Research Lab 2: Parallel Processing

Scope

The goal of Research Lab 3 (RL3) is to provide you a perspective of distributed frameworks in other application domains. We will explore MPI-based application for parallel processing in the High-Performance Computing (HPC) domain. While this domain traditionally supported computational work in the science domains, it is/can be extended to other data intensive applications.

The lab can be viewed in either a science or data approach. Scientific applications are numerous and well documented on the Internet. For instance, diffusion, fluid dynamics, and linear solvers. For applications that may be more familiar with engineering or distributed computing, working with many data files or large data sets easily maps to problems where Hadoop-like engines perform the same work on many sets of data (parallelization).

Choosing which direction your team explores is up to your team. Be aware that like the previous labs, we are looking for in-depth investigation and learning, and while there are many online source repositories of completed MPI-based diffusion or other work. The spirit and scoring of the lab is based on your work.

Technologies

RL3 is inherently C/C++ focused on MPI, GPU, or similar scaling. Use of a front-end to drive the C/C++ is common (e.g., mpi4py) is an option but not the entirety of the body of work for the lab.

A data set (but not limited to) has been uploaded (subset of books) to Canvas, the complete data set is 2+ Gb. There are other sets available (e.g., cal fire, world data bank, or possibilities of social media mining).

References

- See threading and MPI lecture references
- Mpi4py: https://mpi4py.readthedocs.io/en/stable/
- https://e-l.unifi.it/pluginfile.php/625533/mod_resource/content/0/MPI_COURSE.pdf
- https://web.cels.anl.gov/~thakur/sc17-mpi-tutorial/slides.pdf
- Perspective: https://www.dursi.ca/post/hpc-is-dying-and-mpi-is-killing-it