

Evaluating Tensor cores

John Carlsson, Cyprien Toulon de Courtex
Department of Computer Science
University of Salerno
Italy

Abstract— This project was aimed at evaluating Tensor cores in CUDA through experimnts to get a better understanding off high performance computing through different methods. With the ultimate goal to compare the time to calculate a matrix multiplication addition. At the start of the project, we were fairly certain that the fastest algorithm would be the Tensor cores, followed by CUDA with shared memory,CUDA global memory, and at the bottom, an optimized CPU version. However, this was not the case. The results show that CUDA shared memory is the fastest with global memory just behind. For small matrices tensors appear superior.

I. INTRODUCTION

Over the past decade, the increasing demand for deep learning and AI applications has fueled the need for highly efficient Tensor operations. Traditional GPU architectures are powerful but are often limited in their ability to fully exploit the potential of Tensor computations due to the reliance on higher-precision floating-point arithmetic. Tensor cores address this limitation by leveraging mixed-precision arithmetic, combining high throughput with reduced precision.

Tensor cores, introduced in NVIDIA's Volta architecture and further enhanced in subsequent architectures such as Turing and Ampere, have revolutionized the performance of Tensor computations on GPUs. These specialized hardware units offer significant speedups by providing native support for low-precision Tensor operations, specifically matrix multiplications and convolutions. The purpose of this project report is to present an evaluation of Tensor cores, exploring their capabilities compared to CPU and GPU based computation.

In this project, we aim to evaluate Tensor cores using the Matrix Multiplication Accumulate (MMA) with various workloads and compare the results to the CPU and GPU performances.

By conducting this comprehensive evaluation, we aim to provide insights into the capabilities and limitations of Tensor cores. Additionally, the findings from this study will contribute to the broader understanding of GPU acceleration techniques for Tensor operations and their potential impact on high-performance computing.

A. Motivation

The MMA operation plays a critical role in accelerating various computational tasks. It involves three matrices and is fundamental within many computational domains, such as deep learning, scientific simulations, and image processing. However, performing MMA efficiently and at scale is difficult due to the intensity of the operation.

The difficulties that arise when trying to implement a fast version of MMA are due to a combination of factors including memory access patterns, data dependencies, computational intensity, matrix sizes, hardware limitations, and optimization trade-offs.

The objective of this paper is to provide valuable insights for optimizing MMA computations and facilitate the adoption of Tensor cores as a powerful tool for accelerating MMA operations in comparison to CPU and GPU. Our contributions will be as follows:

- We will measure the speedup achieved by Tensor cores measured in actual seconds and explore how performance scales. with the size of the matrices.
- We will compare the performance of Tensor cores with CPU and GPU implementations of MMA using different matrix sizes and precision settings.
- We will analyze the impact of data precision on the performance of Tensor cores and investigate the trade-offs between accuracy and computation time.
- We will explore optimization techniques for maximizing the utilization of Tensor cores and achieving optimal performance.
- We will provide insights into the limitations and challenges of Tensor cores, including memory constraints, data dependencies, and scalability issues.

II. BACKGROUND

A. Tensor cores

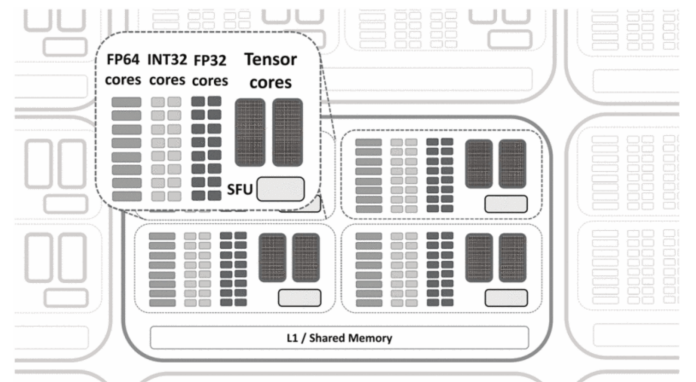


Fig. 1. SM architecture.[1]

Tensor cores are specialized hardware units introduced in NVIDIA GPUs that are designed to accelerate matrix

computations, specifically matrix multiplications and convolutions. They leverage mixed-precision arithmetic, combining high throughput with reduced precision to achieve significant speedups in Tensor operations.

