GPU IMPLEMENTATION (WITH TENSOR CORES)

-

-

DOUBLE-dataype

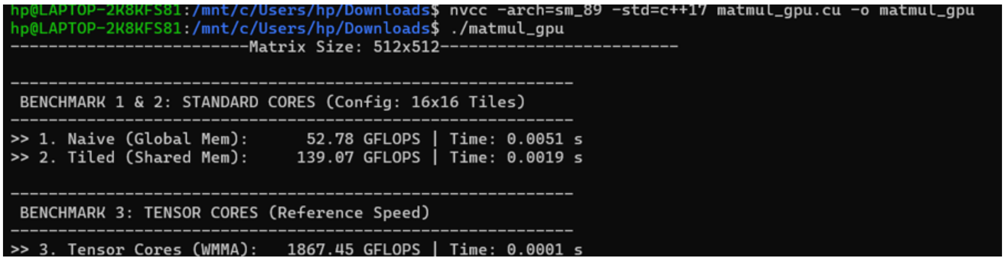
-

512x512

-

Block size-16

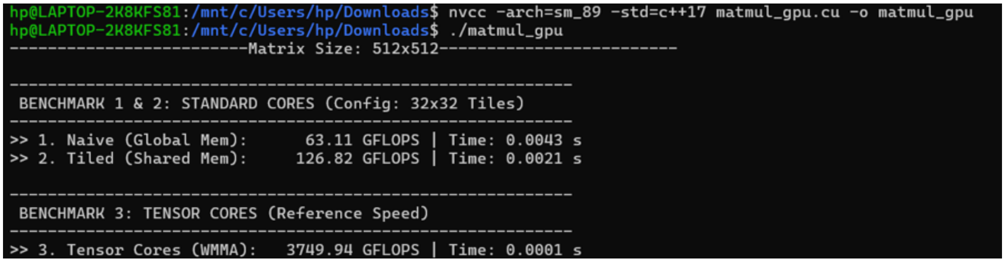
-



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
-------------------------Matrix Size: 512x512-------------------------

-----------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 16x16 Tiles)
-----------------------------------------------------------
>> 1. Naive (Global Mem):      52.78 GFLOPS | Time: 0.0051 s
>> 2. Tiled (Shared Mem):     139.07 GFLOPS | Time: 0.0019 s

-----------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
-----------------------------------------------------------
>> 3. Tensor Cores (WMMA):   1867.45 GFLOPS | Time: 0.0001 s
```

-

Block size-32



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
-------------------------Matrix Size: 512x512-------------------------

-----------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 32x32 Tiles)
-----------------------------------------------------------
>> 1. Naive (Global Mem):      63.11 GFLOPS | Time: 0.0043 s
>> 2. Tiled (Shared Mem):     126.82 GFLOPS | Time: 0.0021 s

-----------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
-----------------------------------------------------------
>> 3. Tensor Cores (WMMA):   3749.94 GFLOPS | Time: 0.0001 s
```

