CSCI 6360: Parallel Computing Lecture Summary - 11

Anirban Das (dasa2@rpi.edu)

February 23, 2018

Summary on ROSS/PDES topic covered in lecture 11:

The lecture covered the performance measure, warp speed (!) or basically, super linear speed up achieved while executing ROSS on 1966080 cores on 'Sequoia', the Blue Gene Q system in LLNL facility . The goal of the project is to push the scaling limits of massively parallel discrete event simulation and to determine whether the Blue Gene/Q would achieve scaling performance level at par with the previous BG/L and BG/P systems.

Sequoia system was a BG/Q, with 1 rack containing 1024 nodes, 16,384 cores and can support upto 4 thread or MPI task per core. BG/Q used the standard 1.6Ghz 16 (+2)core IBM A2 processor, 16GB DDR3 ram per node, 32MB L2 cache @563GB/s anf 42.6 GB/s bandwidth. The networking is implemented in a 5D torus. Sequoia had 96 such racks equivalent to 1,572,864 A2 cores and 1.6 petabytes of ram with 16x16x16x12x2 5-D torus, which they upgraded to 120 racks 'Super Sequoia' by integrating Vulcan in a 20x16x16x12x2 5-D torus. However increasing length of one dimension did not increase bisection bandwidth.

In ROSS des, MPI is used to spawn MPI tasks, and it includes all functions for event scheduling, rollback reverse computation and GVT computations containing AllReduce. The logical processes of this simulation is PHOLD. The authors defined the performance measure in **Warp Speed** which is $\log_{10}(p) - 9$, where p is the event rate/s.

Before the actual run, the authors tested ROSS scaling performance at CCNI BG/Q system , on 2 racks 128K MPI tasks with 40 LP per MPI task. It is found that the maximum performance is obtained by running 4 MPI tasks per core. ROSS achieved, at 2018 nodes, approximately 260% performance increase from 1 to 4 tasks/core with 8B ev/s peak at 65,536 MPI tasks, which is super linear scaling.

When run on 7.8M ranks on 120 racks of Sequoia, ROSS initially had some issues because of IBM's PAMI low level message passing system not allowing jobs larger than 48 racks, but then after IBM's fix, the performance was observed at par with run at CCNI. More superlinear speed up behavior is observed at higher core count, for e.g. the speedup from 1 to 48 racks is 74x , with the system performing at 504B ev/s at 120 racks ! ROSS performance over Sequioa achieved 97x speedup form only 60x more hardware.

This high performance can be attributed to the fact that with more and more ranks, the payload is fragmented in so small chunks (only 82MB per node at 120 racks exec), that it is possible to fit almost entire payload in the 563GBps L2 cache instead of the 42GBps ram. That coupled with advance data prefetch allows almost most computation done from cache and hence the speedup.

The astounding performance achieved shows it is possible to do planetary scale simulations eventually.