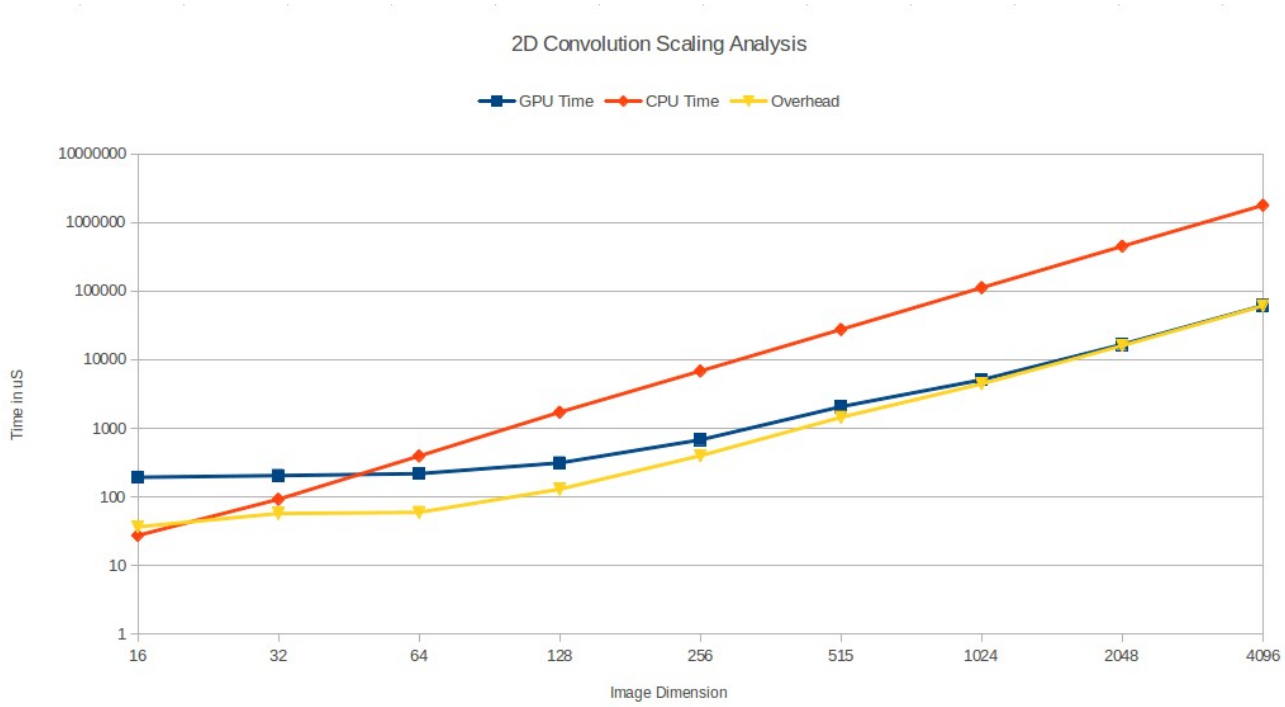
Homework 6
Omkar Deshmukh

Convolution:

- Following plot shows scaling analysis of 2D convolution.
- The time reported for GPU is inclusive time.
- Overhead was calculated from summing up cputime reported in CUDA_PROFILE_LOG for all the corresponding H2D and D2H memcopies. It scales linearly with input on log scale as shown below.