-

-

-

-

-

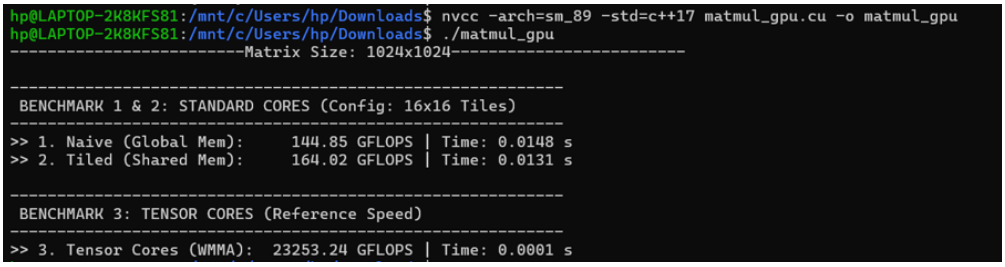
-

1024x1024

-

Block size-16

-

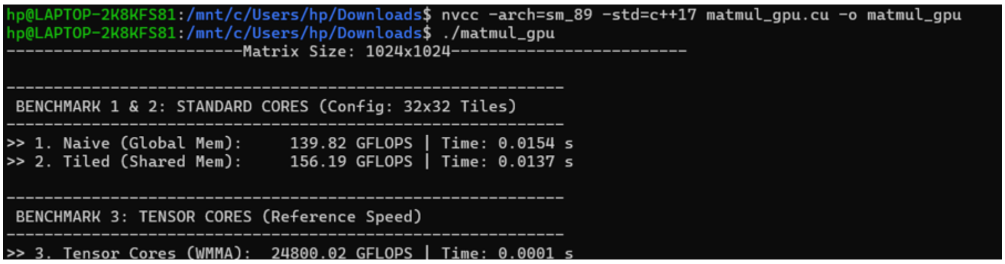


```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
-------------------------Matrix Size: 1024x1024-------------------------

-----------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 16x16 Tiles)
-----------------------------------------------------------
>> 1. Naive (Global Mem):     144.85 GFLOPS | Time: 0.0148 s
>> 2. Tiled (Shared Mem):     164.02 GFLOPS | Time: 0.0131 s

-----------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
-----------------------------------------------------------
>> 3. Tensor Cores (WMMA):  23253.24 GFLOPS | Time: 0.0001 s
```

-

Block size-32

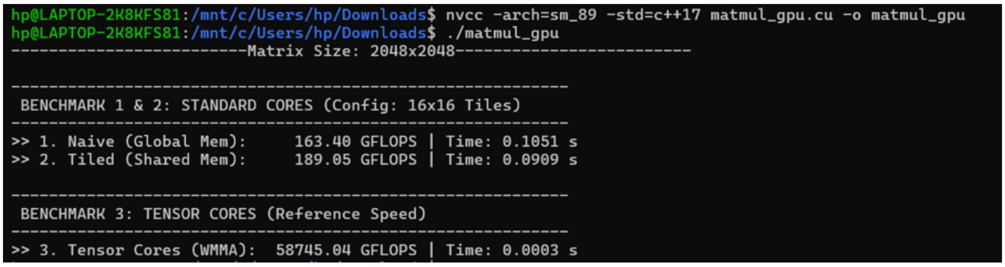
-



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
------------------------Matrix Size: 1024x1024------------------------

------------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 32x32 Tiles)
------------------------------------------------------------
>> 1. Naive (Global Mem):      139.82 GFLOPS | Time: 0.0154 s
>> 2. Tiled (Shared Mem):      156.19 GFLOPS | Time: 0.0137 s

------------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
------------------------------------------------------------
>> 3. Tensor Cores (WMMA):   24800.02 GFLOPS | Time: 0.0001 s
```

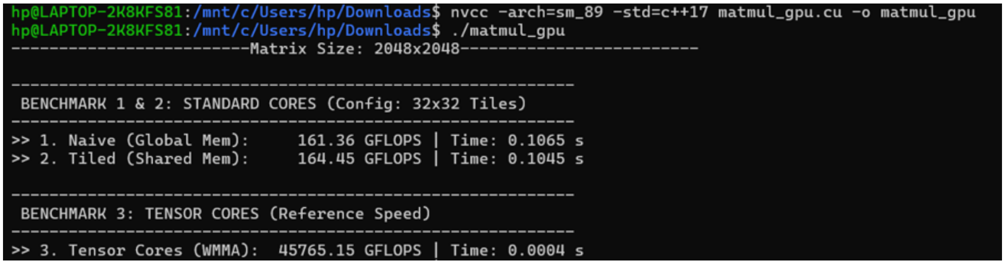
2048x2048

Block size-16



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
------------------------Matrix Size: 2048x2048------------------------

------------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 16x16 Tiles)
------------------------------------------------------------
>> 1. Naive (Global Mem):      163.40 GFLOPS | Time: 0.1051 s
>> 2. Tiled (Shared Mem):      189.05 GFLOPS | Time: 0.0909 s

------------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
------------------------------------------------------------
>> 3. Tensor Cores (WMMA):   58745.04 GFLOPS | Time: 0.0003 s
```

