Introduction

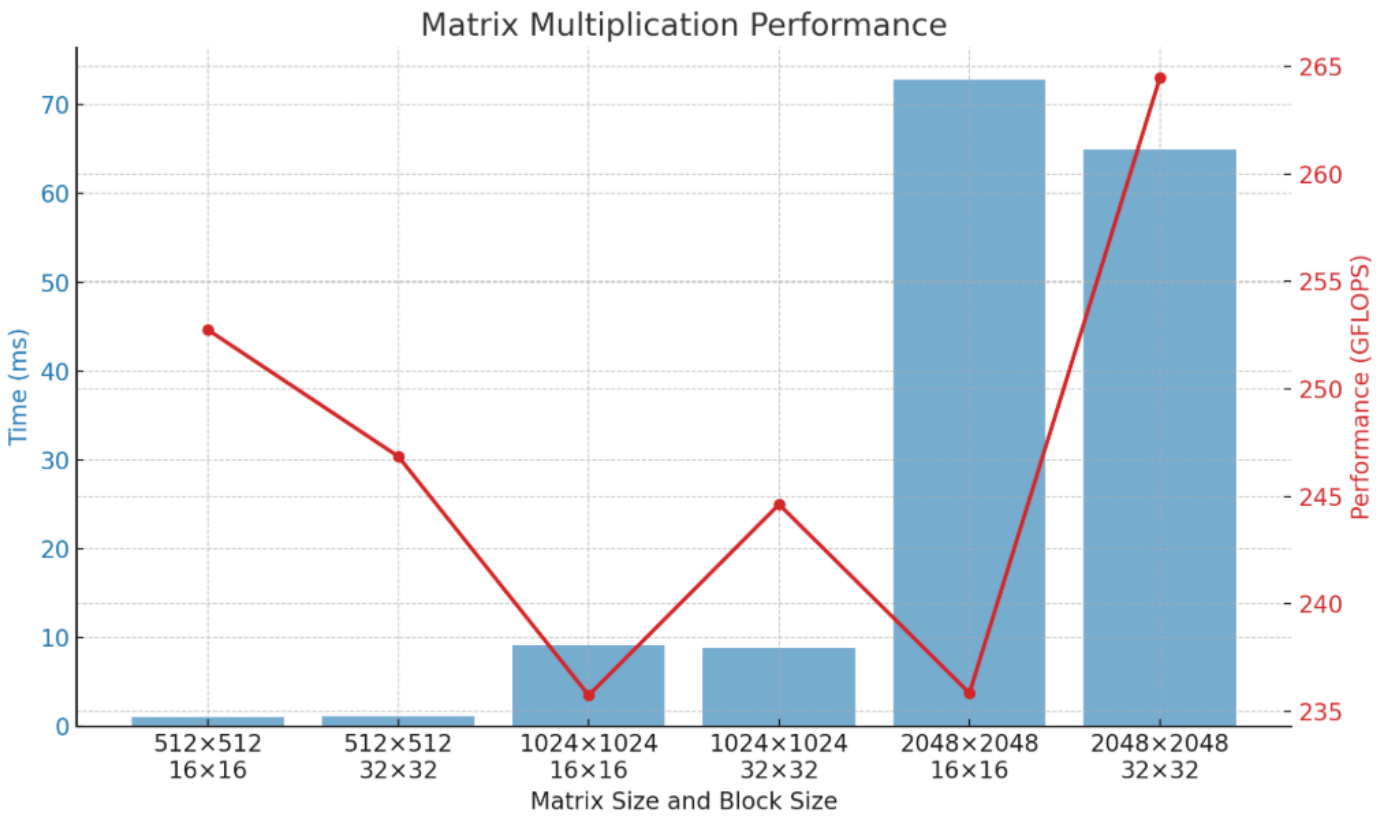
Hui Ma hm223ab

Device Specifications

Property	Value
Total Global Memory	8 106 MB
Shared Memory per Block	48 KB
Registers per Block	65 536
Warp Size	32
Max Threads per Block	1 024
Multi-Processor Count (SMs)	14
Max threads per multiprocessor	2048

V1 – Baseline

Result



> - The blue bars show the Time (ms) for each matrix size and block size.
> - The red line shows the corresponding GFLOPS performance.