The key feature of Tensor cores is their ability to perform mixed-precision operations using half-precision floating-point (FP16) data types. By using FP16, Tensor cores can process a larger number of elements simultaneously, leading to improved throughput at the cost of accuracy.[2]

Tensor cores operate on 4x4 matrix tiles, performing matrix multiplication and accumulation (MMA) operations. These operations are highly parallel and can be efficiently executed on Tensor cores. By exploiting the parallelism of Tensor cores and their ability to process multiple elements simultaneously, significant performance gains can be achieved compared to traditional GPU implementations.

The Tensor operations are warp wise, this means that each Tensor instructions must work on the same data within the same warp.

B. Matrix Multiplication Accumulate (MMA)

Matrix Multiplication Accumulate (MMA) is a key operation in many computational tasks, including deep learning, scientific simulations, and image processing. It involves three matrices: two input matrices (A and B) and an output matrix (C). The operation computes the matrix product of A and B and accumulates the result into C.

Mathematically, the MMA operation can be defined as follows:

$$C \leftarrow \alpha \cdot A \cdot B + \beta \cdot C$$

where A is an $m \times k$ matrix, B is a $k \times N$ matrix, C is an $m \times N$ matrix, and α and β are scalar coefficients.

The MMA operation is computationally intensive and can benefit greatly from hardware acceleration. Tensor cores, with their ability to efficiently perform MMA operations, offer a promising solution for accelerating this operation.

III. CODE

In this section we present the main part of the code. The code shown is the computational part. Since this is deemed most important. We show here the CUDA code and the Tensor code.

A. CPU code

We implement a CPU version of the MMA operation to have a reference to verify our computation results. The CPU code is an implement of the MMA operation that uses the following optimizations:

- 1) Tiling
- 2) Tarallelisation using OpenMP
- 3) Compiler vectorisation for AVX512 using OpenMP
- 4) Compiler optimization flags

B. CUDA code

We implemented multiple versions of the matrix multiplication and addition with CUDA to test out the effect of multiple optimizations techniques. For each implementation, the code is heavily simplified to keep the essence. For full code, see appendix.

1) *Basic MMA in CUDA*: The MMA implement in CUDA is very basic as it uses the base formula of a matrix multiplication and addition and assign one thread per each element of the output matrix.[3]

2) *MMA with shared Memory*: This MMA implement uses shared memory to optimize global memory acces count. The multiplication operation is done using tiling with the size of the tile matching the size of the block. For each tile, within a same block, each threads load a single element of the tile to the shared memory.

```
__global__ void MMACudaShared(float *out,
                              float *A, float *B, float *c, int n)
{
    sum = c[...];
    for (tileNum = 0; tileNum < n/TILE_SIZE;
         tileNum++){
        a_shared[...] = A[...];
        b_shared[...] = B[...];
        __syncthreads();
        for (int k = 0; k < TILE_SIZE; k++)
            {sum += a_shared[...] * b_shared[...];}
    }
    out = sum;
}
```

3) *Mixed precision MMA*: The CUDA API provides the possibility to work with floating point numbers of half the size of regular 32 bits floats. Mixed computation works with A and B being half precision floats matrices and C and The output are regular float matrices. We made two versions of this MMA Mixed precision operation based on the MMA shared memory implementation. One version convert halves to float before doing the computation to see the effect on bandwidth and the other does the conversion after.

```
__global__ void MMACudaMixedA(float *out,
                              half *A, half *B, float *c, int n)
{
    ...
    for (int k = 0; k < TILE_SIZE; k++)
        sum += (float)a_shared[...] *
               (float)b_shared[...]; }
    ...
}

__global__ void MMACudaMixedB(float *out,
                              half *A, half *B, float *c, int n)
{
    ...
```

```

    for (int k = 0; k < TILE_SIZE; k++)
        sum += __hfma( a_shared[...],
                       b_shared[...], sum ); }
    ...
}

