# Parallel Processing

## Assignment 10

**This assignment is individual assignment, every student should submit by himself.**

**Due: Next Section**

1. **To optimize the matrix transpose parallel algorithm we divide the matrix into a set of tiles and transpose each tile separately. This can be done with two methods**
   a. **At the global memory**
   b. **At a shared memory**

   **Which method do you think would produce better performance? Explain.**

2. **What will be the slowdown for each of the following expressions in switch statement?**

```
a. __global__ void foo(){  switch (threadIdx.x%32) case(0…31)}
      kernel <<1024, 1>> ();
b. __global__ void foo(){  switch (threadIdx.x%2)  case(0,1)}
     kernel <<1024, 1>> ();
c. __global__ void foo(){  switch (threadIdx.y%2)  case(0,1)}
     kernel <<16*16, 1>> ();
d. __global__ void foo(){
          for(int i=0; i < threadIdx.x%2; i++)
                bar();
   }kernel <<1024, 1>> ();
```

3. **How to optimize a branch divergence problem?**

4. **How to optimize the mathematical operations in your parallel program?**

5. **If all operations take exactly 1 second, how long to complete the following:**

```
cudaStream_t s1, s2;

a. cudaMemcpyAsync(&d_arr1, &h_arr1, numbytes, cudaH2D, s1)
   A<<<1, 128, s1 >>>(d_arr2);
b. cudaMemcpyAsync(&d_arr1, &h_arr1, numbytes, cudaH2D, s1)
   A<<<1, 128, s2 >>>(d_arr1);
```