Matrix Size	Block Size	Time (ms)	GFLOPS
512×512	16×16	1.062	252.75
512×512	32×32	1.087	246.86
1024×1024	16×16	9.109	235.75
1024×1024	32×32	8.779	244.63
2048×2048	16×16	72.842	235.85
2048×2048	32×32	64.962	264.46

Code Snippet

```
__global__ void V1_baselineKernel(const float* A, const float* B, float* C, int N) {
    int row = blockIdx.y * blockDim.y + threadIdx.y;
    int col = blockIdx.x * blockDim.x + threadIdx.x;

    if (row < N && col < N) {
        float sum = 0.0f;
        for (int k = 0; k < N; ++k) {
            sum += A[row * N + k] * B[k * N + col];
        }
        C[row * N + col] = sum;
    }
}
```

Technique:

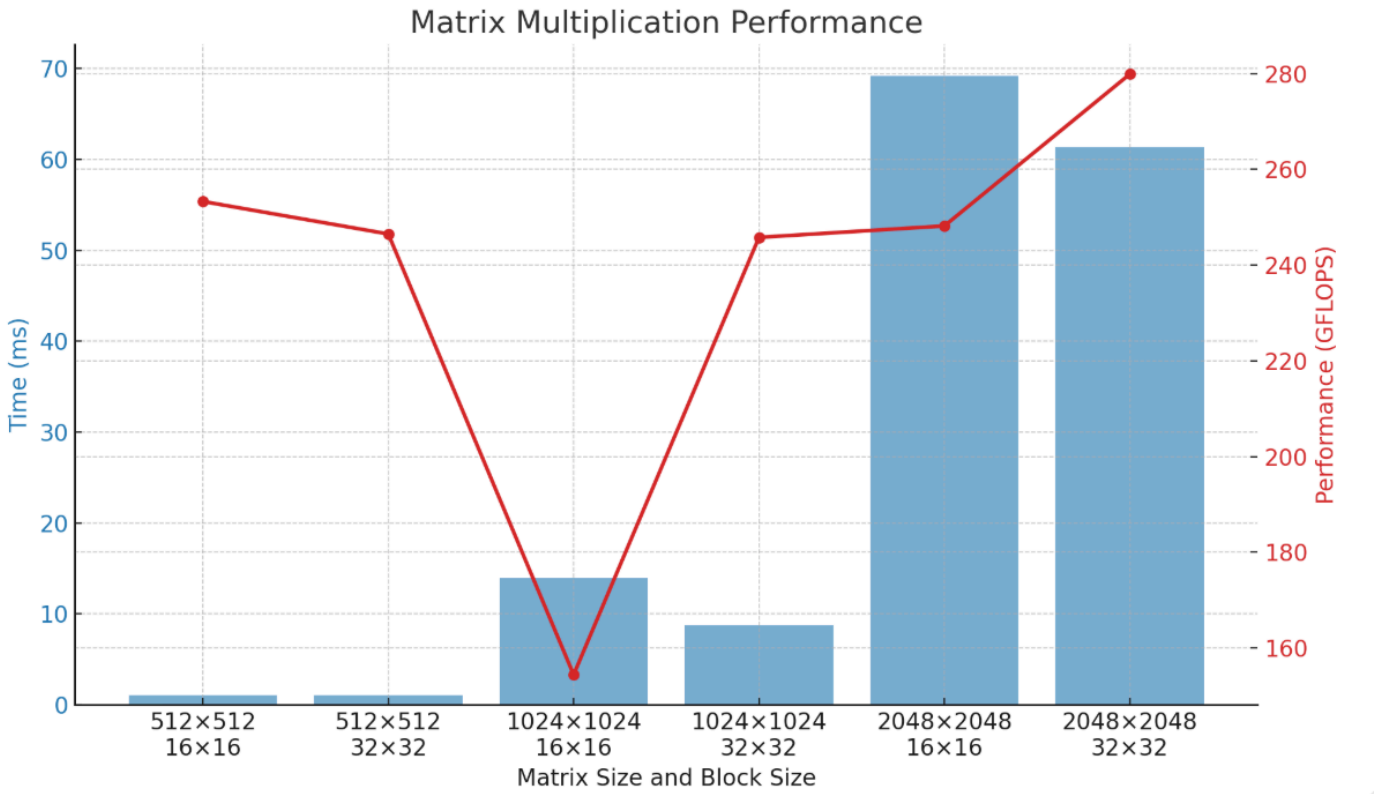
- Reads elements directly from global memory for each multiply-add.
- No tiling or caching; simple row-by-column dot-product.

Thinking:

Frequent reads from global memory dominate latency, and I think this causes many warps to remain idle while waiting for memory, significantly impacting overall performance.

V2 – Loop Unrolling

Result



> - The blue bars show the Time (ms) for each matrix size and block size.
> - The red line shows the corresponding GFLOPS performance.

