Assignment 3 ID5130

Anurag Manoj ee22b003

May 2024

1 Q1

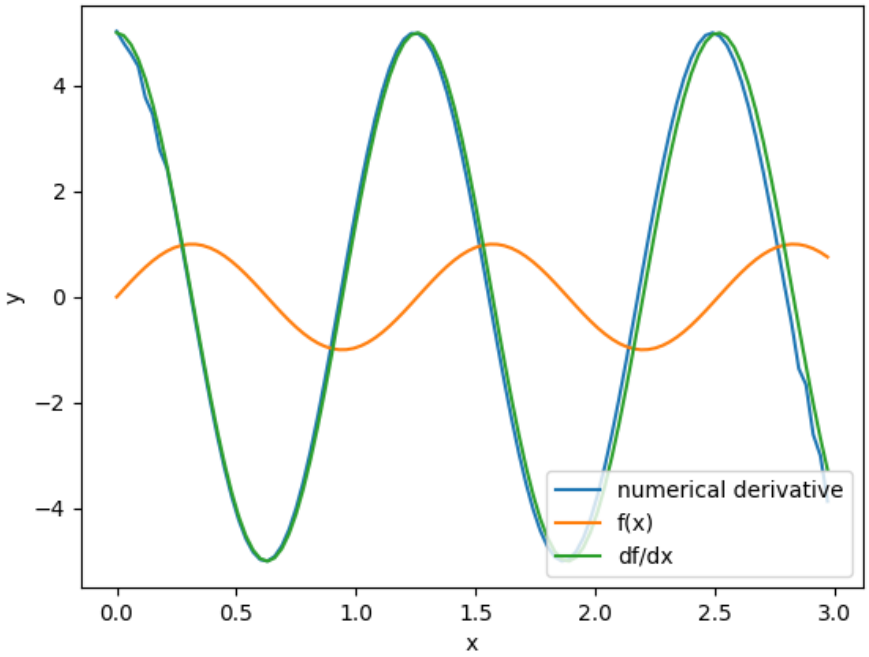

Figure 1: Numerical vs analytical solution for N=100

The code has been optimized to reduce unnecessary data transfers by initializing all memory that is not the output values, that is, the matrix of LU decomposition inside the GPU itself. The code also runs fastest for n=1000 with 100 gangs.

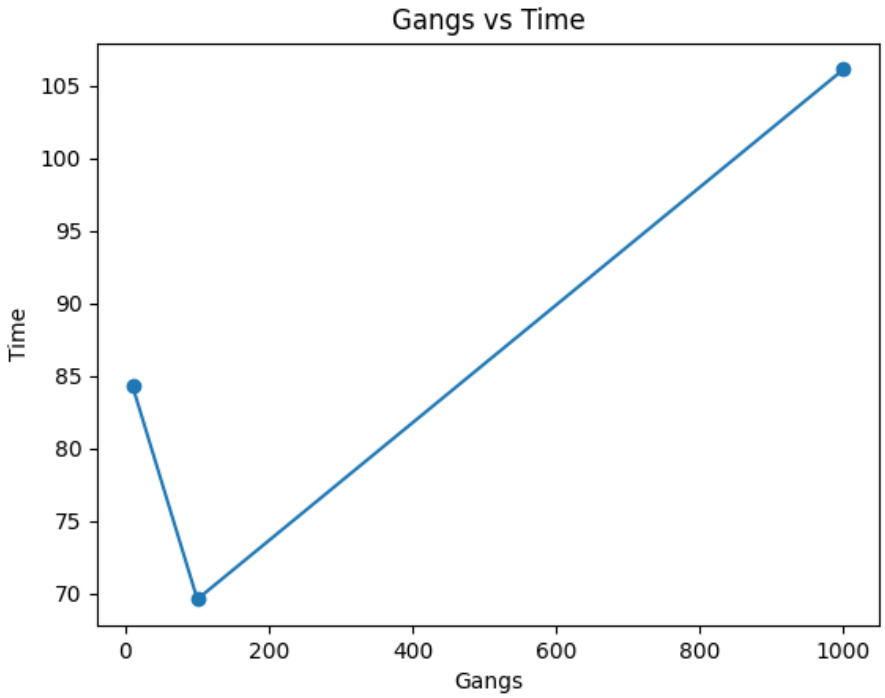

Figure 2: timing of code

2 Q2

2.1 Algorithm

Let the matrix L be the cholesky decomposition of A. If the values of the matrix "L" is known on all columns before the jth column, then all the values on the jth column can be calculated as independent dot products of pairs of rows, of length = j. Taking the dot product of 2 arrays of length j with an acc vector length of v has a time complexity of O(Nlog(v)/v) using the reduction operator, and this can be done parallely over all N-j rows left, in a time O((N-j)/g), where g is the number of gangs, and this has to be repeated N times, that is over all columns. The net time complexity is $O(N^3 \cdot \log(\frac{v}{vg}))$. A smaller v and a larger g is better. v = 32 is optimum based on the warp size of NVIDIA GPU's.

2.2 Speedup

Data transfers have been minimized using a single data region for the cholesky decomposition, N/2 is found to be an estimate for the optimum number of gangs, and 32 as the optimum number of threads. Speedup is shown to increase with N.

