1 Problem 3

1.1 Introduction to problem

For a parallel application with input and output of size n. The computation can be divided between CPU and GPU. Where the ratio of the size computed using the GPU is b. Which is a value between 0 and 1.The time necessary to compute the different parts of the input for CPU and GPU will amount to:

$$t_{GPU} = n_{GPU} R_{GPU} \tag{1}$$

$$t_{CPU} = n_{CPU} R_{CPU} \tag{2}$$

As we divide the input into two parts for CPU and GPU computation, n_{GPU} and n_{CPU} can be set as part of the total input size n:

$$n_{GPU} = bn \tag{3}$$

$$n_{CPU} = (1 - b)n \tag{4}$$

1.2 Finding expression for t_{CPU} and $t_{total_{GPU}}$

Applying (3) and (4) in equations (1) and (2), the amount of time to perform the GPU and CPU computation will then be:

$$t_{GPU} = bnR_{GPU} \tag{5}$$

$$t_{CPU} = (1 - b)nR_{CPU} \tag{6}$$

When using GPU we are required to transfer the data input from the CPU to the GPU. The transfer cost in time is dependent on the bandwidth B and will for a size m amount to:

$$t_{transfer} = mB \tag{7}$$

The total execution time on GPU will be the sum of computation cost and transfer cost:

$$\begin{aligned} t_{total_{GPU}} &= t_{GPU} + t_{transfer} = bnR_{GPU} + bnB \\ &= bn(R_{GPU} + B) \end{aligned} \tag{8}$$

1.3 Finding expression for b to divide work evenly

The GPU are generally faster than the CPU. A evenly distributed system using both CPU and GPU should divide the work between them so they finish simultaneously. This can be achieved by finding an expression for b which makes time to perfrom CPU and GPU calculation equal:

$$t_{CPU} = t_{total_{GPU}} \tag{9}$$

$$(1 - b)nR_{CPU} = bn(R_{GPU} + B)$$

$$b = \frac{1}{1 + \frac{R_{GPU}+B}{R_{CPU}}} \tag{10}$$

1.4 Will it ever be beneficial to not use the GPU at all?

Equation (10) describes the fraction of the input to be computed using the GPU. In this section we will determine if it will ever be meaningful to set $b = 0$, and effectively not using the GPU. Looking at equation (10) we se that there are three parameters which affects the ratio b. Setting $b = 0$ would suggest that the denominator would approach infinity. The only part of the denominator of eq. (10) that can change is the relation between R_{GPU},B and R_{CPU}:

$$\frac{R_{GPU} + B}{R_{CPU}} \Rightarrow \infty$$

$$\frac{R_{GPU}}{R_{CPU}} + \frac{B}{R_{CPU}} \Rightarrow \infty \tag{11}$$

We have already established that the GPU is faster than CPU for all general and relevant cases for this calculations. Which gives:

$$\frac{R_{GPU}}{R_{CPU}} < 1 \tag{12}$$

Using equation (12) in (11) we get:

$$\frac{B}{R_{CPU}} \Rightarrow \infty \tag{13}$$

It will be benefical to not use the GPU and limit our computations to only use the CPU if the bandwidth between the GPU and CPU is much larger than the computation time on the CPU. Meaning if transferring data to the GPU costs much more than computing the same data on the CPU, the GPU should not be used.