INF 559 – Spring 2020

Homework #1: Storage Access Patterns

Due: February 4, Tuesday 11:59 PM PT

100 points

In the lectures you have studied the differences between sequential and random file access. In this assignment, you will read increasing amounts of data using sequential and random access on a large file and plot the results obtained.

1. **Preparation**
    a. Download the Ubuntu 18.04.3 ISO file from here:
    http://releases.ubuntu.com/18.04/ubuntu-18.04.3-desktop-amd64.iso
    This will be used as our test file.
    b. Upload this file to your Google Drive.
    c. Open a new **Python 3** notebook in Google Colab[1] (**Warning:** Python 2 notebooks will not be accepted).
    d. Since this is the first time most of you would be using Google Colab (or even Jupyter notebooks in general), we have provided a starter notebook which you can upload to Colab to get a head start. The notebook outlines the expected layout of your final submission and also contains code to mount your Google Drive and read files from it. This step is not mandatory but is strongly recommended.

[1] Here’s a good guide to get started with Colab: Overview of Colab.

2. Plotting latency and bandwidth [50 points]

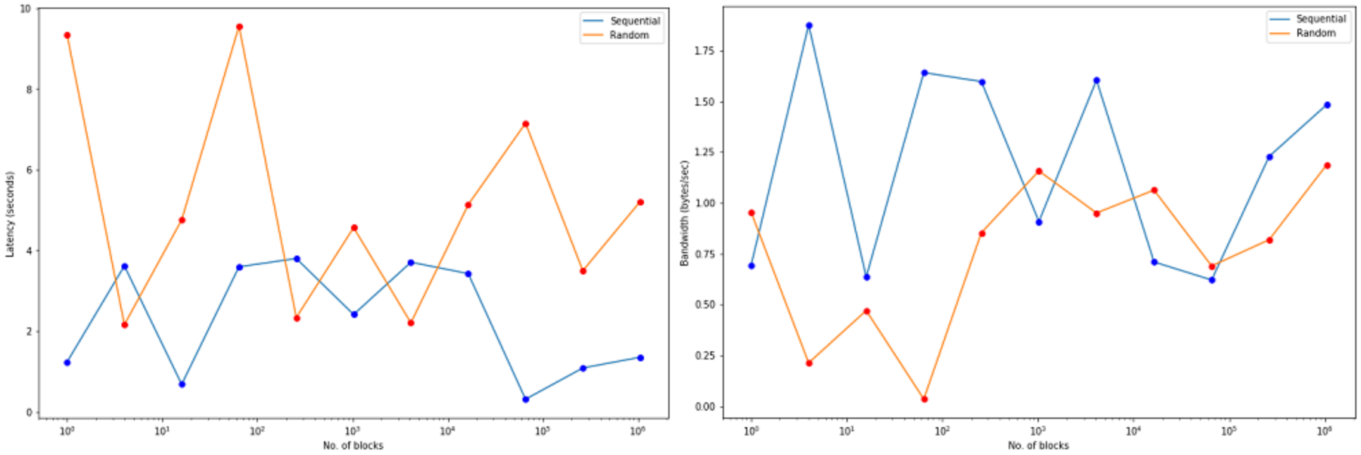
a. Open the test file in **unbuffered** mode[2].

b. Sequentially read [1, 4, 16, 64, 256, 1024, 4*1024, 16*1024, 64*1024, 256*1024, 1024*1024] blocks of data[3]. Use a fixed block size of **4KB**. Measure the latency for each iteration in terms of wall-clock time.

c. Repeat 2b with random reads instead of sequential.

d. Plot the latencies measured in 2b and 2c against the number of blocks read. Both sequential and random results should appear on the same plot and the number of blocks should be scaled logarithmically instead of linearly. Briefly describe your observations from this plot.

e. Calculate the bandwidth for each iteration of 2b and 2c using the latency and amount of data transferred. Plot the results in the same manner as latency and briefly describe your observations.



Sample output for step 2 (generated using random numbers, not suggestive of actual output)

[2] Refer to the documentation to switch buffering off. Also make sure you use the binary read mode while opening.