```

4) *Thread coarsening Mixed precision MMA*: The space gained in the shared memory is not enough to double the tile size of the shared memory. We wanted to maximize the usage of the shared memory to reduce the amount of time each thread has to wait for the memory to arrive to the shared memory between each operation. This operation is based on Mixed precision implementation. This is done by making the threads working on computing multiple elements of the output matrices (see figure 2). In this case NUMTHREADS is the amount of elements each thread has to compute.

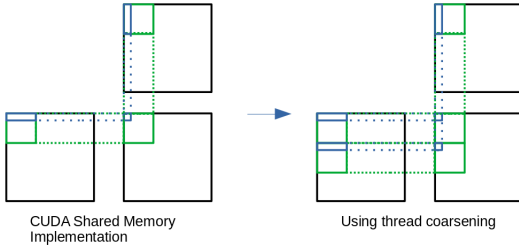


Fig. 2. Performance Comparison total time with deallocation

```

__global__ void MMACoarsening(float *out,
                             half *A, half *B, float *c, int n)
{
    Out accs[NUMTHREADS];
    for (subthread = 0; subthread <
        NUMTHREADS; subthread++)
        accs[subthread] = c[(row + TILE_SIZE *
            subthread)*n + col];
    for (tileNum = 0; tileNum < n/TILE_SIZE;
        tileNum++){
        for (subthread = 0; subthread <
            NUMTHREADS; subthread++)
            a_shared[subthread][...] = A[...];
        b_shared[...] = B[...];
        __syncthreads();
        for (int k = 0; k < TILE_SIZE; k++)
            for (subthread = 0; subthread <
                NUMTHREADS; subthread++)
                {suaccs[subthread]m +=
                    (float)a_shared[subthread][...]
                    * (float)b_shared[...]; }
    }
    for (subthread = 0; subthread <
        NUMTHREADS; subthread++)
        out[...] = accs[subthread];
}

```

C. Tensor code

The Tensor code is implemented using the WMMA, Warped matrix multiplication addition, library. We are working in our case with 16*16*16 fragments that are for row major matrices.

```

using namespace wmma
__global__ void mat_mul_add_tensor(half *a,
    half *b, float *C, float *d, int N){
    // Declare the fragments
    fragment<matrix_a, ...> a_frag;
    fragment<matrix_b, ...> b_frag;
    fragment<wmma::accumulator, ..., float>
        c_frag;
    load_matrix_sync(c_frag, C, N,
        mem_row_major);
    // Loop over k
    for (int i = 0; i < N; i += WMMA_K){
        load_matrix_sync(a_frag, a, N);
        load_matrix_sync(b_frag, b, N);
        // Perform the matrix multiplication
        mma_sync(c_frag, a_frag, b_frag, c_frag);
    }
    store_matrix_sync(d, c_frag, N,
        mem_row_major);
}

```

We tried to shared memory and an implementation with the cuBLASs library but we had terrible accuracy results and did not had time to find a solution to fix them.

IV. EXPERIMENTAL SETUP

In this section, we describe the experimental setup used to evaluate Tensor cores and compare their performance with CPU and GPU implementations of MMA. We did a multitude of tests with different parameters to see how performance scale over matrix size and how conducting many runs consequently affects performance over time.

