



From my experiments, it can be seen that if the matrix is small (approximately less than $10^5 \times 10^5$) computing with CPU will finish faster and vice versa. Because of their architecture and design, GPUs (Graphics Processing Units) can outperform CPUs (Central Processor Units) in huge matrix multiplication. CPUs are built to handle complicated decision-making and sequential processing activities. They have a few cores, each geared for swiftly processing a small number of tasks. GPUs, on the other hand, contain hundreds or thousands of cores that are geared for completing numerous basic operations at the same time. When it comes to matrix multiplication, the algorithm may be divided into many little, easy tasks, each of which can be allocated to a different GPU core. This enables the GPU to do numerous processes concurrently, whereas a CPU would perform the same activities sequentially.

Therefore, the combination of parallel processing and quicker memory access makes GPUs excellent for large-data workloads like matrix multiplication.

Appendix (Example RUN):

Experiments

```
In [25]: np.random.seed(1)
n = 15000
print("Matrix size:", n,"x",n)
x = np.array(np.random.randn(n,n), dtype = np.float32)
y = np.array(np.random.randn(n,n), dtype = np.float32)

start = time.time();
matrix_multi = tf.matmul(x,y) # multiply matrix parallely using GPU
print("GPU:", time.time() - start, "sec")

start = time.time();
x.dot(y); # multiply matrix parallely using CPU
print("CPU:", time.time() - start, "sec")

Matrix size: 15000 x 15000

2023-04-01 21:53:02.164718: W tensorflow/tsl/framework/cpu_allocator_impl.cc:83] Allocation of 900000000 exceeds 1
0% of free system memory.
2023-04-01 21:53:02.813512: W tensorflow/tsl/framework/cpu_allocator_impl.cc:83] Allocation of 900000000 exceeds 1
0% of free system memory.
2023-04-01 21:53:03.461781: W tensorflow/tsl/framework/cpu_allocator_impl.cc:83] Allocation of 900000000 exceeds 1
0% of free system memory.

GPU: 56.84810400009155 sec
CPU: 53.06075859069824 sec
```

Copy the above cell and try to vary the size of the matrix to see the trend of execution time from each device.