Homework 5

- Starting from the provided code, try to optimize it as much as possible
  - Use MPI_Wtime to take time measurements
  - Analyze the performance of your application by drawing stack graphs to show how much time is spent on each stage (computation, moving data between nodes, moving data back and forth to and from the GPU, …)
  - Return the source code, the script to analyze the performance output, and a document describing your solution and a short analysis of the performance (mainly to show you understand how you obtained it).

Example not real data
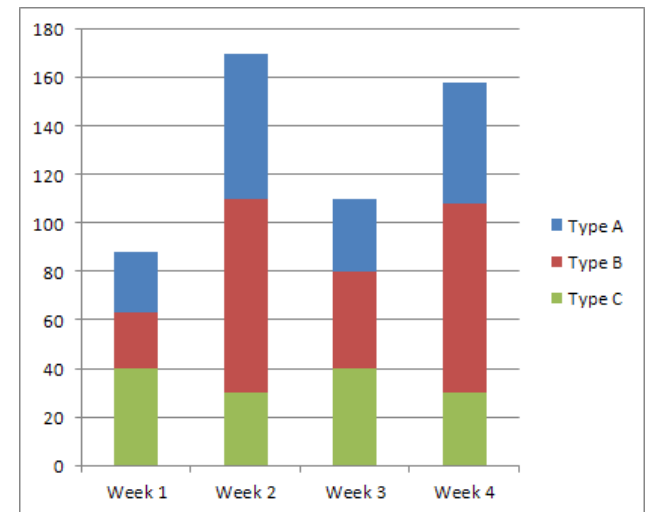
Possible code optimizations

- CUDA:
  - As the computation is symmetrical and highly balanced, one can have a different work distribution and do more computations per thread
  - Use shared memory
  - Divide the computations in 2 parts: what needs external data and what doesn't.
- MPI:
  - Use datatypes
  - Use RMA
- Overlap communication and computations
  - Create a specialized kernel to pack and unpack all the borders in one operation
  - As starting a kernel has a high latency merge this pack/unpack kernel with the updates based on the ghost regions