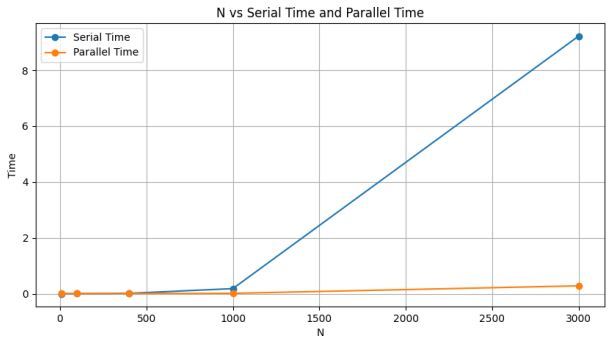

Figure 3: timing of code

2.3 Verifying code

The Cholesky decomposition is shown in the image for N=10, the image compares serial vs parallel solution. There is a descrepancy between the two, and as expected, the discrepancy increases in later columns, this is due to numerical errors accumulating.

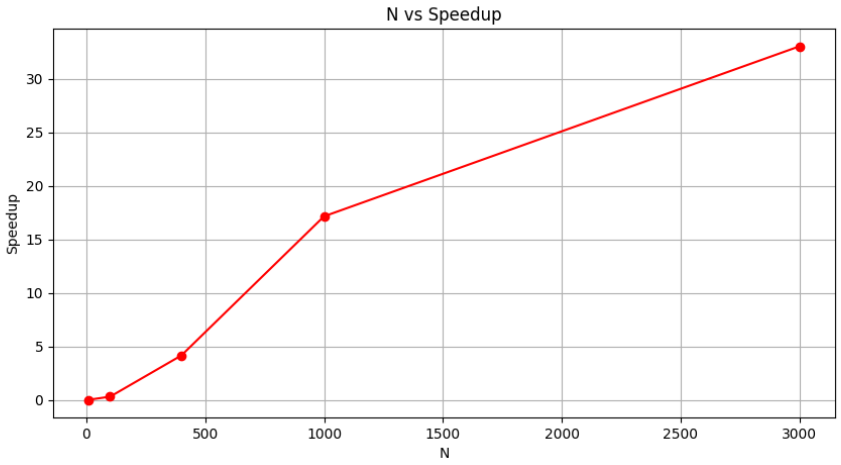

Figure 4: timing of code

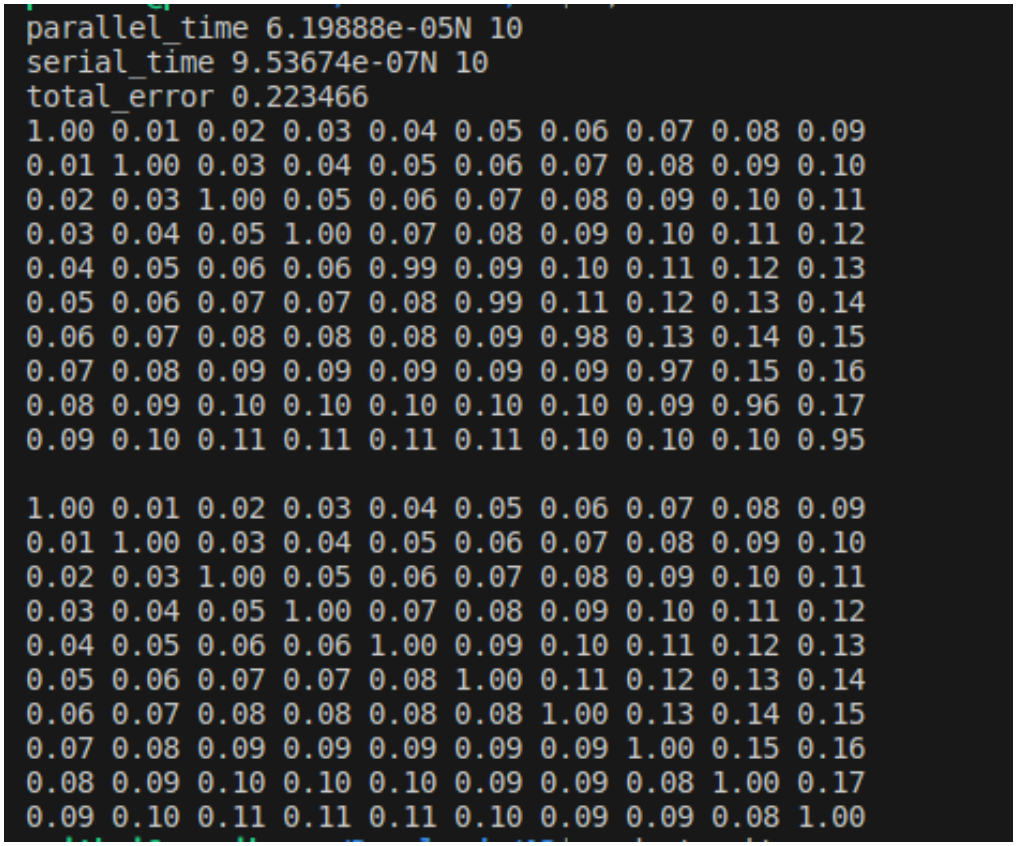

Figure 5: serial(above) vs parallel(below) cholesky decomposition, N=10