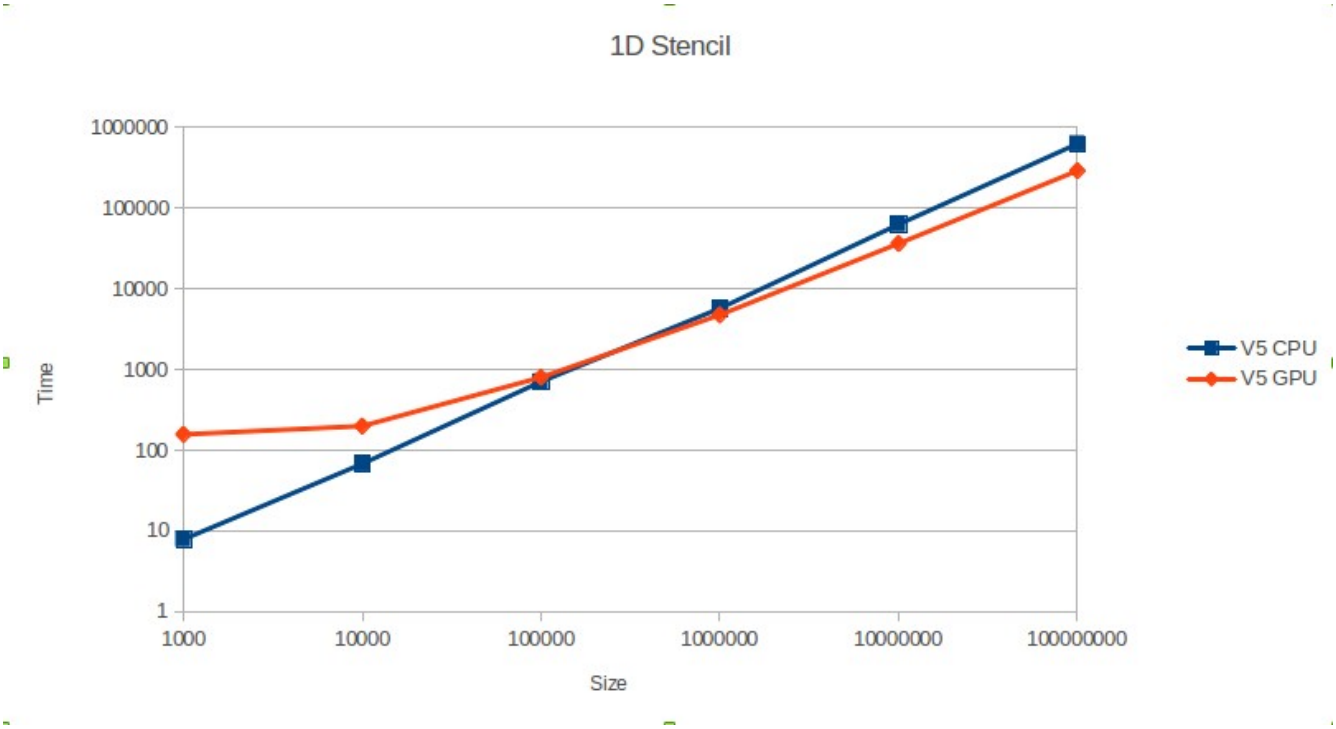


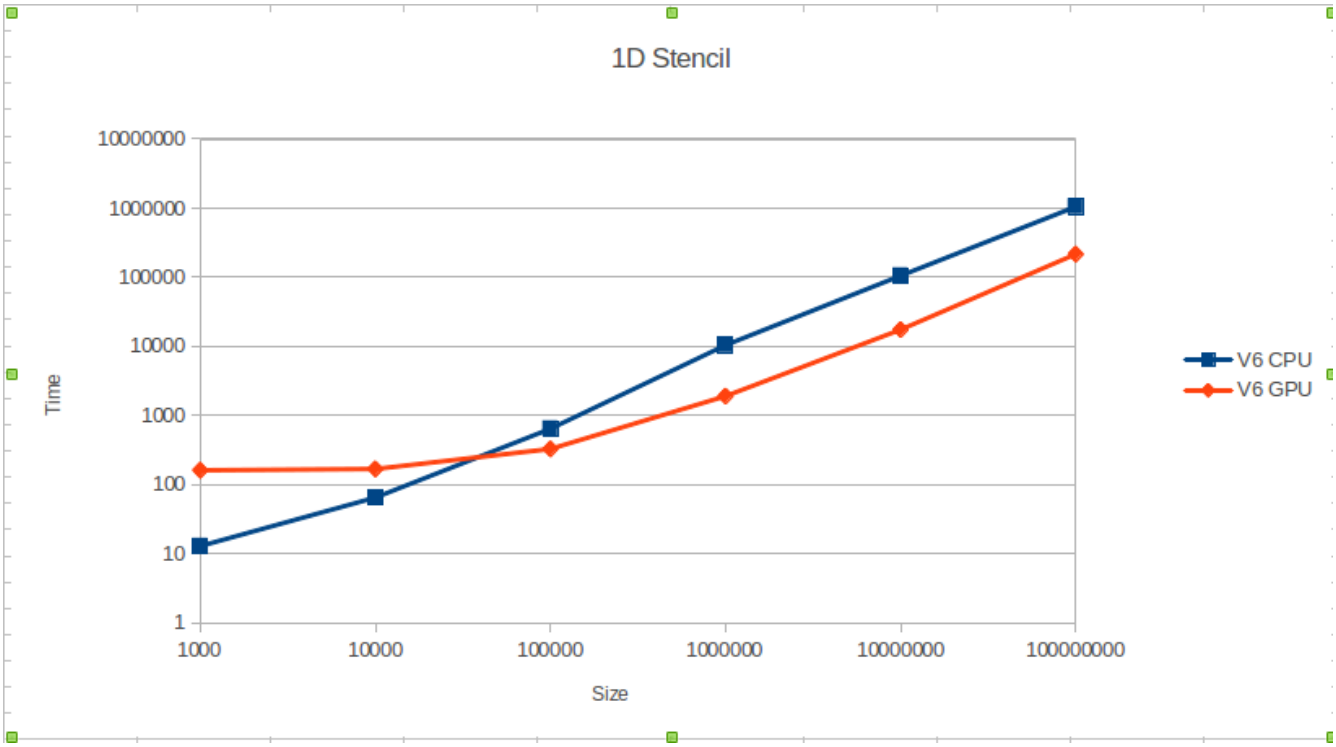
Notes:

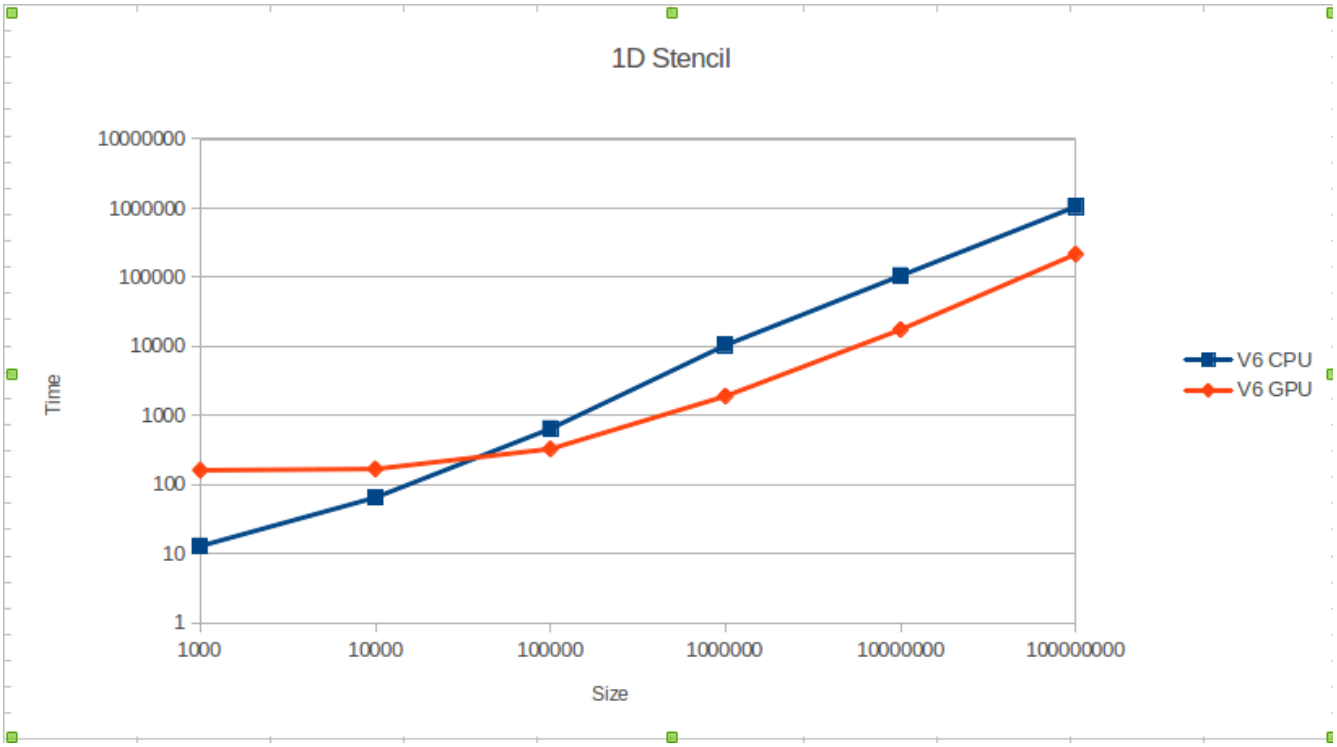
- Please note that the **log** file was generated by running **run.sh** script.
- The program was modified to take **Image Dimension** as input **argv[1].**
- Unlike earlier homework, this program can only convolute square images whose dimensions are power of 2. To remove this restriction, we can pad the matrices/images before passing them to kernel. This was done in earlier homework but not here since analysis did not ask for non square images.

