Data Science with R and pbdR at ORNL: From the CADES Cloud to the OLCF

Exercises

Drew Schmidt and George Ostrouchov

June 18, 2018

Setup

1. Set up your VM in openstack (optional, but encouraged).
2. Install Docker on your VM (see slides if installing on your laptop) `apt-get install docker.io`
3. ssh to your vm and run R
4. If you have R installed on your laptop, pull a remoter container and connect to it.
5. Pull an RStudio container and connect to it.
6. Pull the shiny k-means container and connect to it.

Profiling and Benchmarking

1. For `x <- matrix(rnorm(1000*250), 1000, 250)`, which is faster (single execution):
   - `t(x) %*% x`
   - `crossprod(x)` ?
2. Explore the call stack of `example(glm)` with `Rprof()`.
3. Re-run exercise 2 with `Rprof(memory.profiling=TRUE)`, and examine with `summaryRprof(memory="both")`. See the help files for an explanation of the new output.
4. Which function is faster on average? Try several values of n.

```
f <- function(n)
{
  x <- c()
  for (i in 1:n)
    x[i] <- i*i

  return(x)
}

g <- function(n)
{
  x <- numeric(n)
  for (i in 1:n)
    x[i] <- i*i

  return(x)
}
```

5. Which function is faster on average? Try several values of n.

```
h <- function(n) sapply(1:n, function(i) i*i)

i <- function(n) (1:n)*(1:n)
```

Parallelism

1. Using randomly generated matrices (example in the slides) of varying sizes, compute the principal components with `prcomp()`. Try using a differing number of OpenBLAS threads, and measure the performance.
2. Create a vector containing the square root of the numbers 1 to 100000 using:
   - `lapply()`
   - `mclapply()` with 2 cores
   - `mclapply()` with 4 cores
   - `mclapply()` with 1 core
3. The Monte Hall game is a well known "paradox" from elementary probability. From Wikipedia:

   ```
   Suppose you're on a game show, and you're given the choice of three doors:
   Behind one door is a car; behind the others, goats. You pick a door, say No. 1,
   and the host, who knows what's behind the doors, opens another door, say No. 3,
   which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it
   to your advantage to switch your choice?
   ```

Simulate one million trials of the Monte Hall game on 2 cores, switching doors every time, to computationally verify the elementary probability result. Compare the run time against the 1 core run time.

Services

1. Build the movie-explorer container from source:
   - ssh to your VM
   - Create a new folder and put the Dockerfile in it
   - run `sudo docker build -t movie-explorer .`
   - run `sudo docker run -i -t -p 3838:3838 movie-explorer`
2. The movie-explorer example comes from the shiny-examples repository. Pick another example, modify the above Dockerfile, and rebuild it.