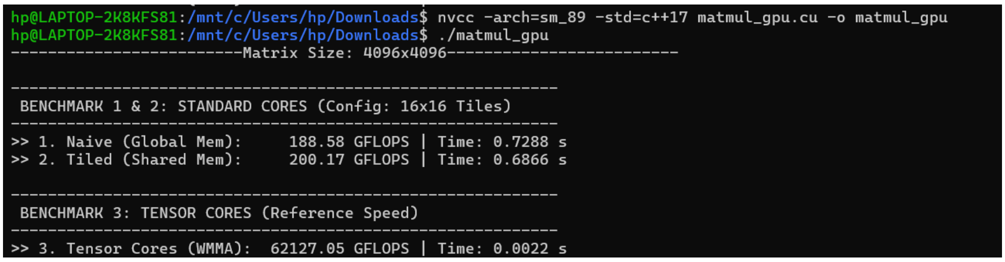
BS=32



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
------------------------Matrix Size: 2048x2048------------------------

------------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 32x32 Tiles)
------------------------------------------------------------
>> 1. Naive (Global Mem):      161.36 GFLOPS | Time: 0.1065 s
>> 2. Tiled (Shared Mem):      164.45 GFLOPS | Time: 0.1045 s

------------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
------------------------------------------------------------
>> 3. Tensor Cores (WMMA):   45765.15 GFLOPS | Time: 0.0004 s
```

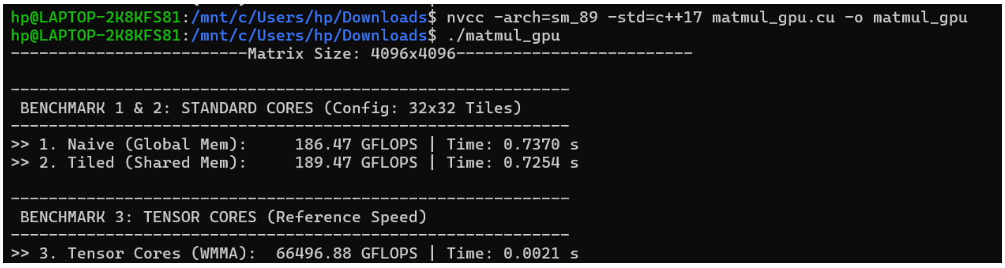
4096x4096

bs=16



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
------------------------Matrix Size: 4096x4096------------------------

------------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 16x16 Tiles)
------------------------------------------------------------
>> 1. Naive (Global Mem):      188.58 GFLOPS | Time: 0.7288 s
>> 2. Tiled (Shared Mem):      200.17 GFLOPS | Time: 0.6866 s

------------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
------------------------------------------------------------
>> 3. Tensor Cores (WMMA):   62127.05 GFLOPS | Time: 0.0022 s
```

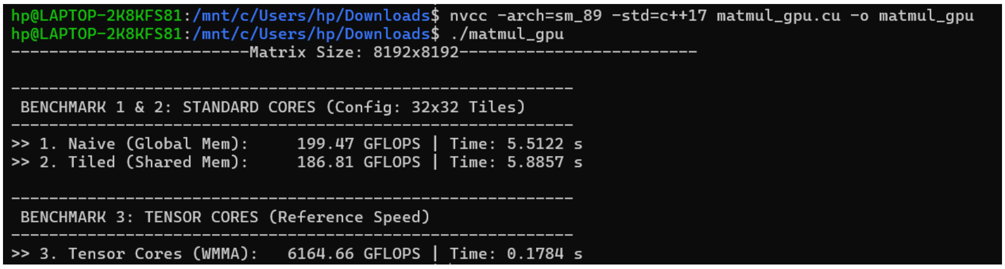
bs=32



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
------------------------Matrix Size: 4096x4096------------------------

---------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 32x32 Tiles)
---------------------------------------------------------
>> 1. Naive (Global Mem):     186.47 GFLOPS | Time: 0.7370 s
>> 2. Tiled (Shared Mem):     189.47 GFLOPS | Time: 0.7254 s