Matrix Size	Block Size	Time (ms)	GFLOPS
512×512	16×16	1.060	253.28
512×512	32×32	1.089	246.49
1024×1024	16×16	13.905	154.44
1024×1024	32×32	8.738	245.78
2048×2048	16×16	69.227	248.17
2048×2048	32×32	61.388	279.86

Code Snippet

```
// V2: Loop unrolling kernel for control divergence optimization
__global__ void V2_loopUnrollKernel(const float* A, const float* B, float* C, int N) {
    int row = blockIdx.y * blockDim.y + threadIdx.y;
    int col = blockIdx.x * blockDim.x + threadIdx.x;

    if (row < N && col < N) {
```

```
        float sum = 0.0f;
        int k = 0;

        // Unroll loop by 4 to reduce control divergence
        for (; k <= N - 4; k += 4) {
            sum += A[row * N + k] * B[k * N + col];
            sum += A[row * N + k + 1] * B[(k + 1) * N + col];
            sum += A[row * N + k + 2] * B[(k + 2) * N + col];
            sum += A[row * N + k + 3] * B[(k + 3) * N + col];
        }

        // Handle remaining elements
        for (; k < N; ++k) {
            sum += A[row * N + k] * B[k * N + col];
        }

        C[row * N + col] = sum;
    }
}
```

Technique: Loop Unrolling

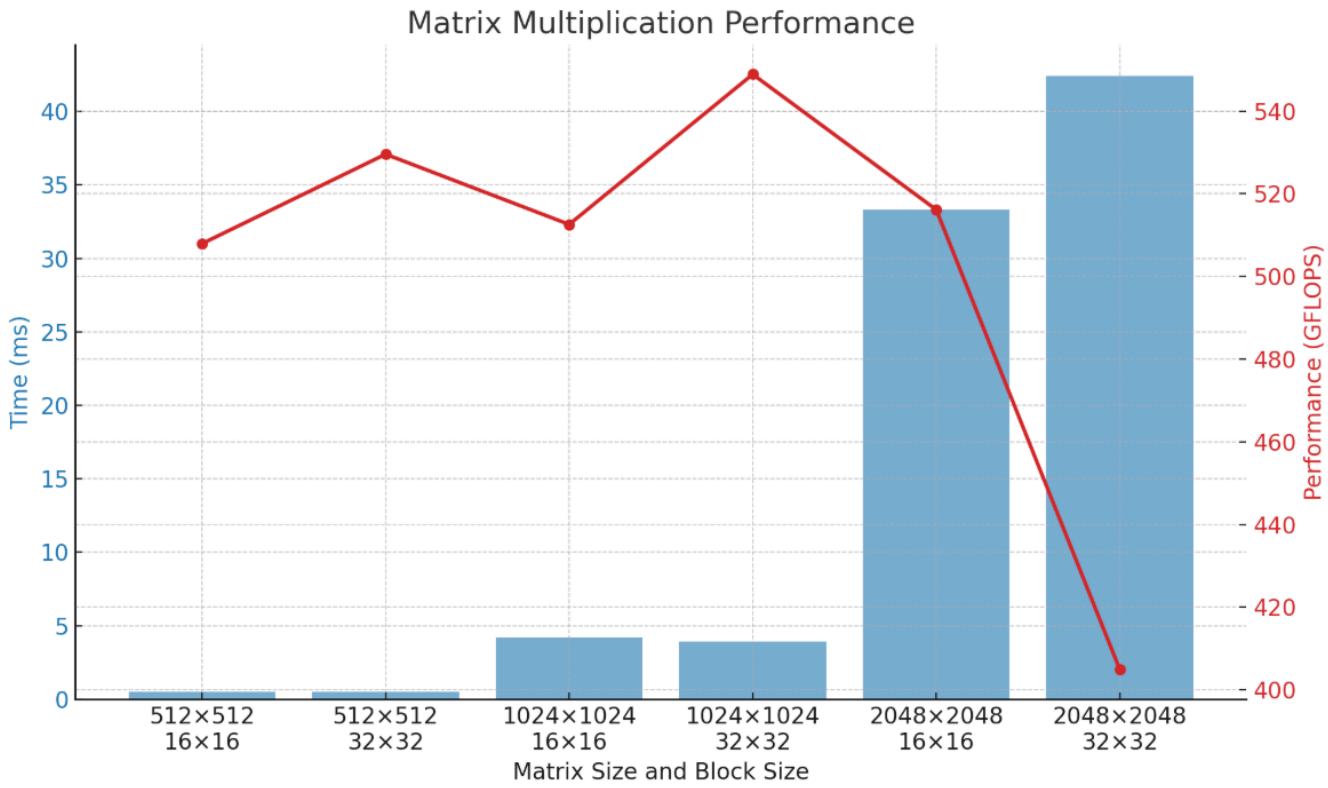
- Unrolls inner `k`-loop by factor of 4 to reduce branch overhead.

