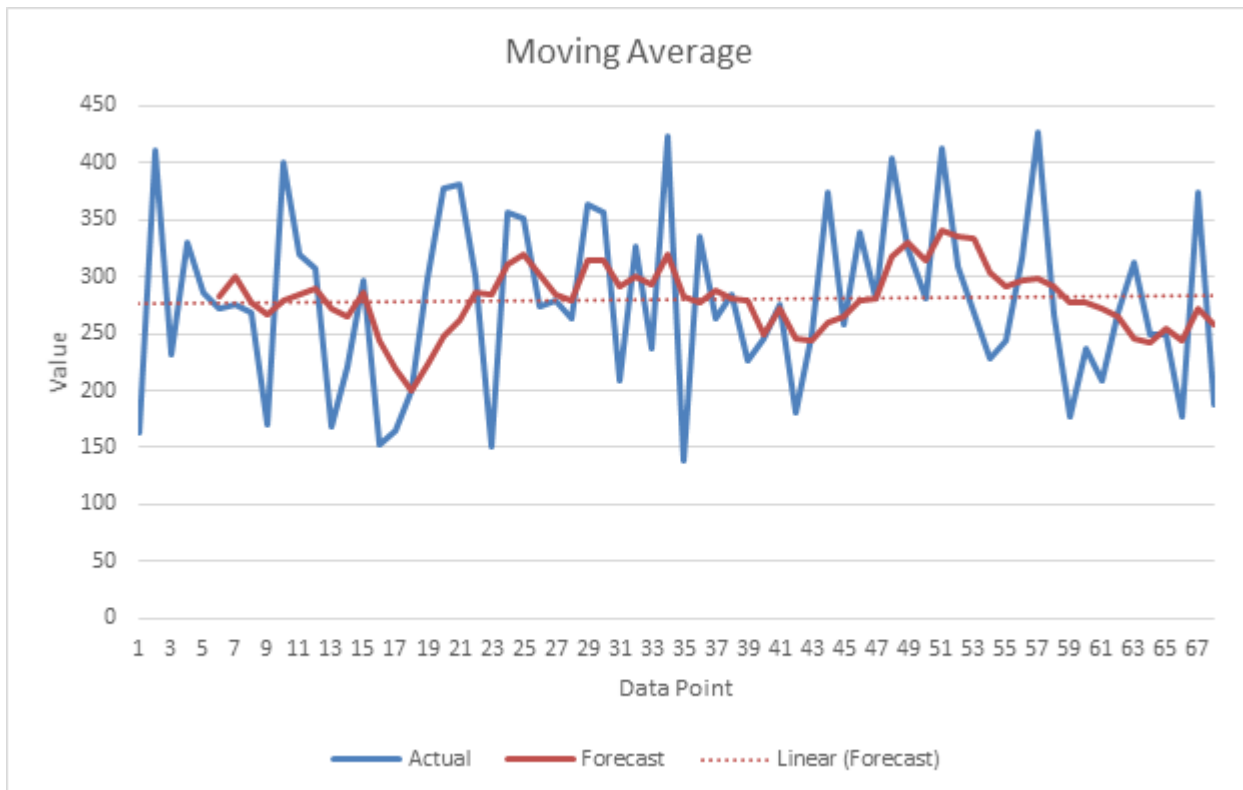


Introduction

Median filtering is a nonlinear digital filtering technique that is often used to remove noise from a data set, a median filter slides over the data item by item, generating a new data set, where each item is replaced by the median of neighboring entries. Few or several data items considered to be filtered are within a filtering window of a certain size.



Aim and predictions

Use the Java Fork/Join framework to parallelize the median filter operation using a divide-and-conquer algorithm, where I filtered parts of the data input at the same time where work is distributed almost equally to unused cores. The naïve approach to achieving the aim is by sequentially going through all the data item-by-item utilizing one core, running order $O(n)$ or not all the computer resources and this can be expensive when n is very large, especially if the filter window is large enough. The parallel algorithm is to be correct and faster relative to serial operations, to give an $O(\log(n))$ runtime.

