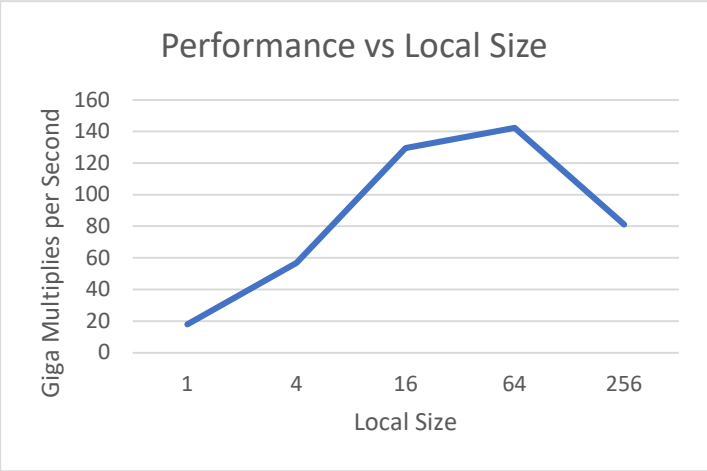
Matrix Multiplication Performance Analysis

Jack Hart | hartjack@oregonstate.edu

I ran this program on the DGX system using the Final-CUDA partition. The graphs below demonstrate the relationships between performance and our independent variables.



With matrix size, it appears as though the performance increases steadily with more data and then plateaus right near the end. This is likely due to it approaching the maximum performance you can gain from this parallelism.

As for the local size, it looks similar to a bell curve with the apex being around 64 work-items. The increase in performance in the beginning is because we are allowing the computer more threads, giving it more room for parallelism. However, near the end, it appears as though we reach the maximum amount of work-items at 64. This is likely the number of work-items each work-group has, thus when we exceed that number, it forces the GPU to synchronize that work-group with other work-groups which will add extra computation and hence, extra compute time.