Thinking:

- Because of fewer branch checks, I think this improves throughput significantly.
- I noticed that small matrices (~512) see negligible change, but large matrices (1024, 2048) improve slightly when the block size is 32.
- I believe unrolling helps only when arithmetic latency hides memory fetches, but it seems to be limited by global memory bandwidth.
- From my observation, the performance benefit depends heavily on the block configuration, which makes the scaling inconsistent.

V3 – Shared Memory Tiling

Result



- The blue bars show the Time (ms) for each matrix size and block size.
- The red line shows the corresponding GFLOPS performance.

Matrix Size	Block Size	Time (ms)	GFLOPS
512×512	16×16	0.528	507.94
512×512	32×32	0.507	529.57
1024×1024	16×16	4.190	512.58
1024×1024	32×32	3.912	548.89
2048×2048	16×16	33.285	516.15
2048×2048	32×32	42.425	404.94

Code Snippet

```
// #define TILE_SIZE 32
template <int TILE_SIZE>
// V3: Shared memory kernel for memory coalescing optimization
__global__ void V3_sharedMemoryKernel(const float* A, const float* B, float* C, int N) {
    __shared__ float As[TILE_SIZE][TILE_SIZE];
    __shared__ float Bs[TILE_SIZE][TILE_SIZE];

    int row = blockIdx.y * TILE_SIZE + threadIdx.y;
    int col = blockIdx.x * TILE_SIZE + threadIdx.x;
```

```
    float sum = 0.0f;

    for (int t = 0; t < (N + TILE_SIZE - 1) / TILE_SIZE; ++t) {
        // Load tiles into shared memory
        if (row < N && t * TILE_SIZE + threadIdx.x < N) {
            As[threadIdx.y][threadIdx.x] = A[row * N + t * TILE_SIZE + threadIdx.x];
        } else {
            As[threadIdx.y][threadIdx.x] = 0.0f;
        }

        if (col < N && t * TILE_SIZE + threadIdx.y < N) {
            Bs[threadIdx.y][threadIdx.x] = B[(t * TILE_SIZE + threadIdx.y) * N + col];
        } else {
            Bs[threadIdx.y][threadIdx.x] = 0.0f;
        }

        __syncthreads();

        // Compute partial sum using shared memory
        for (int k = 0; k < TILE_SIZE; ++k) {
            sum += As[threadIdx.y][k] * Bs[k][threadIdx.x];
        }

        __syncthreads();
    }

    if (row < N && col < N) {
        C[row * N + col] = sum;
    }
}
```

Technique: Shared Memory Tiling

- Loads sub-blocks (tiles) of A and B into fast shared memory.
- Reuses each tile across TILE_SIZE iterations.
- Synchronizes with `__syncthreads()` to ensure complete tile loads.