Predictions:

1. Parallel program should be significantly faster than serial version with the same filter size.

2. An increase in filter size, parallel program should be slower approaching serial version pace.
3. A decrease in SEQUENTIAL CUTOFF should increase speed of parallel program

Methods

Sequential

Designed a sequential median filtering program that takes the following parameters, an array of data items to be operated, filterSize window to utilize in finding median, lastly a lower & upper index on the array to start the filtering process.

A medianFilter method is applied to filter the data according to the given constructor parameters.

```
public SerialFilter(float[] file_data ,int filter_size, int start, int end){}
```

Then a medianFilter method is called and returns a filtered array from lower(start) to upper(end) index is returned.

Parallel

Large data items are divided into smaller ones, which median filtering can be applied to multiple process at the same time. The following compute method parallelizes the program.

The parallelization program constructor takes the same arguments but with one more, the sequential cuttOff, which is a range value within full data items.

```
ParallelFilter(float[] array, int l, int h, int filter_size, int seq_cuttOFF) {  
    ...  
    ...  
}  
  
protected float[] compute() {  
    if ((hi - lo) <= SEQUENTIAL_CUTOFF) {  
        return (new SerialFilter(arr, filter_size, lo, hi)).medianFilter();  
    } else {  
        ParallelFilter left = new ParallelFilter(arr, lo, (hi + lo) / 2,  
filter_size, SEQUENTIAL_CUTOFF);  
        ParallelFilter right = new ParallelFilter(arr, (hi + lo) / 2, hi,  
filter_size, SEQUENTIAL_CUTOFF);  
  
        left.fork();
```

```

        float[] rightAns = right.compute();
        float[] leftAns = left.join();

        float[] both = Arrays.copyOf(leftAns, leftAns.length +
rightAns.length);
        System.arraycopy(rightAns, 0, both, leftAns.length, rightAns.length);
        return both;
    }

```

1. Description of approach

The full data items are inspected whether they are more,less or equal to the sequential cutoff according to upper(hi) and lower(lo) indx

```

if ((hi - lo) <= SEQUENTIAL_CUTOFF) {

```

If it is less than sequential cutoff or equal to, then the sequential(median) filtering is applied to that *piece* of data and a filtered piece is returned by the sequential filtering program,

```

return (new SerialFilter(arr, filter_size, lo, hi)).medianFilter();

```

if it is more than the sequential cutoff, the data is cut into 2 half pieces, by creating the *left* and *right* objects.

```

} else {
    ParallelFilter left = new ParallelFilter(arr, lo, (hi + lo) / 2,
filter_size, SEQUENTIAL_CUTOFF);
    ParallelFilter right = new ParallelFilter(arr, (hi + lo) / 2, hi,
filter_size, SEQUENTIAL_CUTOFF);

```

and the 2 half pieces each run on two separate threads from the main thread where the left object has independent execution [fork()] from the main thread, and the right thread is run on another separate thread[right.compute()], then the left thread waits for the right thread to finish [left.join()].

and repeat the same process of being inspected(at the top) whether each is more,less or equal to the sequential cutoff. This is a recursive process.

```
left.fork()  
float[] rightAns = right.compute();  
float[] leftAns = left.join();
```

since the sequential median filtering program only returns the filtered output from lower and upper index given, i.e if data elements are not on a range to be filtered, then that whole range is not returned)

The two half pieces returned are then pieced together after being filtered, by combining them together, joining the right piece to the right.

```
float[] both = Arrays.copyOf(leftAns, leftAns.length + rightAns.length);  
System.arraycopy(rightAns, 0, both, leftAns.length, rightAns.length);  
return both;
```

2. Speedup was measured by dividing the parallel program runtime by the sequential runtime for each experimental input.

3. Validating the Algorithm

- Debugged the algorithm, inspecting it line by line as to whether it is taking the correct actions when filtering some input data.
- Made use of small data inputs, then increased to large data input.
- Made use of a small data items and confirmed by hand(manually) for all filters windows(3 to 21) and varying the sequential cutoff that the filtering was correct, increased input data items size after some time.
- Applied median filtering to half(the left or the right) of input data items, e.g when there's 50 data items to filter ranging from zero, I input lower index as 25 and upper as 50, and let the parallel program filter the upper half(25-50) without changing the lower half, also vice versa.