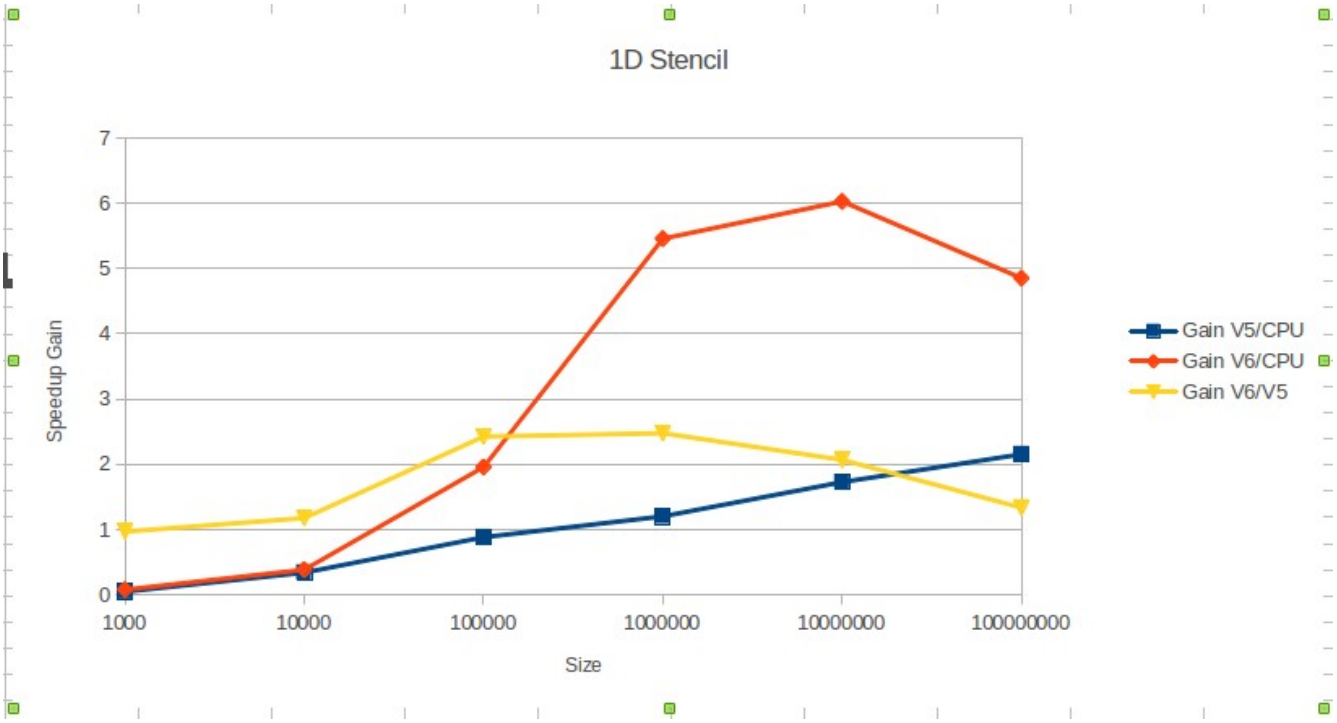
Stencil:

- There are in total 4 plots given below.
- As we can see from the gain plot, V6 with pinned memory offers better GPU performance over V5. This is because pinning the pages to host prevents them from swapping and thus memcopies are faster. Even when compared to CPU, V6 gain is better than V5.
- The data was collected by running the script **run.sh.**
- The program was modified to take **Size** as input **argv[1].**


1D Stencil
1000000
100000
10000
1000
100
10
1
Time
1000
10000
100000
1000000
10000000
100000000
Size
V5 CPU
V5 GPU


1D Stencil
10000000
1000000
100000
10000
1000
100
10
1
Time
1000
10000
100000
1000000
10000000
100000000
Size
V6 CPU
V6 GPU


1D Stencil
10000000
1000000
100000
10000
1000
100
10
1
Time
1000
10000
100000
1000000
10000000
100000000
Size
V6 CPU
V6 GPU


1D Stencil
7
6
5
4
3
2
1
0
Speedup Gain
1000
10000
100000
1000000
10000000
100000000
Size
Gain V5/CPU
Gain V6/CPU
Gain V6/V5