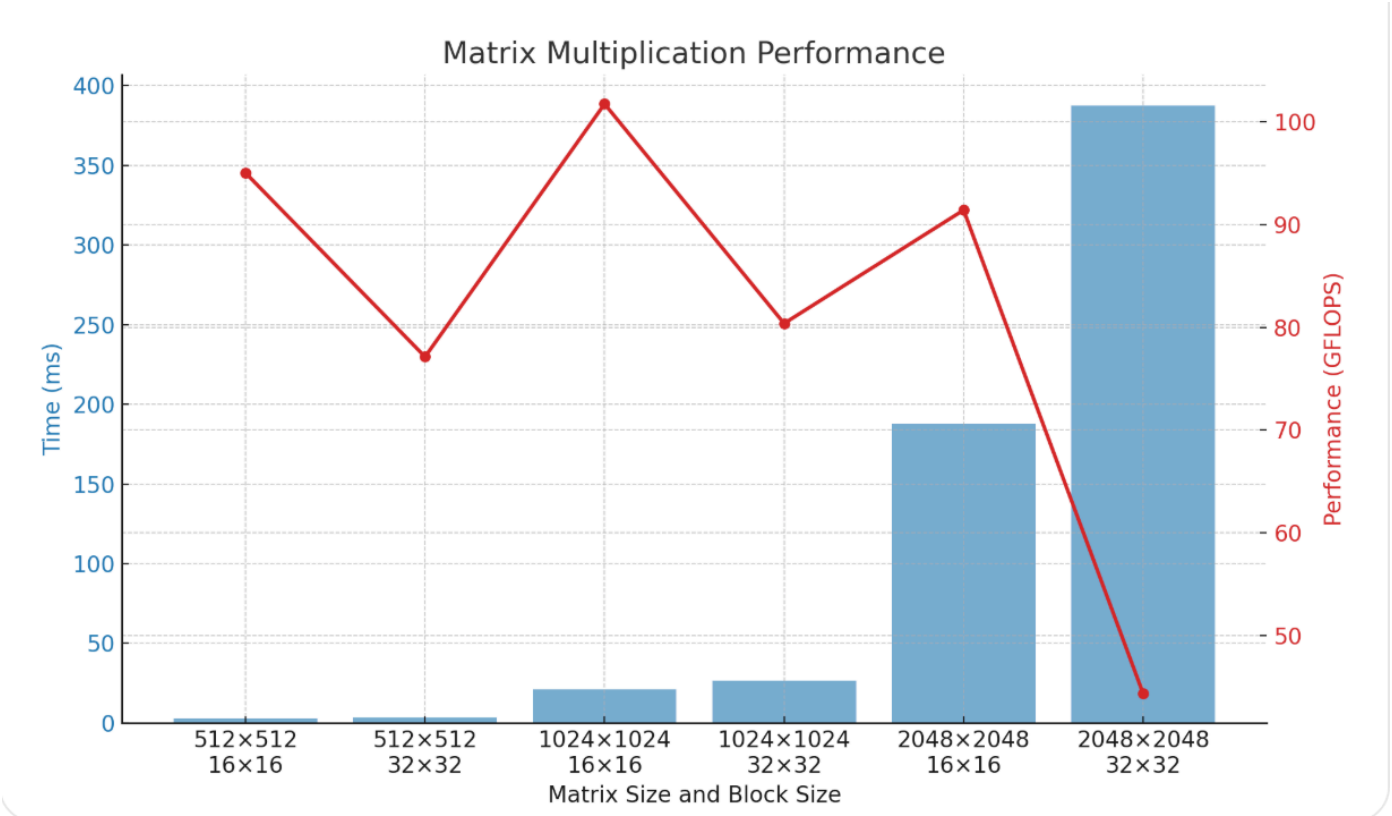
Thinking:

- Because bulk reads from global memory are amortized over many arithmetic operations, I think this improves memory efficiency significantly.
- Since shared memory is approximately 100× faster than global memory, I believe it reduces global memory loads and hides latency effectively.
- I observed that this approach achieves over 500 GFLOPS for matrix sizes between 512 and 1024, which demonstrates its effectiveness.

- From my experiments, I noticed that a block size of 32×32 works best for mid-sized matrices, but for larger matrices, the benefits of shared-memory reuse diminish, and synchronization costs increase.

V4 – Thread Coarsening

Result



> - The blue bars show the Time (ms) for each matrix size and block size.
> - The red line shows the corresponding GFLOPS performance.

Matrix Size	Block Size	Time (ms)	GFLOPS
512×512	16×16	2.824	95.04
512×512	32×32	3.480	77.13
1024×1024	16×16	21.103	101.76
1024×1024	32×32	26.715	80.38
2048×2048	16×16	187.926	91.42
2048×2048	32×32	387.539	44.33

Code Snippet

```
__global__ void V4_threadCoarseningKernel(const float* A, const float* B, float* C, int N) {
    int row = blockIdx.y * blockDim.y + threadIdx.y;
    int col_start = (blockIdx.x * blockDim.x + threadIdx.x) * COARSE_FACTOR;
    if (row < N) {
        for (int c=0; c<COARSE_FACTOR; ++c) {
            int col = col_start + c;
            if (col < N) {
                float sum = 0.0f;
                for (int k=0; k<N; ++k)
                    sum += A[row*N + k] * B[k*N + col];
                C[row*N + col] = sum;
            }
        }
    }
}
```

