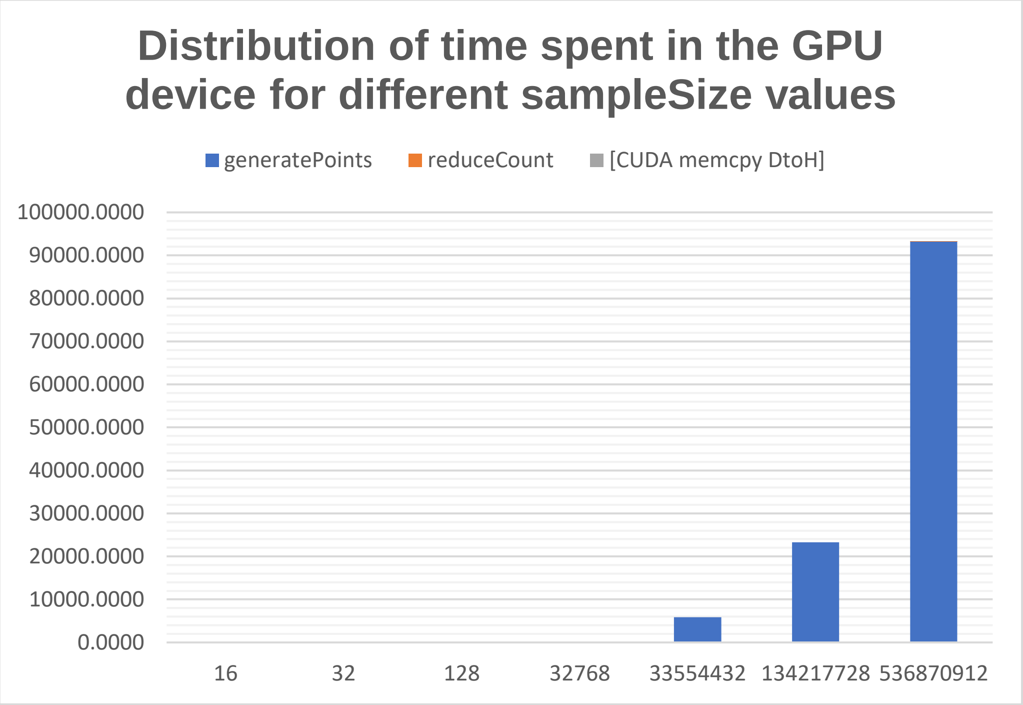
For generateThreadCount = 1024, sampleSize =1000000, reduceThreadCount = 32 and reduceSize = 32, the pie chart of distribution of time taken in GPU is shown below. The maximum time is spent in the generatePoints kernels as there is no data transfer from host to device.



The following graph shows the time distribution as the generateThreadCount is varied while keeping the other arguments as constant. The temporal behaviour of the GPU is the same. Most of the time is spent in the generatePoints function. The function reduceCount forms very small part of the computation time.



The following graph shows the time distribution as the sampleSize is varied while keeping the other arguments as constant. The temporal behaviour of the GPU is the same. Most of the time is spent in the generatePoints function. The function reduceCount forms very small part of the computation time.


Distribution of time spent in the GPU device for different sampleSize values
generatePoints
reduceCount
[CUDA memcpy DtoH]