---------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
---------------------------------------------------------
>> 3. Tensor Cores (WMMA):   66496.88 GFLOPS | Time: 0.0021 s
```

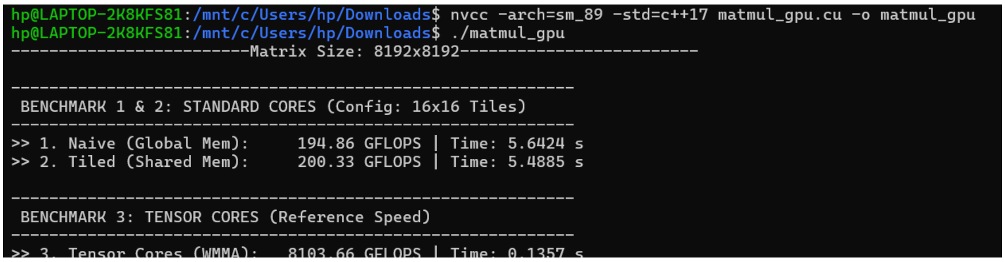
8192x8192

Block size-32



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
------------------------Matrix Size: 8192x8192------------------------

---------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 32x32 Tiles)
---------------------------------------------------------
>> 1. Naive (Global Mem):     199.47 GFLOPS | Time: 5.5122 s
>> 2. Tiled (Shared Mem):     186.81 GFLOPS | Time: 5.8857 s

---------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
---------------------------------------------------------
>> 3. Tensor Cores (WMMA):   6164.66 GFLOPS | Time: 0.1784 s
```

Block size-16



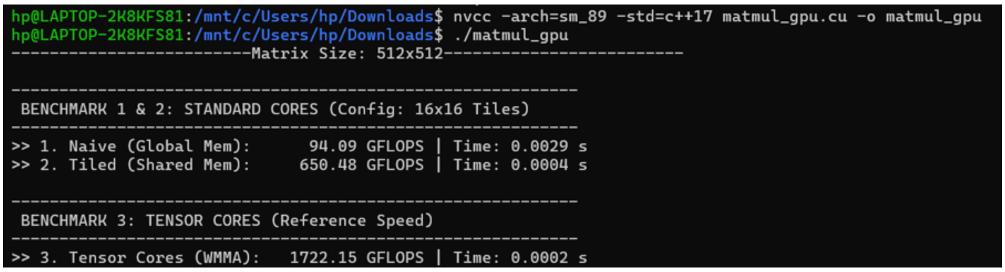
```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
------------------------Matrix Size: 8192x8192------------------------

---------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 16x16 Tiles)
---------------------------------------------------------
>> 1. Naive (Global Mem):     194.86 GFLOPS | Time: 5.6424 s
>> 2. Tiled (Shared Mem):     200.33 GFLOPS | Time: 5.4885 s

---------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
---------------------------------------------------------
>> 3. Tensor Cores (WMMA):   8103.66 GFLOPS | Time: 0.1357 s
```

GPU IMPLEMENTATION (WITH TENSOR CORES)

Float-dataype

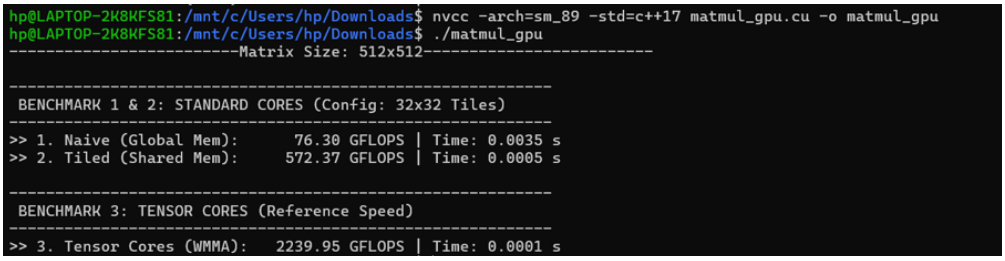
512x512

Block size-16



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
-------------------------Matrix Size: 512x512-------------------------

-----------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 16x16 Tiles)
-----------------------------------------------------------
>> 1. Naive (Global Mem):      94.09 GFLOPS | Time: 0.0029 s
>> 2. Tiled (Shared Mem):     650.48 GFLOPS | Time: 0.0004 s