4. Timing the Algorithm.

Timing was restricted to the median filter operation (only), excluding the time to read in the file with data input beforehand, and print the cleaned data afterwards, as well as other unnecessary operations.

Two timing methods were created *tick()* and *tock()*.

tick() records the current time translated to milliseconds.

```
private static void tick(){startTime = System.currentTimeMillis();}
```

tock() also records the current time, but subtracts it from *startTime* so to get the difference, divided by 1000 to return the time in seconds.

```
private static float tock(){return (System.currentTimeMillis()-startTime)
/ 1000.0f;}
```

The algorithm was run 20 times, finding timing average of the 20 runs across a range of data sizes and number of threads (sequential cut-off).

```
for (int indx = 0; indx < 20; indx++) {
    System.gc();
    tick();
    float[] b = toParallelFilter(file_data, filter, seq); //
seq=sequentialCutOff
    a = a + tock();
}
a = a / 20;
```

System.gc() is called to minimize the likelihood that the garbage collector will run during the execution.

A *Experiment.java* program ran the algorithm 20 times for each 5% increments from data input length of sequential cut-off for all the filter windows i.e 3-21. e.g If data input with 100 items is ran on the algorithm, then a sequential cut-off start at 5% which is 5, and runs the algorithm 20 times for each filter size averaging timing for that filtersize with that sequential length, then increment sequential with another 5% which will be 10, then again runs the algorithm 20 times for each filter size averaging timing for that filtersize with that sequential length.

5. Machine architecture.

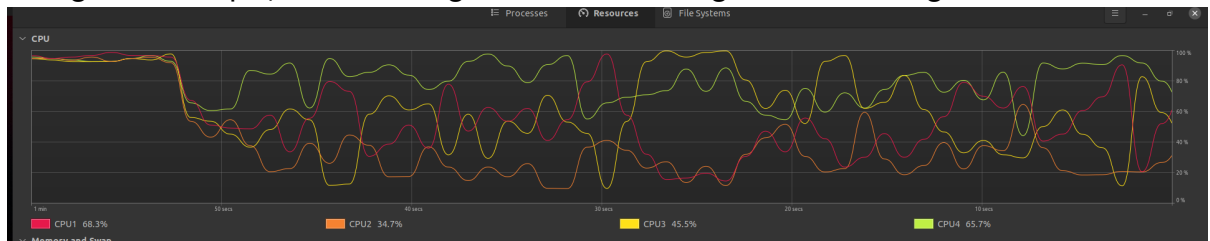
Memory	7.1 GiB
Processor	Intel® Core™ i3-7020U CPU @ 2.30GHz × 4
Graphics	Mesa Intel® HD Graphics 620 (KBL GT2)
Disk Capacity	500.1 GB

6. Problems encountered.

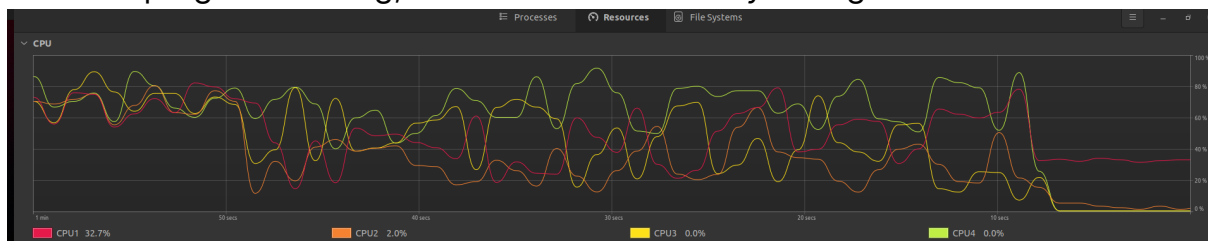
1. Nondeterminism, recieved a **huge** timing difference results for the same data input size, with the same sequential cutoff and the same filtersize, even after I have warmed up the computer by running warm up test, as much as this is an research assignment, the timing difference should not be too much off, there's a small range of allowed difference. After investigation, my algorithm was not the problem, I found difference in cpu spikes.

