CS4200-001 Computer Architecture I
Parallel Compute Assignment
Sam Allen

---

i. Setup

For this assignment, I am using Google Colab for my parallel compute system instead of INCLINE. I have created a new notebook containing my code which is included in my submission for your reference.

I am also using the UCCS Blanca server in place of my local compute system, as mkl is not supported on my 2020 MacBook Air which uses an M1 chip instead of Intel.

All code given in the example notebook provided runs as expected.

ii. lscpu

```
!lscpu
```

Because I am using Google Colab instead of INCLINE, I used the command above to print information about the CPU and cache for my parallel compute system.

```
lscpu
```

Similarly, I used the command above to print information about the CPU and cache on the UCCS Blanca server.

After comparing the output of these commands, the first thing I observe is that my parallel compute system and my local compute system use a very similar model of CPU, and have the same number of cores.

However, there are some significant differences. The greatest difference between the two is cache size. For both L1 cache and L2 cache, the UCCS Blanca server has about double the amount of cache than Google Colab. The L3 cache of the UCCS Blanca server is also greater, though only by 5 MiB. Another difference is that Google Colab has 4 total threads, while the Blanca server only has 2.

Google Colab:

CPU model: Intel(R) Xeon(R) CPU @ 2.20GHz
CPU cores: 2
Threads: 4 (2 per core)
Cache L1d size: 32 KiB
Cache L1i size: 32 KiB
Cache L2 size: 256 KiB

Cache L3 size: 55 MiB

UCCS Blanca server:
CPU model: Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz
CPU cores: 2
Threads: 2 (1 per core)
Cache L1d size: 64 KiB
Cache L1i size: 64 KiB
Cache L2 size: 512 KiB
Cache L3 size: 60 MiB

As you can see above, my parallel compute system utilizes hyperthreading, as there are more logical processors (threads) than physical processors (cores).

iii. Compute

I have captured the data from all computations for both my parallel compute system and my local compute system. This data is provided below:

Google Colab:

# Threads	1
# Elements	**Time (s)**
10	0.0008218288421630859
50	0.001476287841796875
100	0.0012285709381103516
250	0.0019161701202392578
500	0.00733494758605957
750	0.021157026290893555
1000	0.05231451988220215
2500	0.5696053504943848
5000	3.9575250148773193
7500	7.75937557220459
10000	15.506006717681885

# Threads	2
# Elements	
10	0.001676321029663086
50	0.001369476318359375
100	0.0011417865753173828
250	0.004548072814941406
500	0.016391515731811523
750	0.030424833297729492
1000	0.04295659065246582
2500	0.5269920825958252
5000	4.071028232574463
7500	7.0238611698150635
10000	15.22873044013977

# Threads	1
# Elements	Time (s)
10	0.0008218288421630859
50	0.001476287841796875
Average:	2.53443291

# Threads	2
# Elements	
10	0.001676321029663086
50	0.001369476318359375
Average:	2.449920047

UCCS Blanca server:

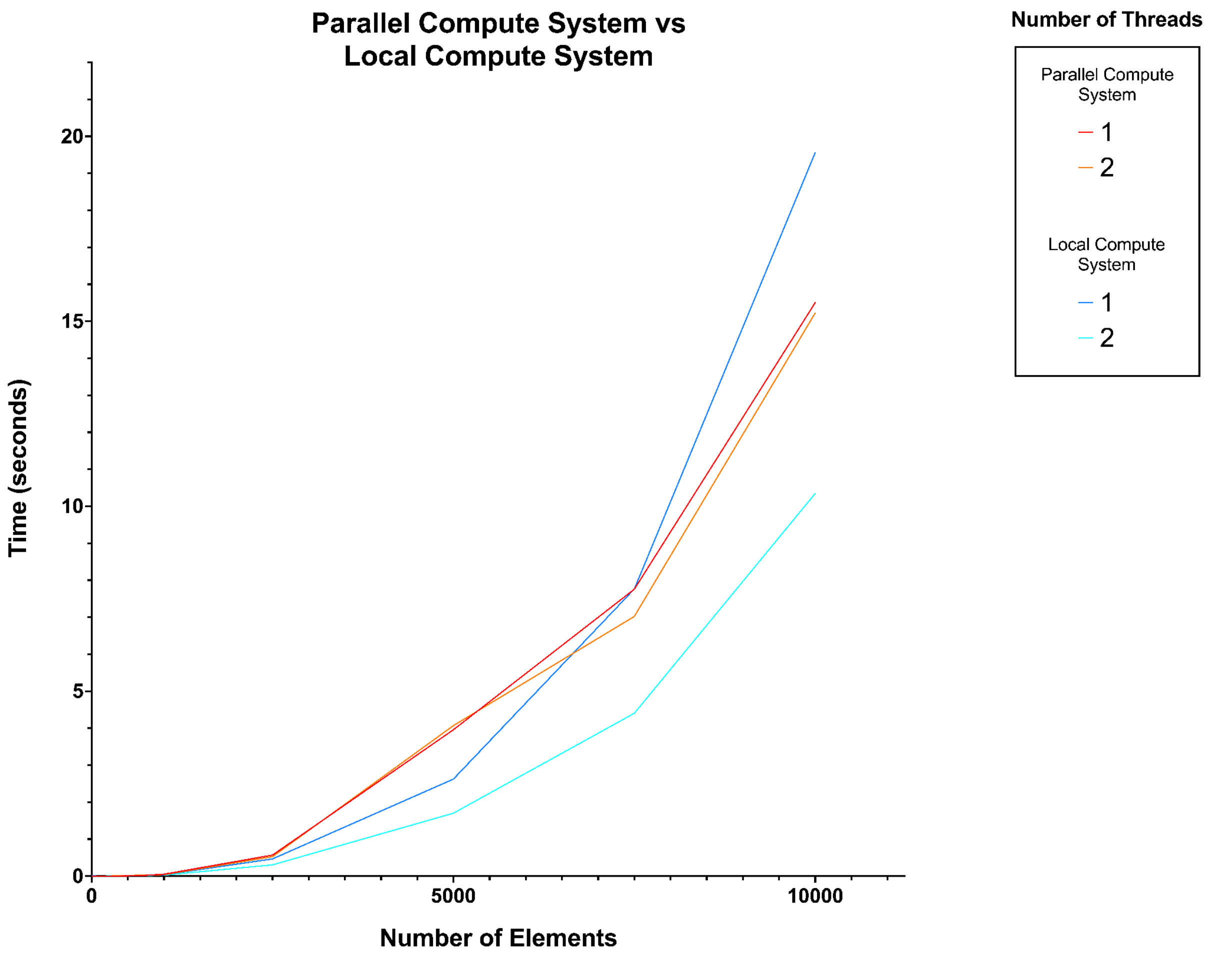
# Threads	1
# Elements	Time (s)
10	0.007241725921630859
50	0.005825042724609375
100	0.001413583755493164
250	0.005063295364379883
500	0.009050369262695312
750	0.026053190231323242
1000	0.038248538970947266
2500	0.4660756587982178
5000	2.6264007091522217
7500	7.7626330852508545
10000	19.56199359893799
Average:	2.773636254

# Threads	2
# Elements	Time (s)
10	0.003007650375366211
50	0.001653432846069336
100	0.001077890396118164
250	0.0063877105712890625
500	0.007876157760620117
750	0.020916461944580078
1000	0.028319835662841797
2500	0.3062717914581299
5000	1.7030730247497559
7500	4.404102325439453
10000	10.348954916000366
Average:	1.5301492

v. Graph

I have created a graph of the performance data for both my parallel compute system and local compute system.

(I have also included a full-size PDF document of the graph in my submission for reference.)



v. Summary

As you can see from the data tables and the graph above, there are significant differences in computation performance between the parallel compute system (Google Colab, Intel Xeon) and the local compute system (UCCS Blanca server, Intel Xeon).

For example, when increasing from 1 thread to 2 threads using Google Colab, the overall performance stayed mostly the same, with some computations even increasing in runtime. In contrast, the UCCS Blanca server displays a significant improvement in performance when threads are increased. At 2 threads, the overall runtime of the UCCS Blanca server marks the lowest of the all data collected.

Since we are only using up to 2 threads and the CPUs of both systems are incredibly similar, this can be attributed to the cache size difference between the two systems. Considering that the UCCS Blanca server has about double the size of Google Colab, this means that threads are much more productive, as they don't have to travel as far down the memory hierarchy to obtain the data needed for computations. This is especially true for larger matrices, and this is reflected in the graph above.