CSE 5541 Spring'22 Assignment #1 – Basic Concepts and Principles of Parallel Computing

Neng Shi

2022/02/22

1. Parallel overhead is required execution time that is unique to parallel tasks, as opposed to that for doing useful work.

Parallel Overhead = Parallel performance on one processor - Sequential Time

2. embarrassingly parallel

3. Scalability refers to a parallel system's (hardward and/or software) to demonstrate a proportionate increase in parallel speedup with the addition of more resources.

Strong scaling:

- The total problem size stays fixed as more processors are added.
- Goal is to run the same problem size faster.
- Perfect scaling means problem is solved in 1/P time (compared to serial).

Weak scaling:

- The problem size per processor stays fixed as more processors are added. The total problem size is proportional to the number of processors used.
- Goal is to run larger problem in same amount of time.
- Perfect scaling means problem Px runs in same time as single processor run.

4.

$$Speedup == \frac{1}{\frac{P}{N} + S} = \frac{1}{\frac{1 - 0.19}{128} + 0.19} = 5.09$$

5.

p	56	64	128	∞
Efficiency	2.765	2.777	2.816	2.857

6. In Distributed Memory Model, tasks exchange data through communications by sending and receiving messages.

In Data Parallel Model, tasks is typically organized into a common structure, and a set of tasks perform the same operation on a partition of the same data structure. Typically, there is not many communications.

7. CPU utilization is the amount of CPU time taken for sending/receiving massages. Best networks should have very low CPU utilization, thus the system spend most of time on useful work.

8. Benefits:

- Facilicates loda balcancing.

Deficiencies:

- Relatively small amounts of computational work are donw between communication events.
- Low computation to communication ratio.
- Implies high communication overhead and less opportunity for performance enhancement.

9.

- Step 1: $1 \to 9$
- Step 2: $1 \to 5, 9 \to 13$
- Step 3: $1 \to 3, 5 \to 7, 9 \to 11, 13 \to 15$
- Step 4: $1 \to 2, 3 \to 4, 5 \to 6, 7 \to 8, 9 \to 10, 11 \to 12, 13 \to 14, 15 \to 16$

10. The executing time of jth communication step is:

$$t_j = t_s + 2(q - j + 1)t_h + 2m \sum_{i=1}^{q-j+1} t_{wi},$$

where $q = log(N)$.

So, the total time of one-to-all broadcast is:

$$t_{bcast} = \sum_{j=1}^{q} t_j.$$

We assume that leaf nodes are at level 0, and a level-i node's parent is at level $i+1$. Because in a fat tree, the bandwidth of level $i+1$ branches is twice of the level i branch, so the per-word transfer time of level $i+1$ branches is half of the level i branches. Thus, $t_{wi} = (\frac{1}{2})^{i-1} t_w, t \in [1,4]$. So, the one-to-all broadcast time in a fat tree is:

$$t_{bcast} = \sum_{j=1}^{q} [t_s + 2(q-j+1)t_h + 2m \sum_{i=1}^{q-j+1} (\frac{1}{2})^{i-1} t_w]$$
$$= qt_s + q(q+1)t_h + 4m[q-1+(\frac{1}{2})^q]t_w$$
$$= t_s log(N) + t_h(log(N)+1)log(N) + 4mt_w(logN - 1 + \frac{1}{N}).$$