For example *(these images are not original to when i discovered the problem but relate)*.

At first time running the experiment, I had Visual Studio Code, and Firefox browser(watching YouTube video) running with many tabs open, thus it was using a lot of cpu, and slowing down the the algorithm making it slow.



Later that day, after powering on my computer I ran the experiment again, with no other program running, the results dramatically changed.

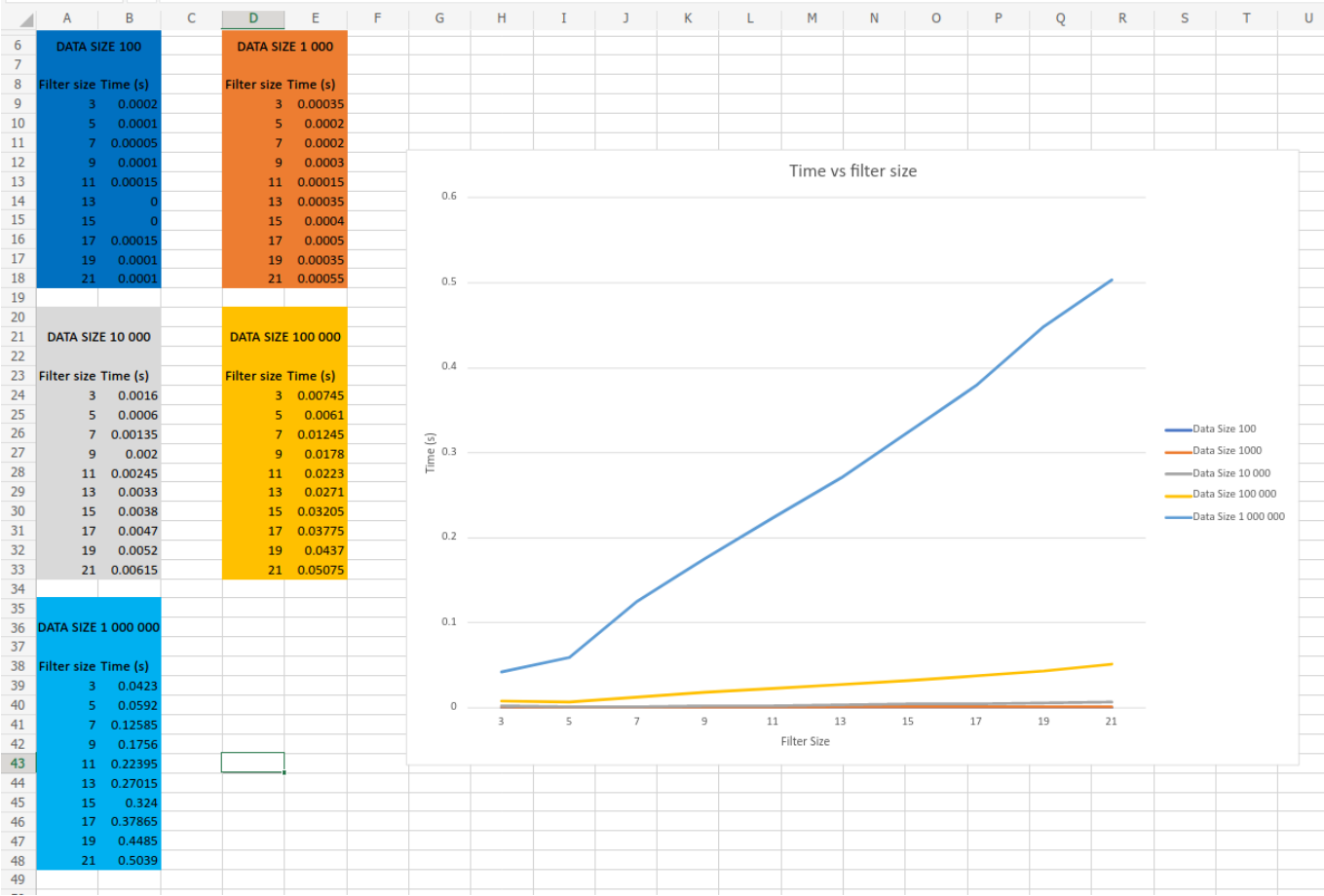


Experimental Results

Serial Program

For every data size, every filter window size that program has been run 20 times and the time of execution has been averaged.