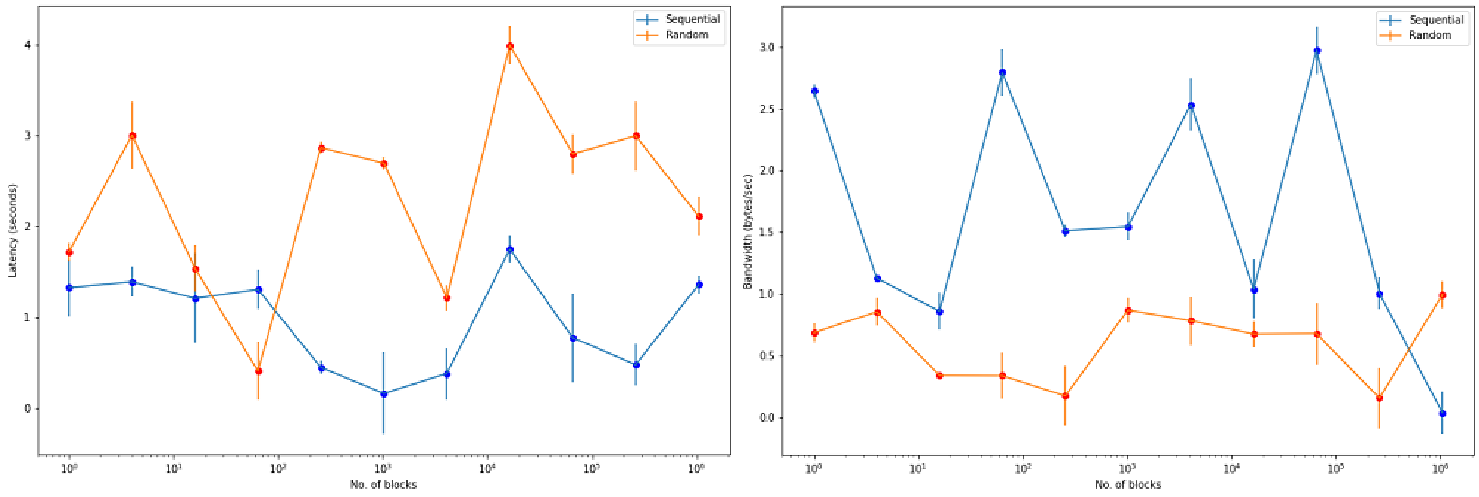
[3] It may be possible that you reach the end of the file prematurely during sequential access. Make sure to seek to the start of the file again and continue reading in this case.

3. Monte-Carlo simulations [50 points]

a. Run 10 simulations of steps 2b and 2c and store the results.

b. For each of the 11 iterations, calculate the mean and standard error[4] over the 10 simulations.

c. Use the results of 3b to generate errorbar plots for latency and bandwidth. Again, both sequential and random results should be on the same plot and the number of blocks should be scaled logarithmically. Briefly describe your observations from these plots.



Sample output for step 3 (generated using random numbers, not suggestive of actual output)

[4] The standard error is defined as the standard deviation divided by the square root of the number of observations.

Submission:

1. Submit a single file on Blackboard **FirstName_LastName_hw1.ipynb**
2. The only file format accepted is **.ipynb**. You do not need to submit the plots and explanations separately. Describe your observations using text cells (Jupyter notebooks allow both code and text cells). Your final notebook will have both the plots and explanations as part of it.
3. Make sure to mention your name and USC ID in your notebook (as done in the starter notebook).

Grading Criteria:

1. Late submissions (up to 24 hours) will be penalized by 20%. No credit will be given after 24 hours of submission deadline.
2. As mentioned above, Python 2 notebooks will receive no credit.
3. The submitted notebook must have all its cells executed and outputs visible (if applicable). Notebooks without outputs will be penalized by **30%**.
4. You may use any Python internal library, but the only external libraries allowed are **numpy** and **matplotlib**. Both these libraries are already installed in Colab and you just need to import them. Use of any external library other than these will be penalized by **20%**.

Important Notes:

1. Submitted work must be your own. Don't share your code with anyone.
2. The Monte-Carlo simulations may take up to 5 minutes to execute. This is expected behavior given such large reads. To make debugging easier, you may decrease the simulations to 2 or 3 but make sure to run all 10 simulations before the final submission.
3. Start early, and make sure to visit the TA's during office hours to make sure you're on the right track or if you need help.