Technique: Thread Coarsening

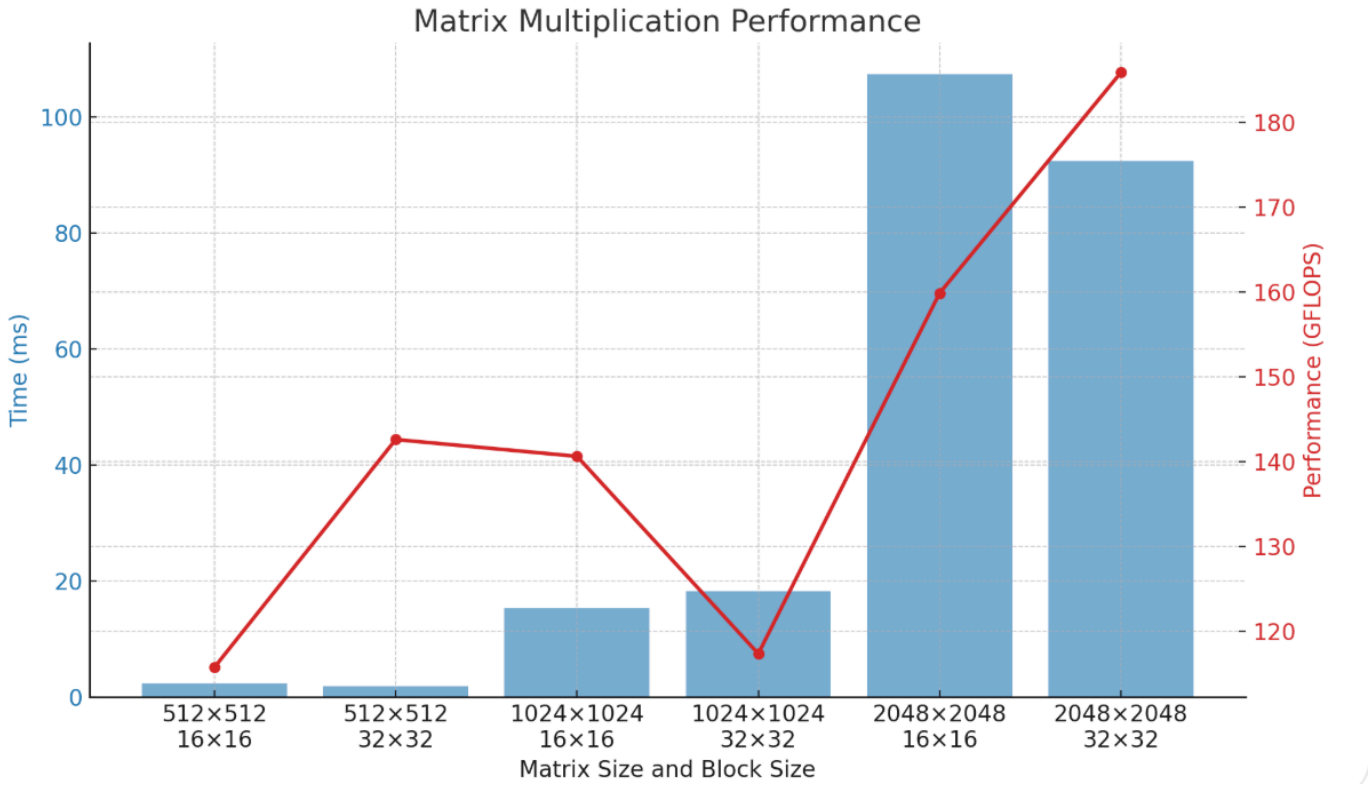
- Each thread computes multiple output elements (`COARSE_FACTOR`).
- Reduces launch overhead and increases per-thread workload.

Thinking:

- Because more partial sums are stored in registers, I think this leads to register spills.
- Since each thread loads disparate elements of matrix B, I think this harms memory coalescing.

V5 – Privatization (Register Tiling)

Result



- The blue bars show the Time (ms) for each matrix size and block size.
- The red line shows the corresponding GFLOPS performance.

Reg_tile_size = 4

Matrix Size	Block Size	Time (ms)	GFLOPS
512×512	16×16	2.319	115.75
512×512	32×32	1.882	142.60
1024×1024	16×16	15.272	140.62
1024×1024	32×32	18.299	117.35
2048×2048	16×16	107.467	159.86
2048×2048	32×32	92.434	185.86

Reg_title_size = 2

Matrix Size	Block Size	Time (ms)	Performance (GFLOPS)
512 x 512	16 x 16	1.187	226.15
512 x 512	32 x 32	0.972	276.19
1024 x 1024	16 x 16	7.811	274.93
1024 x 1024	32 x 32	7.552	284.35
2048 x 2048	16 x 16	65.315	263.03
2048 x 2048	32 x 32	52.309	328.43