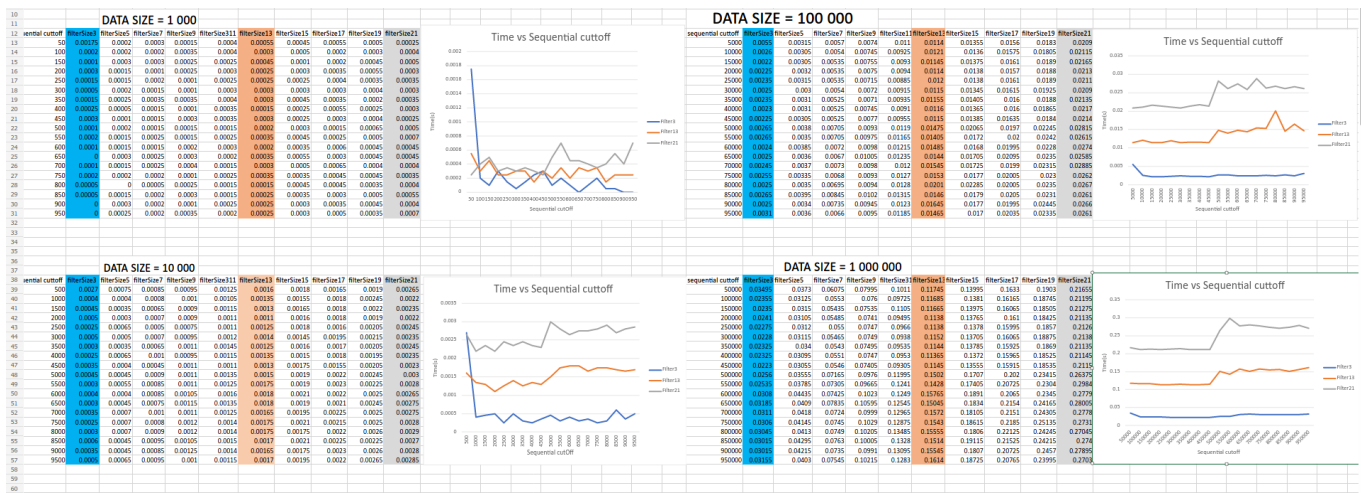
Each data size has a different colour relating to the data table on the left.



The extrapolated graphs show that for every data size, an increase in the filtering window increases the runtime of filtering of data.

Parallel Program

The sequential cut-off of every data size has been made of increments of 5% of the full data size, and for every sequential cut-off , every filtering window size has been run 20 times and the time of execution has been averaged, i.e for every sequential cut-off and filtering window size the program has run 20 times, then the timing was avaraged.