We used square matrices with the sizes of 32, 64, 128, 256, 512, 1024, 2048, 4096 and 8192.

we conducted single comparison runs to measure the computation error and then 200 runs to get a better average for time. Each implement works on the same input matrices that are made of float elements that are initialised with random values between 0 and 1 and are not allocated in an alligned way. The cpu implement will allign the matrices before the computation and each mixed precision implementation will do a CPU pass do the conversion they need before sending the data to the GPU, these conversions and memory managements are not mesured in the computation measurements, however it is taken in account for the total operation execution time.

A. Hardware Configuration

The experiments were conducted on a system equipped with an NVIDIA GPU that supports Tensor cores. The specific GPU model is seen below. The GPU belongs to a group of compute capability 7.5.

TABLE I
GPU SPECIFICATIONS[4]

| | |
|--------------------|------------------------|
| GPU Model | NVIDIA Quadro 4000 RTX |
| CUDA cores | 2304 |
| Tensor cores | 288 |
| Memory interface | 256-bit |
| CUDA version | 11.8 |
| Compute capability | 7.5 |
| shared memory size | 64 kb |

B. Software Configuration

The experiments were conducted using the following software tools and frameworks:

- CUDA Toolkit version 11.8
- WMMA
- C++17 compiled with g++ 11.3.0

C. implementations parameters

Each of our CUDA implementations, except for the thread coarsening one uses threadblocks of $32 * 32$. The thread coarsening version is executed twice with a different value (2 and 3) for NUMTHREADS. This mean that the thread coarsening version will be allocating twice and a third of the amount of threads by the other CUDA implementations. The Tensor implementation work block that handle 4 by 4 matrices of 16 by 16 elements, therefore the block size is 128 by 4.

The MMACudaMixedB implementation is not present in the results as it had some major issues that we could not correct by the time we made the benchmarkings.

D. Benchmarking Methodology

To evaluate the performance of Tensor cores, we performed a series of experiments using different matrix sizes and precision settings. For each experiment, we measured the execution time of the MMA operation for CPU, GPU (without Tensor cores), and GPU (with Tensor cores) implementations.

We varied the matrix sizes from small to large to analyze the impact of matrix dimensions on performance. Additionally, we tested different precision settings, including full precision (FP32) and mixed precision (FP16).

To ensure accurate measurements, we repeated each experiment multiple times and calculated the average execution time. We also recorded the peak memory usage during each experiment to analyze the memory requirements of Tensor cores.

V. RESULTS AND ANALYSIS

In this section, we present the results of our experiments and analyze the performance of Tensor cores compared to CPU and GPU implementations of MMA.

A. Performance Comparison

First we look at the computational time for defferent kinds of matrices, the results are plotted on a log scale on both axes (See figure 5). And the result show clearly that In terms of pure computation time, the CPU implementation is

inferior to the other implementations even with optimizations. The Tensor implementation seem to be the one with the fastest computation time.

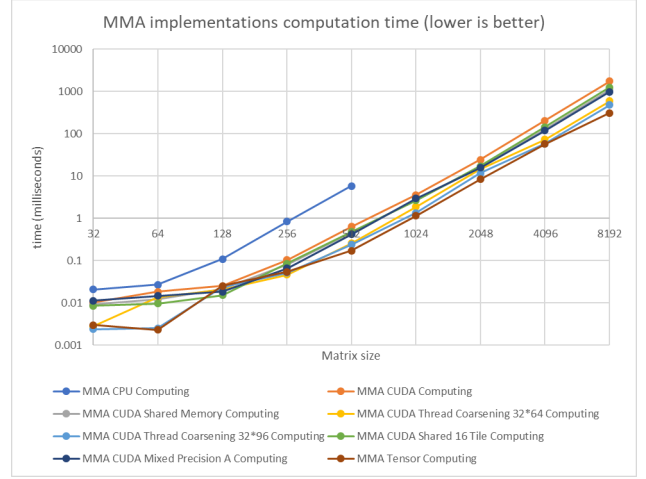


Fig. 3. Performance Comparison

The second graph shows relative speedup to CPU of the execution of full MMA operation. This include the memory allocation, the memory movement, and for some versions, the conversion of the data from one type to another, requiered by every mixed precision implementations. (See figure 4) We can see that under a matrix size of 64, the CPU implementation is faster than all the others and that it is not until a size of 256 that every otimeplementaionher implementation beat CPU computation. We can also see that every mixed precision operations are slower to run. This is probably caused by our method of converting the memory from one type to another.

