Project 2B

Tianyi Mo, 875556

Improvement of Project 2A

The previous approach has the height or width of each block equal to the height or width of whole matrix divided by maximum number of threads available. It is improved by using smaller block size (decreasing granularity) to increase the utilization of cores.

Summary of Project 2B

Calculation of each pair of genes can be view as a task and the 2B part is about achieving best performance by scheduling these tasks on multiple nodes. It is important to reduce the load imbalance the while keep the overhead as small as possible. The tasks are assigned to nodes in decreasing order of the task size and each task uses all 16 cores. Since the executing time is not known in advance, the matrix size of task is used as an approximation.

Static assignment has less communication overhead, but in test the slowest nodes is about 30% slower than the fastest one because our approximation is not accurate. Therefore, **dynamic assignment** (Bryant, 2018) is used to schedule tasks and **long tasks are schedule before** small tasks.

For example, 1 large task is allocated to each node at the beginning. When the node finishes the calculation of that task, it sends the result to root node and receive another task from root node until all tasks are scheduled.

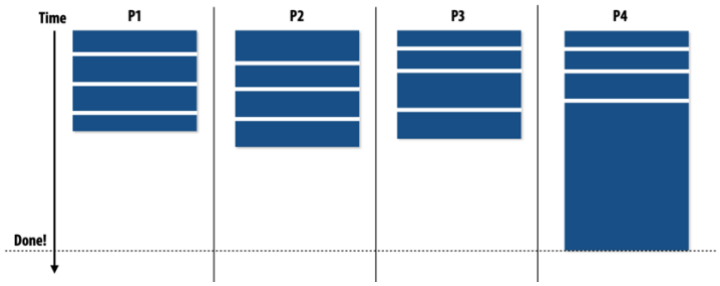

Figure 1. Load imbalance if long task scheduled at last

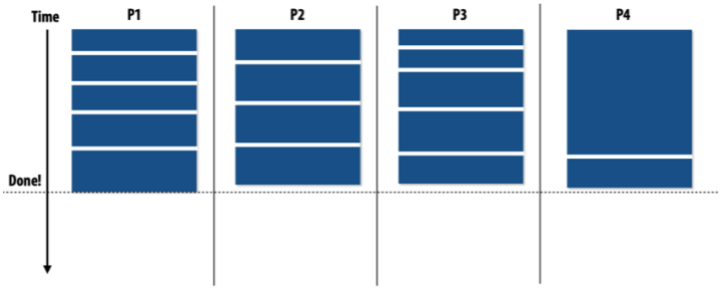

Figure 2. Dynamic assignment of tasks

Bibliography

Bryant, R. (2018). *Performance Optimization Part 1: Work Distribution and Scheduling*. Retrieved from http://www.cs.cmu.edu/afs/cs/academic/class/15418-s18/www/lectures/05_progperf1.pdf