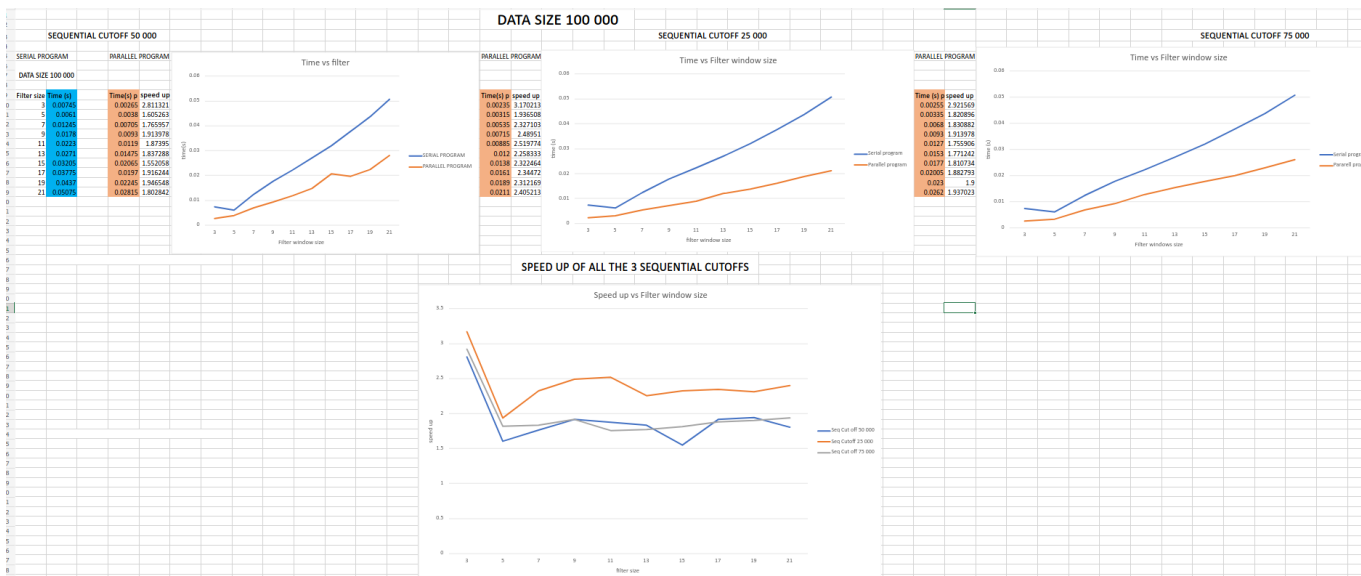


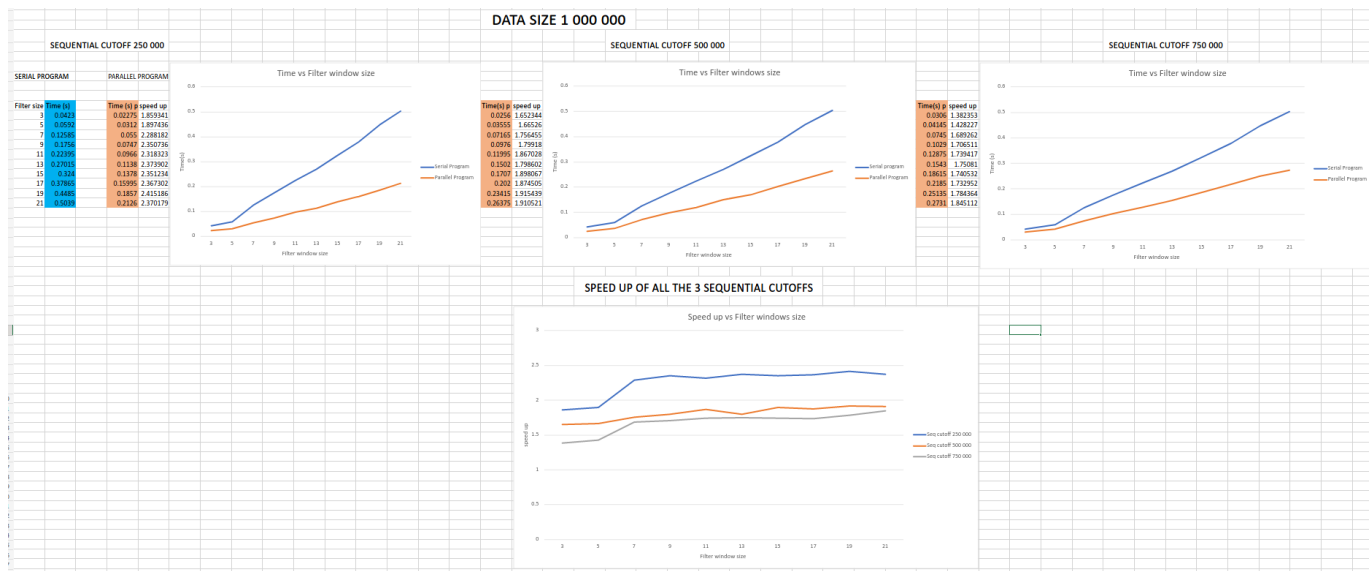
The extrapolated graphs show that for every data size,

1. the lower the sequential cut-off and the lower the filtering window size then the shorter the amount of time the program takes to filter the data, i.e the program speed increase.
2. Once the sequential cut-off has reached half of the data size, there's spike(increase) in the time the program filters the data regardless of what the filter window is.

