HPC - An EBS Perspective

Ze-Yu Zhong

Wednesday 27/09/2023

Introduction to HPC

HPC can be well suited for some scientific workloads.
Overview + hands on session specifically for R

- Monte Carlo study for the IV estimator

Prerequisites

HPC account + Project

- Various computational projects available - MonARCH is specifically built for Monash staff/students

ssh client

- putty (Windows), terminal (macOS or Linux)

ftp client to access HPC files

- MobeXterm (Windows), Cyberduck (macOS), Nautilus/Dolphin (Linux)

Login Node Setup

1. Login to MonARCH
   - ssh user@monarch.erc.monash.edu
2. Load R modules
   - module load R
3. Setup folder for R libraries
   - asd
4. Install R libraries
   - R
   - install.packages("tidyverse")

Code Setup

Recommended to set up git repository containing all code

- cd project/user
- git clone

Job Script (Bash)

Parallelization + Splitting Workloads

HPC can be well suited for parallelized/split workloads.
This is done via the ARRAY environment variable

Monte Carlo Example

IV Regression Example:

$$y = x_1\beta_1 + e_1, \quad e_1 \sim N(0,1) \tag{1.1}$$
$$x_1 = \gamma z + e_2, \quad e_2 \sim N(0,1) \tag{1.2}$$
$$cor(e_1, e_2) = \rho \tag{1.3}$$

Interested in properties of IV estimator $\widehat{\beta_1}$ across different

- Instrument strength $\gamma \in \{0, 0.25, 0.5\}$
- Endogeneity $\rho \in \{0, 0.25, 0.5\}$
- Sample size $N \in \{100, 200, 500\}$

Record $\widehat{\beta}_1$ and $se(\widehat{\beta}_1)$ for each specification, for $R = 1000$ replications

Code

1. Code up a minimum working example that runs on your local machine, e.g. for a small number of replications
2. Take note of how long, extrapolate how much time it would take to run on the HPC cluster
3. Convert the local code to something that is distributed across different HPC arrays
   - Typically, letting each array handle a different DGP specification is most straightforward
4. Prepare job script, and submit

Practical Advice

Do not request too many resources - this can take a long time to be allocated.
Do not mess with job priority unless you have a legitimate reason - this is bad etiquette

Extra Resources

Data Fluency Workshops (free for students!):

- Introduction to Bash/Shell Scripts
- Introduction to HPC

Advanced Issues

rcpp

- Compiling rcpp code on one array and asking other arrays to use this is inconsistent - no guarantee that different arrays are of same architecture
- Solution: ask explicitly for same compute instance nodes OR compile code for each array (inefficient, but not usually not prohibitively so)