Due Friday, November 18, 2013

Thread Creation and Destruction

The object of this program is to examine the time it takes to create and destroy an individual thread based on how many other threads are being created. There timer starts right before the first thread is created, and ends after the call to `pthread_join` returns. Each of the threads does no work and returns immediately, so we can reasonably assume that each one takes no time to execute. Each timing was done 1,000 for every given number of threads, so the times shown are pretty good averages. The results of the timings versus thread size are shown below.

Number of Threads	Total Time(s)	Average Time(s)
1	0.000517027	0.000517027
2	0.000646301	0.000323150
3	0.000666475	0.000222158
4	0.000782081	0.000195520
5	0.000851409	0.000170282
6	0.000903886	0.000150648
7	0.000880552	0.000125793
8	0.000961013	0.000120127
9	0.000938553	0.000104284
10	0.001036311	0.000103631
11	0.001011882	0.000091989
12	0.001089187	0.000090766
13	0.001091393	0.000083953
14	0.000915015	0.000065358
15	0.001463961	0.000097597
16	0.001634600	0.000102163
17	0.001385159	0.000081480
18	0.001310196	0.000072789
19	0.001337876	0.000070415
20	0.001384097	0.000069205
25	0.001657484	0.000066299
30	0.001788521	0.000059617
35	0.002004404	0.000057269
40	0.002324859	0.000058121
45	0.002483848	0.000055197
50	0.002571187	0.000051424

Table 1: The results of the total and average creation time of 1 to 50 threads

While this data is great, there's a lot of it, and it isn't all that illuminating at first glance. Basically, the cost of starting and stopping a thread is high enough so that for a few threads, it is easily seen. However, as the number of threads increases, the creation/destruction time begins to converge to about 0.00005 s, or 50 μs. There are anomalies in the data; the trend line isn't a completely smooth curve. However, there is enough data to show that over time, the values do approach a constant.

The plot below shows how even though there is a high cost for small numbers of threads but the average

cost converges suprisingly quickly.

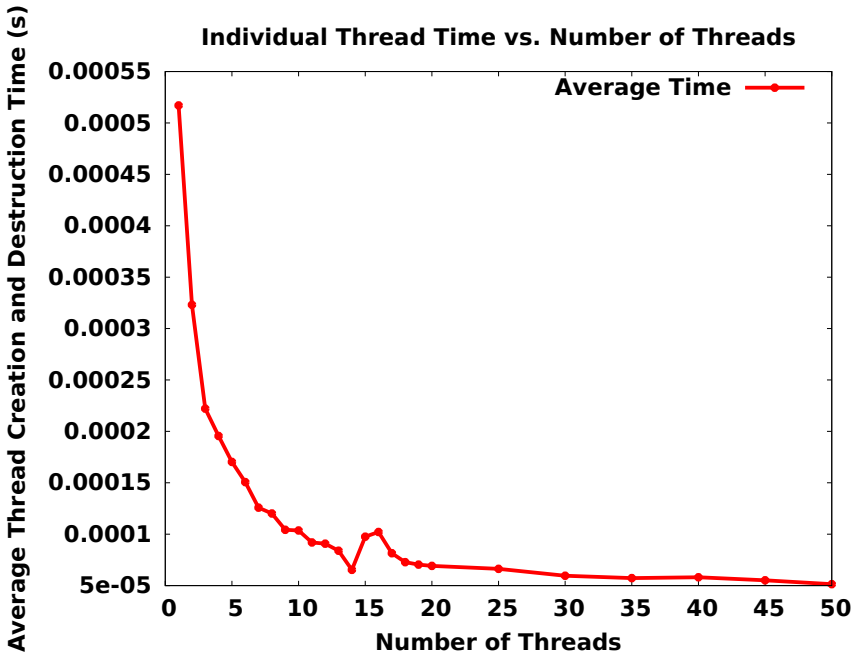

Figure 1: The plot showing the same data as Table 1.

The code that creates, destroys, and measures the time of each thread is included with this submission as `q1.c`. If you wish to run the program for yourself and collect your own data, I have a script that does so, outputs the results, and makes plots and tables for you. I will also include this with the submission as `proj2.sh`. Be aware that the script collects enormous amounts of data so that it can get good averages due to random system scheduling variables that frequently throw off the timings. Depending on the machine you run it on, it may take you anywhere from 3 to 30 minutes.

Histogram problem

We started with a serial implementation of a histogram problem provided to us by Dr. Ribbens. The goal was to use the `pthreads` library to parallelize the code. Further requirements were that the code had to run in $O(n/t + b)$ time where n is the number of data elements, t is the number of threads, and b is the number of bins to sort the data into.

There were two main aspects to parallelizing the code. The first had to do with breaking the initial data array up into n/t parts, and having each thread do some bin sorting on each part. Once that is accomplished, all threads wait until all sorting has been done. Then they work to consolidate all the bins into one giant count of bins, each thread summing up an index of each local bin.

That's the basic idea. The source code might be more illuminating. Describing what the code does at a higher level doesn't quite do it justice, because there's a lot of small details that went into making sure edge cases were handled correctly, and also making sure that the runtime fit the requirements for full credit. THe source code for part 2 is included with the submission as `histogram.c`. You will also need to link the math library (`-lm` flag) in order to compile it. If you run the `proj2.sh` script that I wrote, it should compile it for you so you won't have to worry about anything. I suggest you do that. Keep in mind that because of the enormous amounts of data collection from problem 1, it will take a long time for the script to run. You can either comment out some of the lines or change the number of repititions to make it faster. It's all up to you.