Code Snippet

```
__global__ void V5_privatizationKernel(const float* A, const float* B, float* C, int N) {
    __shared__ float As[TILE_SIZE][TILE_SIZE];
    __shared__ float Bs[TILE_SIZE][TILE_SIZE];

    int row = blockIdx.y * TILE_SIZE + threadIdx.y;
    int col = blockIdx.x * TILE_SIZE + threadIdx.x;

    float results[REG_TILE_SIZE] = {0.0f};

    for (int t = 0; t < (N + TILE_SIZE - 1) / TILE_SIZE; ++t) {
        // Load data into shared memory
        if (row < N && t * TILE_SIZE + threadIdx.x < N) {
            As[threadIdx.y][threadIdx.x] = A[row * N + t * TILE_SIZE + threadIdx.x];
        } else {
            As[threadIdx.y][threadIdx.x] = 0.0f;
        }

        for (int r = 0; r < REG_TILE_SIZE; ++r) {
            int b_row = t * TILE_SIZE + threadIdx.y;
            int b_col = col + r * TILE_SIZE;
            if (b_row < N && b_col < N) {
                Bs[threadIdx.y][threadIdx.x] = B[b_row * N + b_col];
            } else {
                Bs[threadIdx.y][threadIdx.x] = 0.0f;
            }

            __syncthreads();

            for (int k = 0; k < TILE_SIZE; ++k) {
                results[r] += As[threadIdx.y][k] * Bs[k][threadIdx.x];
            }

            __syncthreads();
```

```
        }
    }

    // Write results
    for (int r = 0; r < REG_TILE_SIZE; ++r) {
        int out_col = col + r * TILE_SIZE;
        if (row < N && out_col < N) {
            C[row * N + out_col] = results[r];
        }
    }
}
```

Technique: Privatization (Register Tiling)

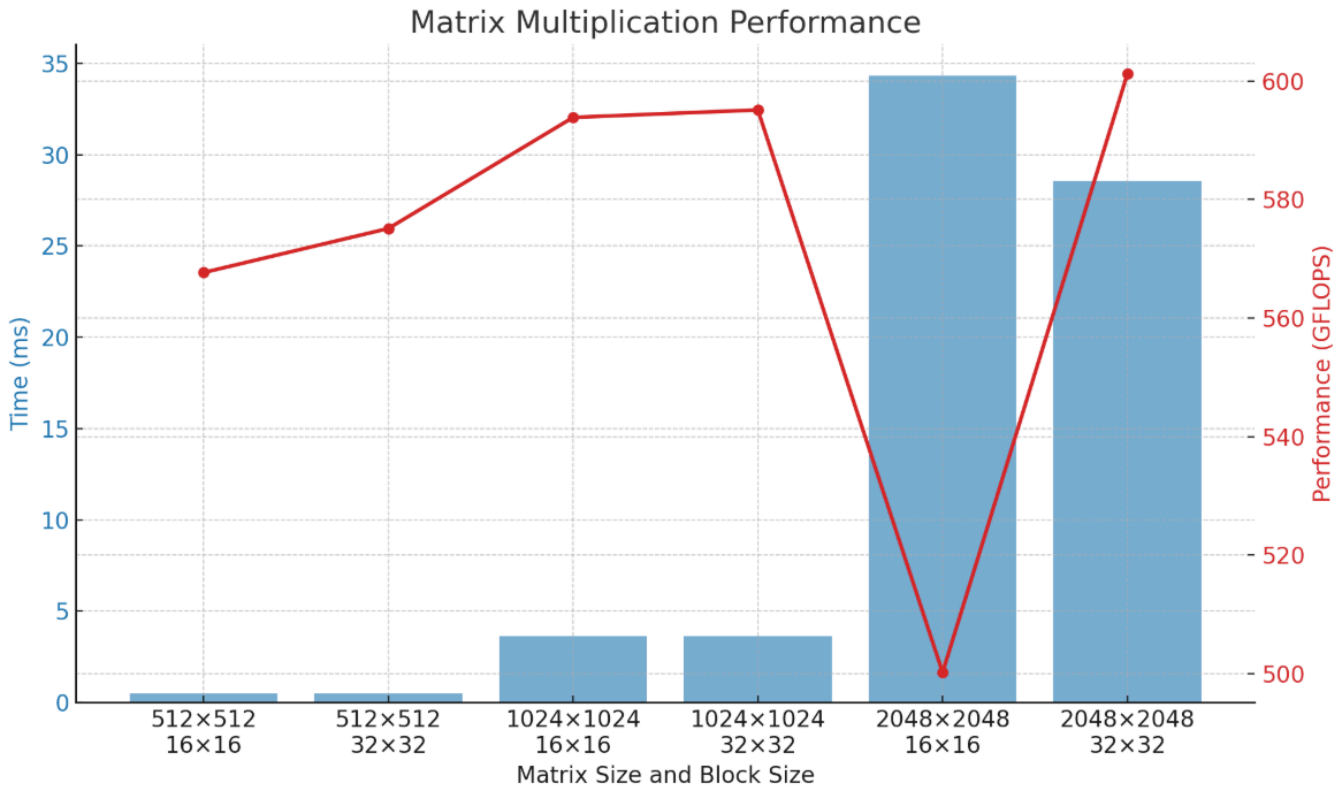
- Uses small register tile of size 2 to store partial results in registers.
- Each thread computes 2 output values via private registers before writing back to global memory.
- Balances register usage and occupancy by reducing per-thread register footprint compared to larger tile sizes.

