ECE 1782

CUDA Programming Assignment

September 2020

1. Program Specification

Write a CUDA program that adds two matrices of the same size containing elements of the type `float`. Your CUDA program should accept two arguments:

1. An integer that specifies the number of columns in the matrices.
2. An integer that specifies the number of rows in the matrices.

Each time the CUDA program is invoked, it shall invoke a GPU kernel called `f_matadd`(), exactly once.

Your CUDA program should output 4 numbers **on one line** terminated with the newline character and then exit:

```
<total_time> <CPU_GPU_transfer_time> <kernel_time> <GPU_CPU_transfer_time><nl>
```

Each of these numbers are defined as follows (and the code provided below shows how to obtain these numbers):

- `<total_time>`: the time in seconds from the point just before data is transferred to the GPU to just after the result data is transferred back from the GPU (`timeStampD-timeStampA` in the code below).

- `<CPU_GPU_transfer_time>`: the time in seconds it takes to transfer the two matrices to the GPU (`timeStampB`–timeStampA in the code below).

- `<kernel_time>`: the time in seconds it takes the GPU to execute the kernel (`timeStampC-timeStampB` in the code below).

- `<GPU_CPU_transfer_time>`: the time in seconds it takes the GPU to transfer the result matrix back to the CPU (`timeStampD-timeStampC` in the code below).

The numbers should be output with 6 digits of precision (`"%.6f"`). All four numbers must be output only one line and be separated by white space.

If an error occurs, then your program should output **one line** starting with "`Error: `" followed by a description of the error before exiting.

The two matrices that will be added, here called `h_A` and `h_B`, should be initialized host-side (before copying them to GPU global memory) as follows:

- `h_A[i,j] = (float)  (i+j)/3.0 ;`
- `h_B[i,j] = (float) 3.14*(i+j) ;`

The result matrix produced on the GPU, here called d_C, should be copied back to CPU memory and (to check correctness) should be compared against the result of a corresponding matrix addition, h_hC, computed CPU side. Your code must do this comparison, and if there is a discrepancy, an error message must be output (instead of the numbers).

Please ensure that your program can correctly handle matrices of the following sizes: 16,384 x 16,384 (a nice square matrix), 32,768 x 8,192, and 30 x 8,947,850 . Your program should take significantly less than 60 seconds to execute --- when we test your code, we will cut off your execution after 60 seconds

When running your experiments, you will notice that transferring the matrices over the PCIe bus dominates the total time. This is because matrix addition is not very compute intensive. That is OK, as this is just an exercise for you to familiarize yourself with CUDA programming and for getting a feel for some of the things that affect performance. Optionally, you may want to also run versions of the above code (but not submit) where matrices are of type `int` and `double` (instead of type `float`) to see what difference it makes to your measured execution times.

2. Different variants of your program.

You should implement four variants of your program and collect timing information for each:

1. **Variant 1**: have each thread compute one element of the output matrix, C, and have two adjacent threads, t_i and t_{i+1} compute row-wise adjacent elements of C.

2. **Variant 2**: same as Variant 1, except have two adjacent threads compute column-wise adjacent elements of C.

3. **Variant 3**: same as Variant 1 but in this case allocate space for the three matrices in global memory to each be one element larger than what is needed for the matrix itself, and then have the matrices each be located GPU-side starting at word 1 of the allocated space (instead of word 0).

4. **Variant 4**: same as Variant 1, but have each thread compute four elements of the output matrix C.

3. Deliverables

You will have to submit (i) the timings you were able to obtain for each of the four variants of your CUDA program, and (ii) the code of Variant 4, all contained in one file, and implemented in C. Your submitted CDA file should be named `<your_student_no>.cu`.

The details on how to submit your submission will be provided at a later time.

4. Possible rough program sketch

Here is one possible program structure:

```
// time stamp function in seconds
double getTimeStamp() {
    struct timeval  tv ;
    gettimeofday( &tv, NULL ) ;
    return (double) tv.tv_usec/1000000 + tv.tv_sec ;
}

// host side matrix addition
h_addmat(float *A, float *B, float *C, int nx, int ny){ … }

// device-side matrix addition
__global__ f_addmat( float *A, float *B, float *C, int nx, int ny ){
    // kernel code might look something like this
    // but you may want to pad the matrices and index into them accordingly
    int ix = threadId.x + bloackId.x*blockDim.x ;
    int iy = threadId.y + bloackId.y*blockDim.y ;
    int idx = iy*ny + ix ;
    if( (ix<nx) && (iy<ny) )
        C[idx] = A[idx] + B[idx] ;
}

int main( int argc, char *argv[] ) {
    // get program arguments
    if( argc != 3) {
        printf( "Error: wrong number of args\n" ) ;
        exit() ;
    }
    int nx = atoi( argv[2] ) ; // should check validity
    int ny = atoi( argv[3] ) ; // should check validity
    int noElems = nx*ny ;
    int bytes = noElems * sizeof(float) ;
    // but you may want to pad the matrices…



    // alloc memory host-side
    float *h_A = (float *) malloc( bytes ) ;
    float *h_B = (float *) malloc( bytes ) ;
    float *h_hC = (float *) malloc( bytes ) ; // host result
    float *h_dC = (float *) malloc( bytes ) ; // gpu result

    // init matrices with random data
    initData( h_A, noElems ) ; initData( h_B, noElems ) ;

    // alloc memory dev-side
    float *d_A, *d_B, *d_C ;
    cudaMalloc( (void **) &d_A, bytes ) ;
    cudaMalloc( (void **) &d_B, bytes ) ;
    cudaMalloc( (void **) &d_C, bytes ) ;

    double timeStampA = getTimeStamp() ;
```

```
    //transfer data to dev
    cudaMemcpy( d_A, h_A, bytes, cudaMemCpyHostToDevice ) ;
    cudaMemcpy( d_B, h_B, bytes, cudaMemCpyHostToDevice ) ;
    // note that the transfers would be twice as fast if h_A and h_B
    // matrices are pinned

    double timeStampB = getTimeStamp() ;

    // invoke Kernel
    dim3 block( 32, 32 ) ;   // you will want to configure this
    dim3 grid( (nx + block.x-1)/block.x, (ny + block.y-1)/block.y ) ;

    f_addmat<<<grid, block>>>( d_A, d_B, d_C, nx, ny ) ;

    cudaDeviceSynchronize() ;

    double timeStampC = getTimeStamp() ;

    //copy data back
    cudaMemCpy( h_dC, d_C, bytes, cudaMemCpyDeviceToHost ) ;

    double timeStampD = getTimeStamp() ;

    // free GPU resources
    cudaFree( d_A ) ; cudaFree( d_B ) ; cudaFree( d_C ) ;
    cudaDeviceReset() ;

    // check result
    h_addmat( h_A, h_B, h_hC, nx, ny ) ;
    // h_dC == h+hC???

   // print out results
}
```