-----------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
-----------------------------------------------------------
>> 3. Tensor Cores (WMMA):   1722.15 GFLOPS | Time: 0.0002 s
```

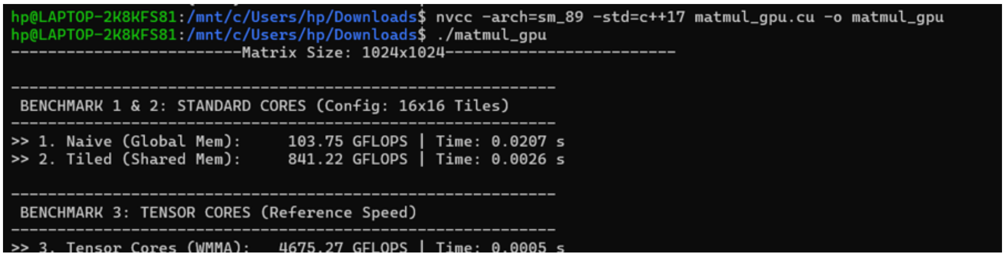
Block size-32



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
-------------------------Matrix Size: 512x512-------------------------

-----------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 32x32 Tiles)
-----------------------------------------------------------
>> 1. Naive (Global Mem):      76.30 GFLOPS | Time: 0.0035 s
>> 2. Tiled (Shared Mem):     572.37 GFLOPS | Time: 0.0005 s

-----------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
-----------------------------------------------------------
>> 3. Tensor Cores (WMMA):   2239.95 GFLOPS | Time: 0.0001 s
```

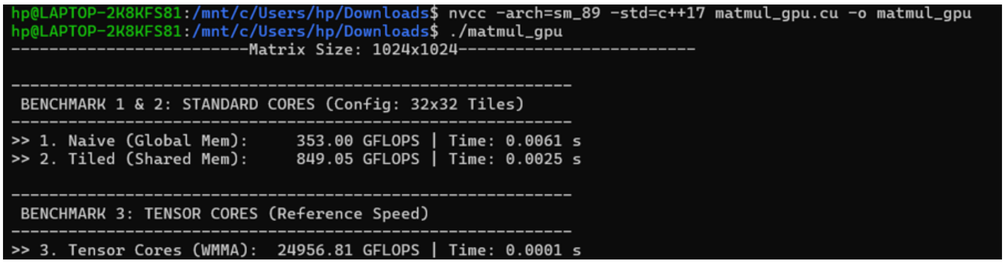
1024x1024

Block size-16



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
-------------------------Matrix Size: 1024x1024-------------------------

-----------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 16x16 Tiles)
-----------------------------------------------------------
>> 1. Naive (Global Mem):     103.75 GFLOPS | Time: 0.0207 s
>> 2. Tiled (Shared Mem):     841.22 GFLOPS | Time: 0.0026 s

-----------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
-----------------------------------------------------------
>> 3. Tensor Cores (WMMA):   4675.27 GFLOPS | Time: 0.0005 s
```

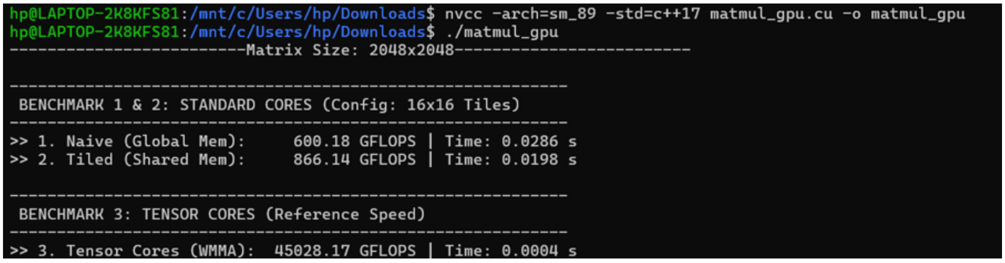
Block size-32



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
-------------------------Matrix Size: 1024x1024-------------------------

------------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 32x32 Tiles)
------------------------------------------------------------
>> 1. Naive (Global Mem):     353.00 GFLOPS | Time: 0.0061 s
>> 2. Tiled (Shared Mem):     849.05 GFLOPS | Time: 0.0025 s

------------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
------------------------------------------------------------
>> 3. Tensor Cores (WMMA):  24956.81 GFLOPS | Time: 0.0001 s
```