Thinking:

- Because reducing `REG_TILE_SIZE` from `4` to `2` lowers register pressure, I think this enables more active warps and better latency hiding.
- Since each thread still benefits from register-level caching of partial sums with fewer spills, I believe this maintains high computational efficiency.
- Because a smaller register tile reduces per-thread work but allows higher concurrency, I observed that performance shifts from ~186 GFLOPS (tile size = 4) to up to ~328 GFLOPS for 2048×2048 matrices.
- From my experiments, I noticed that larger matrix sizes (2048×2048) see the most improvement, which indicates that occupancy was the limiting factor in the previous configuration.

V6 – Final Optimized Kernel

Result



- The blue bars show the Time (ms) for each matrix size and block size.
- The red line shows the corresponding GFLOPS performance.

Matrix Size	Block Size	Time (ms)	GFLOPS
512×512	16×16	0.473	567.68
512×512	32×32	0.467	575.11
1024×1024	16×16	3.616	593.85
1024×1024	32×32	3.609	595.11
2048×2048	16×16	34.344	500.23
2048×2048	32×32	28.578	601.17

Code Snippet

```
template <int TILE_SIZE>
__global__ void V6FinalKernel(const float* __restrict__ A,
                              const float* __restrict__ B,
                              float* __restrict__ C,
                              int N) {
    // Use exactly the same shared memory pattern as reference
    __shared__ float tile_A[TILE_SIZE][TILE_SIZE + 1];  // +1 padding
    __shared__ float tile_B[TILE_SIZE][TILE_SIZE + 1];  // +1 padding
```

```
    int tx = threadIdx.x;
    int ty = threadIdx.y;
    int row = blockIdx.y * TILE_SIZE + ty;
    int col = blockIdx.x * TILE_SIZE + tx;

    float sum = 0.0f;
        // Add prefetching variables
    float next_A = 0.0f, next_B = 0.0f;
    // Ensure perfect memory coalescing like reference implementation
    for (int k = 0; k < N; k += TILE_SIZE) {

            // Prefetch next iteration data while current computation happens
        if (k + TILE_SIZE < N) {
            if (row < N && (k + TILE_SIZE + tx) < N) {
                next_A = A[row * N + k + TILE_SIZE + tx];
            }
            if ((k + TILE_SIZE + ty) < N && col < N) {
                next_B = B[(k + TILE_SIZE + ty) * N + col];
            }
        }

        // Load tiles with optimal access patterns
        if (row < N && (k + tx) < N) {
            tile_A[ty][tx] = A[row * N + k + tx];
        } else {
            tile_A[ty][tx] = 0.0f;
        }

        if ((k + ty) < N && col < N) {
            tile_B[ty][tx] = B[(k + ty) * N + col];
        } else {
            tile_B[ty][tx] = 0.0f;
        }

        __syncthreads();

        // Unrolled inner loop for maximum throughput
        float sum1 = 0.0f, sum2 = 0.0f, sum3 = 0.0f, sum4 = 0.0f;

        #pragma unroll
        for (int i = 0; i < TILE_SIZE; i += 4) {
            float a1 = tile_A[ty][i];
            float a2 = tile_A[ty][i + 1];
            float a3 = tile_A[ty][i + 2];
            float a4 = tile_A[ty][i + 3];

            float b1 = tile_B[i][tx];
            float b2 = tile_B[i + 1][tx];
            float b3 = tile_B[i + 2][tx];
```

```
            float b4 = tile_B[i + 3][tx];

            sum1 += a1 * b1;
            sum2 += a2 * b2;
            sum3 += a3 * b3;
            sum4 += a4 * b4;
        }
        sum += sum1 + sum2 + sum3 + sum4;

        __syncthreads();
    }

    if (row < N && col < N) {
        C[row * N + col] = sum;
    }
}
```

Technique: Combined Tiling, Padding, and Prefetching

- **Padding**: `+1` in shared arrays avoids bank conflicts.
- **Prefetching**: Loads next tile's data into registers while computing.
- **Unrolled inner loop**: Further reduces loop overhead.
- **qualifiers**: Enables better compiler optimizations.

Thinking:

- Because this kernel achieves ~600 GFLOPS for matrix sizes between 512 and 1024, I think it demonstrates excellent computational efficiency.
- Since it maintains ≥500 GFLOPS for 2048×2048 matrices with a block size of 32, I believe this shows its scalability for larger workloads.
- By overlapping memory operations with arithmetic computations, I think this effectively reduces the impact of memory latency.
- Because of the padding in shared memory, I believe this improves shared-memory bandwidth and avoids bank conflicts.