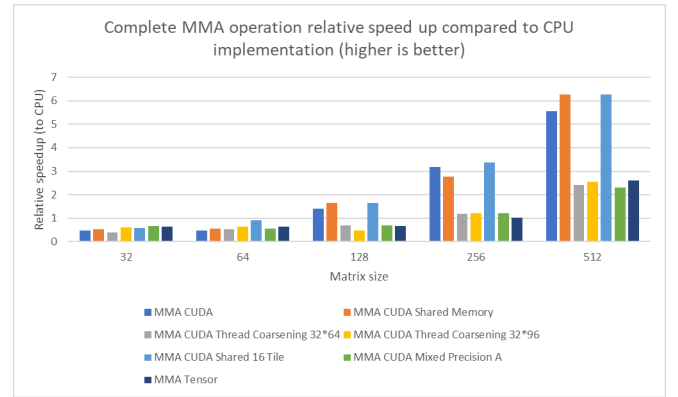


Fig. 4. Performance Comparison total time with deallocation

Now we look at the relative computational speedup compared to a CUDA implementation (See figure 5). We chose to compare to the CUDA with global memory since the performance is closer to the other implementations and a comparison to CPU would be irrelevant since all of the implementations outscale the CPU by large. We see here that Tensor cores without any specific optimization is the fastest implementation

for all matrix sizes above 256 by 256. CUDA implementation with thread coarsening (32*64 tiling) represent a huge speedup compared to the simpler CUDA implementations, the CUDA implementation with thread coarsening (32*96 tiling) also improves the performances but, the gap between 32*64 is not as big as the gain 32*64 offers. Thread coarsening has the goal of saturation the memory bandwidth but has the effect of increasing the register pressure. We may have reached a limit. Still, the CUDA implementation with thread coarsening (32*96 tiling) is not that far behind from the Tensor core implementation. however as we will see in figure 6, the accuracy for smaller matrices is not good.

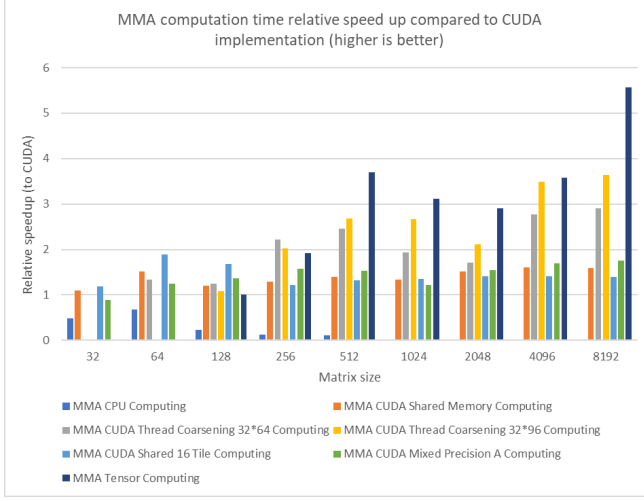


Fig. 5. Relative speedup to CUDA with global memory

B. Impact of Precision

To look at the precision of our implementations we compare to the CPU how much each value differ in percentage. We used this formula:

$$error = \frac{abs(OUT - CPU)}{CPU}$$

By looking at the figure 6 we can see that each thread coarsening implementations as well as the Tensor implementation have an exceptionally high error compared to the other implementations under a matrix size of 128*128. For smaller matrices the tiling is just not effective and will overlap and hence cause a big error. We also see that the error decreases with size. This may be attributed to the number used in our testing, which is random floating point numbers between 0 and 1.

