

High Performance Computing

Assignment 2

Theory Questions

1. Total Number of Nodes = 260

Number of GPU Nodes = 64

Number of large memory capacity nodes = 36

Number of nodes required for high priority jobs = 128

Number of regular nodes left = 32

- a. A normal partitioning scheme could have been dividing all 64 GPU nodes into one partition, similarly all 36 large memory capacity nodes as a partition, all 128 nodes needed for high priority jobs as one and last partition of 32 nodes. But that would not be the best way, probably.

Instead, one partition could be of 128 non-accelerated nodes that are exclusively for high priority jobs. This will ensure all high priority jobs can be executed as quickly as possible without interference. The one partition overlap could be if low priority jobs are running on this partition. But that can be resolved by pre-emption. That is, if a new high priority job comes, then the running low priority jobs can be pre-empted and continued when the high priority job is completed.

Other partitions will include a few regular nodes, few GPU nodes and few large memory nodes. Making these partitions to be used for normal scheduling, effective resource utilization and also GRES (I know this is SLURM specific, but it could be useful for high computational power jobs without disrupting the scheduling of other jobs). So, maybe 2 partitions, one could be TST env other PROD env, could be of 16 regular nodes, 30 GPU Nodes, and 15 large memory capacity. Another one partition (DEV env) could be left over 4 GPU Nodes and 6 nodes of large memory capacity nodes.

- b. Some provisions that should be implemented to facilitate parallel job debugging would be:
- Use DEV and TST for development and testing debugging is done in TST env
 - Using the partitions with regular nodes, GPU and high memory nodes to parallelize jobs effectively
 - Implement parallel programming models and libraries like OpenMP, MPI to be able to actually run parallelization code for all resource management suite (SLURM,

- Hadoop YARN, OpenLava, PBS). Also as administrator, implementing any tools that help in optimizing the code and identifying performance issues would be beneficial
- Nodes with similar configurations and properties should be assigned for parallelization in each partition to ensure maximum and efficient parallelization
 - When assigning nodes, as an administrator, providing all necessary details of configurations to ensure optimal resource utilization and effective parallelization
- c. Some strategies and features of some schedulers that can minimize the impact of conflicts between jobs of different priorities are (Not SLURM specific):
- Event Triggered Scheduling: Whenever a new event is added or system configurations are changed (trigger), only a few jobs at the front of the queue are considered and scheduled. This is to ensure no starvation of low priority jobs
 - A feature that can be added to this is the considering of ALL the jobs in the queue to make scheduling decisions. But this requires a lot of overhead, therefore can only be done at infrequent intervals
 - Gang Scheduling: This is an efficient scheduling to ensure no starvation and proper resource allocation based on priority and job requirements. The jobs that have similar properties and configurations are allocated similar resources and the resources are alternated between them. Time slicer is added to keep checking if any job has been suspended for too long and then the active jobs are suspended and added to back of queue. This is to ensure the low priority job that has been suspended for too long gets the chance to start and finish
 - Pre-emption: A variant of gang scheduling, This allows high priority jobs to interrupt any running low priority jobs if there is a resource allocation conflict. The low priority jobs are suspended and continue once the high priority jobs are completed

2. Required: peak 100Tflops machine

Unit1 = dual socket CPU with 12-Core processor operating at 3.4GHz

Unit2 = dual socket CPU with 20-Core processor operating at 2.5GHz

Number of Floating point operations per clock cycle = 4

Floor Space = 32U for nodes in each rack

Nodes per rack = 16

Floor space taken by 1 rack = 1U

Max Number of nodes possible = 32

- a. For Unit 1,
- Per clock cycle = 4 floating point operations
 - Clock Speed = 3.4 GHz
 - Number of cores = 12

Peak Performance per node = $(3.4\text{GHz} * 4\text{FLOPS/cycle} * 12\text{cores})\text{GFLOPS} = 0.1632\text{TFLOPS}$

For 16 nodes per rack = 2.6112 TFLOPS

For Unit 2,

Per clock cycle = 4 floating point operations

Clock Speed = 2.5 GHz

Number of cores = 20

Peak performance per node = $(2.5\text{GHz} * 4\text{FLOPS/Cycle} * 20\text{cores})\text{GFLOPS} = 0.2000\text{TFLOPS}$

For 16 nodes per rack = 3.2 TFLOPS

The 20-core processors use less floor space and provide higher TFLOPS. Therefore, recommended option would be **20-Core Processor 2.5 GHz Dual socket CPUs**

b. Required performance = 100 TFLOPS

1 rack = 3.2 TFLOPS

Number of racks = $100/3.2 = 31.25$ Racks

Instead of getting 0.25 racks, we would need **32 Racks**. That means it will occupy exactly 32U floor space.

c. 32 Racks * 3.2 TFLOPS = 102.4 TFLOPS

We get computational performance of 102.4 TFLOPS.

102.4 tera floating point operations can be performed every second taking 32U floor space.

3. Execution of a 1 million -instruction program takes 2.5 ms on a 2.5GHz core. The Hardware monitor reports a cache miss ratio of 6% for the application. Main Memory access takes on average 80 ns, while cache access has a latency of 800ps. Given that all ALU instructions are executed effectively in a single clock cycle.