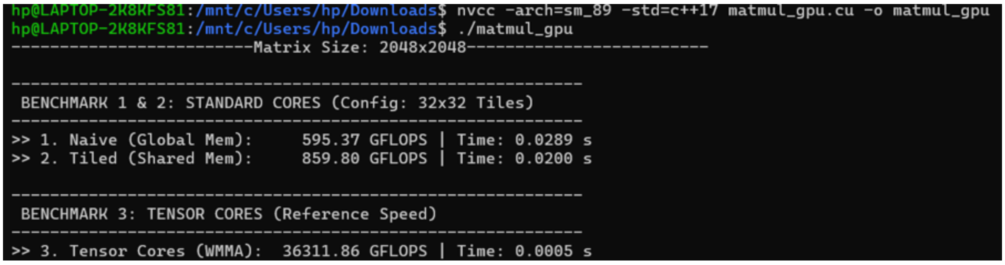
2048x2048

Block size-16



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
-------------------------Matrix Size: 2048x2048-------------------------

------------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 16x16 Tiles)
------------------------------------------------------------
>> 1. Naive (Global Mem):     600.18 GFLOPS | Time: 0.0286 s
>> 2. Tiled (Shared Mem):     866.14 GFLOPS | Time: 0.0198 s

------------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
------------------------------------------------------------
>> 3. Tensor Cores (WMMA):  45028.17 GFLOPS | Time: 0.0004 s
```

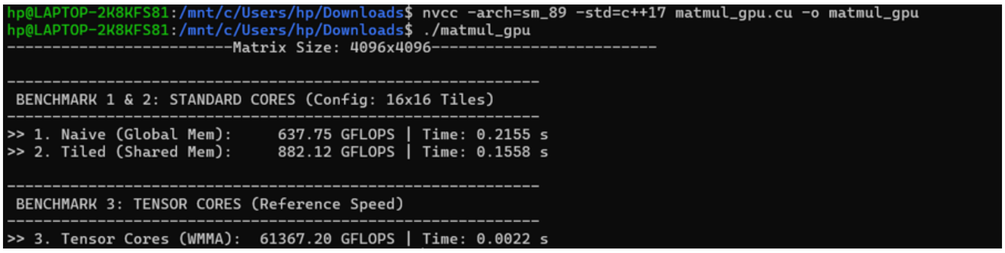
BS=32



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
-------------------------Matrix Size: 2048x2048-------------------------

------------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 32x32 Tiles)
------------------------------------------------------------
>> 1. Naive (Global Mem):     595.37 GFLOPS | Time: 0.0289 s
>> 2. Tiled (Shared Mem):     859.80 GFLOPS | Time: 0.0200 s

------------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
------------------------------------------------------------
>> 3. Tensor Cores (WMMA):  36311.86 GFLOPS | Time: 0.0005 s
```

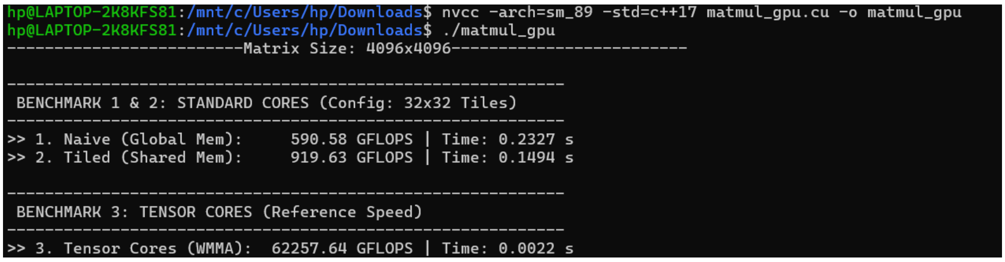
4096x4096

bs=16



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
-------------------------Matrix Size: 4096x4096-------------------------

------------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 16x16 Tiles)
------------------------------------------------------------
>> 1. Naive (Global Mem):     637.75 GFLOPS | Time: 0.2155 s
>> 2. Tiled (Shared Mem):     882.12 GFLOPS | Time: 0.1558 s