The results show that using mixed precision (FP16) with Tensor cores can provide significant performance improvements compared to full precision (FP32). This is because Tensor cores can process a larger number of elements simultaneously in FP16, leading to higher throughput. This accuracy also improves with time for the matrix sizes. Using FP16 then introduces a degree of numerical approximation, affecting the accuracy of the computation. Therefore, the choice of

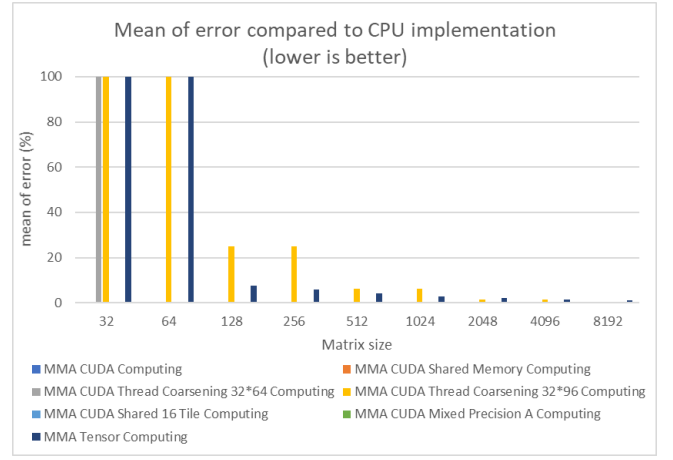


Fig. 6. Impact of Precision on mixed precision computation and Tensor cores Performance

computing should be carefully considered based on the specific requirements of the application.

VI. LIMITATIONS AND CHALLENGES

The main limitations faced in this project was the limited information regarding how to program Tensor cores. The cuBLASs library, which is an optimized version of MMA with Tensor cores, developed by NVIDIA is not open source, so there is no real way of obtaining it. This led to experimenting with different setups to achieve the fastest computation possible. Another challenge to achieve fast computation for larger matrices with tensors are the fact that the GPU only has 64Kb of memory for each work group. With the tensors, the sum of shared memory is then $N*N*(16+16)$. Which corresponds to the matrix size, and the byte size of the elements in the A and B matrix. This entails that the largest matrices that can fit this is of 32 x 32 elements of half precision.

VII. CONCLUSION

In this paper, we have explored the performance of NVIDIA Tensor cores for matrix multiplication and accumulation operations. We conducted experiments to compare the performance of Tensor cores with CPU and GPU implementations, considering different matrix sizes and precision settings.

From our result we can see that Tensor cores without any optimization have significantly better computational power than the other, optimized approaches. It is then our belief that to optimize the Tensor core version would decrease the computational time further. The total time comparison shows that allocation of memory has a significant impact on the time from input to output. Therefore using Tensor cores is optimal in problem where the same data is handled continuously, for example in convolutional neural network and image processing. The error seems to decrease with matrix size which can be attributed to what values we are using when testing. Using only integers, the error is 0.

REFERENCES

- [1] S. Markidis, S. W. D. Chien, E. Laure, I. B. Peng, and J. S. Vetter, "Nvidia tensor core programmability, performance & precision," in *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2018, pp. 522–531.
- [2] P. Blanchard, N. J. Higham, F. Lopez, T. Mary, and S. Pranesh, "Mixed precision block fused multiply-add: Error analysis and application to gpu tensor cores," *SIAM Journal on Scientific Computing*, vol. 42, no. 3, 2020.
- [3] Nvidia, "Cuda c++ programming guide," <https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html>, 2007, (Accessed: 03 June 2023).
- [4] —, "Volta Tuning guide," <https://docs.nvidia.com/cuda/volta-tuning-guide/index.html>, 2017, (Accessed: 31 May 2023).