T = total execution time

T_{cycle} = time for a single processor cycle = 2.5ms

I_{count} = total number of instructions = 1000000

I_{ALU} = number of ALU instructions (e.g. register – register)

I_{MEM} = number of memory access instructions (e.g. load, store)

CPI = average cycles per instructions = 2.5GHz

CPI_{ALU} = average cycles per ALU instructions = 1

CPI_{MEM} = average cycles per memory instruction = 5.6

r_{miss} = cache miss rate = 6%

r_{hit} = cache hit rate = 94%

$CPI_{MEM-MISS}$ = cycles per cache miss = 80ns

$CPI_{MEM-HIT}$ = cycles per cache hit = 800ps = 0.8ns

M_{ALU} = instruction mix for ALU instructions = 0.67

M_{MEM} = instruction mix for memory access instruction

$$CPI_{MEM} = CPI_{MEM-HIT} + r_{miss} * CPI_{MEM-MISS}$$

$$= 0.8 + 0.06 * 80$$

$$= 0.8 + 4.8$$

$$= 5.6$$

$$M_{ALU} + M_{MEM} = 1$$

$$M_{MEM} = 1 - M_{ALU}$$

$$CPI = (M_{ALU} * CPI_{ALU}) + (M_{MEM} * CPI_{MEM})$$

$$2.5 = (M_{ALU} * 1) + ((1 - M_{ALU}) * 5.6)$$

$$2.5 = M_{ALU} + 5.6 - 5.6M_{ALU}$$

$$4.6M_{ALU} = 5.6 - 2.5$$

$$M_{ALU} = 3.1 / 4.6 = 0.67$$

a. Fraction of ALU instructions

$$I_{ALU} = M_{ALU} * I_{COUNT} = 0.67 * 1000000 = 670000$$

b. Cache Size = 16KB

To decrease execution time by 50%, cache size should be increased and miss rate therefore decreased. We know, miss rate decreases by 1% upon doubling cache size => cache size increases to the power 2^1

Since they are directly proportional,

$$T_1 / (0.5 * T_1) \sim r_{miss} / r_{miss-new}$$

$$r_{miss-new} = \sim r_{miss} * 0.5$$

$$r_{miss-new} = \sim 6\% * 0.5 = 3\%$$

since doubling cache size decreases miss rate by 1%, to decrease miss rate TO 3%, doing some math, the size should double almost 3 times (6% to 3%) = 2^3

c. Applying the final formula to calculate the program runtime if all accessed data fits in cache, meaning run rate is 3%

$$T = I_{count} * [(M_{ALU} * CPI_{ALU}) + M_{MEM} * (CPI_{MEM-HIT} + r_{miss} * CPI_{MEM-MISS})] * T_{cycle}$$

$$T = 1000000 * [(0.67 * 1) + 0.33 * (3.2)] * 2.5$$

$$T = 1000000 * 1.726 * 2.5$$

$$T = 4315000 = 4315s$$

Programming Questions

4. I used the math directly from the question, the program runs but somehow my values are going VERY NAN. I tried debugging and the values are just very small and close to 0. The interesting thing is that I added a condition that simulation ends when all values are 12 or less, and still somehow simulation ends with a matrix full of "NAN" values.
- So instead of 100 by 100 matrix, I am doing for 20X20 matrix to avoid too small values. And make output seem like it is broken

Screenshot :

[illegible]


```

1.425  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
-nan 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 -3.98844 0 58.8897 -3.98844 0 0 0 0 0 0 0 0 0 0
0 0 0 -3.98844 0 -3.98844 -3.98844 -7.29891 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 -nan 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Number of Iterations done: 2

```

- I used the `pragma omp parallel shared default with reduction` on the value of 'y'. The value changed clearly. Y value depends on the previous y value so yes, I don't know how you parallelize without breaking the calculations.

```

● punyajamishra@LAPTOP-786BHK21:~/4350$ g++ a2q5.cpp -o a2q5_output
● punyajamishra@LAPTOP-786BHK21:~/4350$ ./a2q5_output
The value of y at x is : 1.10364
● punyajamishra@LAPTOP-786BHK21:~/4350$ export OMP_NUM_THREADS=2
● punyajamishra@LAPTOP-786BHK21:~/4350$ g++ a2q5.cpp -o a2q5_output -fopenmp
● punyajamishra@LAPTOP-786BHK21:~/4350$ ./a2q5_output
The value of y at x is : 1.81959
● punyajamishra@LAPTOP-786BHK21:~/4350$ export OMP_NUM_THREADS=4
● punyajamishra@LAPTOP-786BHK21:~/4350$ g++ a2q5.cpp -o a2q5_output -fopenmp
● punyajamishra@LAPTOP-786BHK21:~/4350$ ./a2q5_output
The value of y at x is : 1.81625
● punyajamishra@LAPTOP-786BHK21:~/4350$ export OMP_NUM_THREADS=8
● punyajamishra@LAPTOP-786BHK21:~/4350$ g++ a2q5.cpp -o a2q5_output -fopenmp
● punyajamishra@LAPTOP-786BHK21:~/4350$ ./a2q5_output
The value of y at x is : 1.72476
○ punyajamishra@LAPTOP-786BHK21:~/4350$

```