COMP 4901Q: High Performance Computing (HPC)

Homework 2

Due date: ~~20~~ **24 April 2022 23:59**

Consider a machine learning (ML) problem that can be solved by the stochastic gradient descent (SGD) algorithm. The ML problem is described as follows: we need to find a set of parameters (denoted by w, which is a *d*-dimension vector, i.e., $w \in \mathbb{R}^d$) such that the objective function $f: \mathbb{R}^d \to \mathbb{R}$ is minimized, i.e.,

$$\text{minimize } f(w, D),$$

where D is a data set which consists of *n* samples. The SGD algorithm solves the above objective function by the following steps:

1. Initialize the value of w as w_0.
2. Randomly select *m* samples (denoted as $D_1, D_2, \ldots, D_m$) from D.
3. Calculate the gradient of the objective function using
$$\nabla w = \sum_{i=1}^{m} g(w, D_i),$$
where $g: \mathbb{R}^d \to \mathbb{R}^d$ is a function to calculate the gradient with one data sample.
4. Update the parameter using: $w = w - \alpha \cdot \nabla w$, where α is a scalar.
5. Goes to step 2 till some conditions are satisfied.

Step 2-4 is also called one iteration, and SGD typically takes a large number of iterations to find the optimal solution. We use *N* to denote the number of iterations that the algorithm takes.

In a traditional serial computer, each step at every iteration should be sequentially executed. We use t_1, t_2, t_3, and t_4 to denote the time used in step 1, 2, 3, and 4 respectively. Then, in a serial computer with one processor without any parallelism, the SGD algorithm takes

$$T_{serial} = t_1 + N(t_2 + t_3 + t_4)$$

to find the optimal solution.

Given a *P*-processor cluster, which consists of P_1 servers and each server has P_2 processors, you are required to write a report to describe a possible parallel solution to accelerate the above SGD algorithm. You should include the following key components in your parallel solution:

1. Which parts can be parallelized, and which parts cannot be parallelized?
2. Clearly describe how to parallelize the algorithm.
3. Describe what are the parallel programming models of your proposed parallel solutions.
4. Write out the time equation of your parallel solution and compare to T_{serial}. Note that you can introduce any extra notations to represent the time.
5. Determine whether your parallel solution is strong-scaling or weak-scaling, and plot the maximal speedup of your parallel solution over the serial solution with the increased number of processors. You can assume some numerical numbers for t_1, t_2, t_3, and t_4 if necessary.

Hint: Each step is possible to be parallelized and between steps can also be parallelized.