------------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
------------------------------------------------------------
>> 3. Tensor Cores (WMMA):  61367.20 GFLOPS | Time: 0.0022 s
```

bs=32

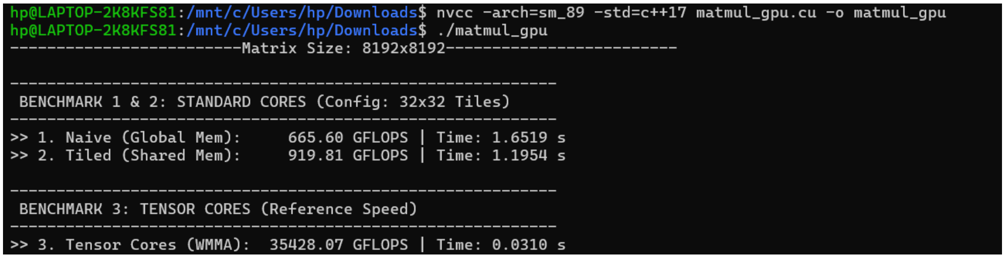


```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
-------------------------Matrix Size: 4096x4096-------------------------

------------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 32x32 Tiles)
------------------------------------------------------------
>> 1. Naive (Global Mem):     590.58 GFLOPS | Time: 0.2327 s
>> 2. Tiled (Shared Mem):     919.63 GFLOPS | Time: 0.1494 s

------------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
------------------------------------------------------------
>> 3. Tensor Cores (WMMA):  62257.64 GFLOPS | Time: 0.0022 s
```

8192x8192

-

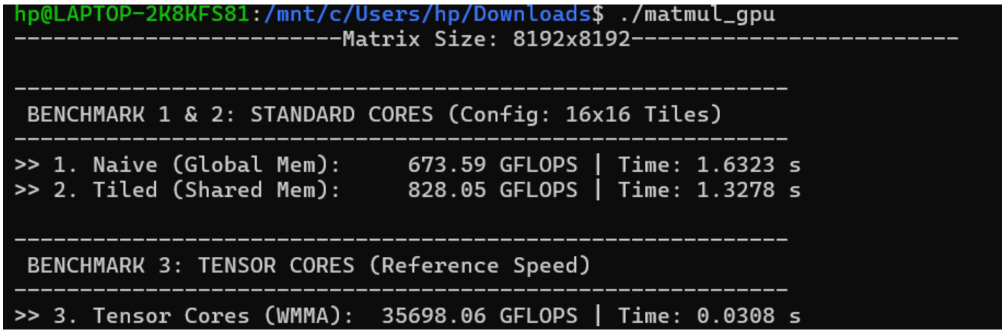
Block size-32



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ nvcc -arch=sm_89 -std=c++17 matmul_gpu.cu -o matmul_gpu
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
------------------------Matrix Size: 8192x8192------------------------

-----------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 32x32 Tiles)
-----------------------------------------------------------
>> 1. Naive (Global Mem):     665.60 GFLOPS | Time: 1.6519 s
>> 2. Tiled (Shared Mem):     919.81 GFLOPS | Time: 1.1954 s

-----------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
-----------------------------------------------------------
>> 3. Tensor Cores (WMMA):  35428.07 GFLOPS | Time: 0.0310 s
```

Block size-16

-



```
hp@LAPTOP-2K8KFS81:/mnt/c/Users/hp/Downloads$ ./matmul_gpu
------------------------Matrix Size: 8192x8192------------------------

-----------------------------------------------------------
 BENCHMARK 1 & 2: STANDARD CORES (Config: 16x16 Tiles)
-----------------------------------------------------------
>> 1. Naive (Global Mem):     673.59 GFLOPS | Time: 1.6323 s
>> 2. Tiled (Shared Mem):     828.05 GFLOPS | Time: 1.3278 s

-----------------------------------------------------------
 BENCHMARK 3: TENSOR CORES (Reference Speed)
-----------------------------------------------------------
>> 3. Tensor Cores (WMMA):  35698.06 GFLOPS | Time: 0.0308 s
```

-

-

-

-