Speed up

To determining the speed up, a serial program and parallel program were run on two data sizes, but for a parallel program the 100 000 data size with sequential cut-off of 25 000, 50 000 & 75 000 and for a data size of 1 000 000 with sequential cut-off of 250 000, 500 000 & 750 000.





From the data collected and graphs extrapolated, and for all filter sizes, the parallel program is more efficient relative to the serial program, from the speed up graph below the parallel program is twice faster than the serial program, with the highest speed up value of 3.17 on the 100 000 data size collected data, this result is far from the expected speed up, the highest speed up can be obtained when the data size is low as possible. This suggest that an increase in data size, will slow the program for both the serial and the parallel program.

The optimal sequential cutoff for the 100 000 data set is 50 000, and 250 000 for the 1 000 000 data set. This shows that the more the work is distributed up and processed at the same time equally, the faster the program becomes for all data sizes but the optimal sequential cut-off will vary.

Conclusions

- My first prediction was wrong, the parallel program did not run on order $O(\log(n))$, infact the obtained results grew lineary from the serial program.
- The lower the filter and with a suitable sequential cut-off results in the high peak of the speed up of the program, the parallel program is almost 3x efficient than the serial program when the data size is not very much large, which is in the line of Amdahl's law, speed up isn't much when problem size is small.
- An increase in data size, results in a little decrease in speed up of the program to when the data size is almost 10x smaller.
- Parallelising is effective but not signicficatly, an increase in computer cores does not significantly improve the speed.

Appendices

GIT LOG

commit 5701152313e4f6c34c0f11e72305817c73044136 (HEAD -> master)

Author: myname <myemail@example.co.za>

Date: Fri Aug 27 23:15:39 2021 +0200

Created and added both parallel and serial programs used for experiment and for collecting data

commit 9258a609cfb8a6ccd8c5af02b974850cceae814a

Author: myname <myemail@example.co.za>

Date: Fri Aug 27 23:04:25 2021 +0200

Created a Parallel Algorithm testing program.

commit 9386ba8671ef8a2e1ad6b6895ec007e24cca223b

Author: myname <myemail@example.co.za>

Date: Wed Aug 25 14:27:14 2021 +0200

Added description and comments to ParallelFilter program.

commit 398a0262803905208d849592b2336b42e357a3ff

Author: myname <myemail@example.co.za>

Date: Wed Aug 25 14:04:24 2021 +0200

Added SerialFilter description (fix spelling mistake).

commit f02c373f088dcdbd1e613c13963e873bb7c8c5886

Author: myname <myemail@example.co.za>

Date: Wed Aug 25 13:59:57 2021 +0200

Added SerialFilter description.

commit 66c449f56d2a4ec58a99bac4d312894c919c485b

Author: myname <myemail@example.co.za>

Date: Wed Aug 25 13:41:26 2021 +0200

Created and Finished ParallelFilter program.

commit e56d48d76b86150d848d33a84e79905085a44266

Author: myname <myemail@example.co.za>

Date: Wed Aug 25 13:40:12 2021 +0200

Added comments to SerialFilter program

commit a7a03e9bed1f50964a4991c35c9303473471dc7e

Author: myname <myemail@example.co.za>

Date: Wed Aug 25 13:12:05 2021 +0200

Finished program creation.

commit 6736b227b5f85e2f18265082704d192cef271134

Author: myname <myemail@example.co.za>

Date: Mon Aug 23 20:14:24 2021 +0200

Created serialFilter java file.

(END)