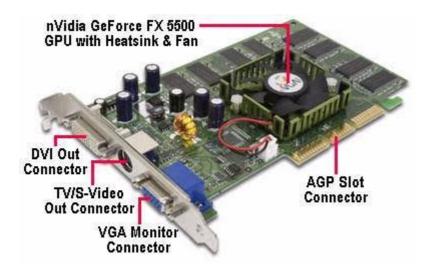
Excercise 5 Introduction to CUDA

สิ่งที่ต้องใช้ 1. WinSCP 2. Terminal 3. Code Editor 4. CUDA (nvcc or Nvidia compiler) 5. สไลด์อาจารย์ด้วยก็ดีนะ

GPU คืออะไร?

GPU หรือ Graphic Processing Unit มันเกิดมาเพื่อประมวลผลกราฟิก 3 มิติโดยเฉพาะมีสองบริษัทตอนนี้คือ Nvidia และ AMD/ATI ในสมัยก่อน CPU (Central Processing Unit) นั้นจะทำการประมวลผลกราฟิกเป็นหลักทำให้ CPU ทำงาน ไม่พอต่องานด้านอื่นๆ จึงเกิด GPU ขึ้นมาเพื่อลดภาระของ CPU ได้มากขึ้นและทำงานได้ดียิ่งขึ้นกว่าเดิม

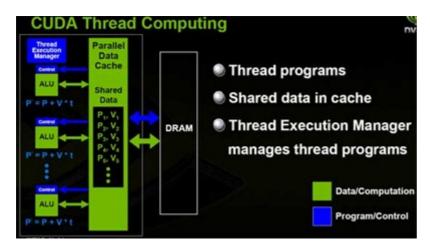


สมัยก่อนจะมีพอร์ตเฉพาะสำหรับกราฟิก ส่วนพอร์ตเชื่อมต่อนั้นจะเรียกว่า AGP (Accelerated Graphics Port) ก่อนจะ กลายมาเป็น PCIe (Peripheral Component Interconnection Express) ซึ่งจะมีตั้งแต่ x16 x8 x4 x1 ช่องเชื่อมต่อ สมัยก่อนเรียกว่า VGA (Video Graphic Array) ซึ่งช่องต่ออันนี้จะแสดงภาพออกมาในรูปแบบ Matrix Array ซึ่งช่อง แต่ละช่องจะเรียกว่าพิกเชลซึ่งบรรจุค่าในอาเรย์เป็นสีทั้งหมด 3 สี แดง น้ำเงิน เหลือง ค่าต่างกันรวมกันจึงเกิดเป็นภาพขึ้น มา

แล้วอะไรคือ CUDA กันนะ?



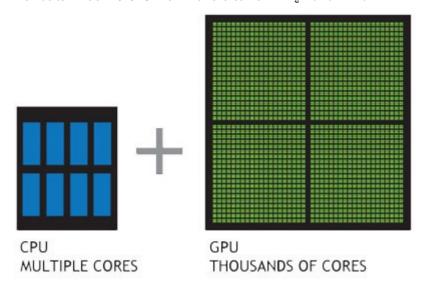
CUDA (Compute Unified Device Architecture) คือ การประมวลผลแบบคู่ขนานและ Application Programing Interface (API) นี้เกิดจากบริษัท Nvidia เพื่อให้โปรแกรมเมอร์สามารถดึงประสิทธิภาพของการ์ดจอออกมาได้เต็มที่จาก GPU (Graphic Processing Unit) หรือ GPGPU (General-Purpose computing on Graphics Processing Units)



แล้วทำไมต้องใช้ GPU แทน CPU กันล่ะ

CPU นั้นจะประกอบไปด้วย core เพียงไม่กี่ core จึงเหมาะกับงานในลักษณะที่เรียกว่า Sequential serial processing หรือ การประมวลผลแบบลำดับ ในขณะที่ GPU จะประกอบด้วย core ขนาดเล็กจำนวนมากและถูกออกแบบให้มีการกระจายการ

ทำงานในลักษณะ Parallel หรือการประมวลผลแบบคู่ขนานแบบนี้



ซึ่ง GPU เหมาะกับงานประเภท Single Instruction Multiple Data stream (SIMD)

วิธีการดูเมทริกซ์ใน CUDA

เอาชื่อตัวแปรก่อนเลย

- 1. threadIdx.x // threadID
- 2. blockDim.x // block Dimension
- 3. blockId.x // blockID
- 4. gridDim.x // grid Dimension

threadIdx.x คือ ตัว Pointer ที่จะชี้/ระบุตำแหน่งภายใน block แกน X ซึ่ง ตรง .x สามามารถเปลี่ยนเป็น Y หรือ Z ได้

blockDim.x จำนวนเธรด (Thread) ในบล็อกแกน X ซึ่ง ตรง .x สามามารถเปลี่ยนเป็น Y หรือ Z ได้

blockld.x คือ Pointer หรือตัวชี้บล็อก (block) ที่ข้างในนั้นมีเธรดบรรจุอยู่ ในแกน X ซึ่ง ตรง .x สามามารถเปลี่ยนเป็น Y หรือ Z ได้

gridDim.x คือ จำนวน threadblocks ในแกน X ซึ่ง ตรง .x สามามารถเปลี่ยนเป็น Y หรือ Z ได้

มาดูตัวอย่าง **อาเรย์ 1** มิติ ใน CUDA กัน

```
threadIdx.Z มีค่า 0
_blockDim.x = 4, _blockDim.y = 1, blockDim.z = 0 จะมี thread ทั้งหมด 4*1 = 4 เธรด
```

มาดูตัวอย่าง **อาเรย์ 2 มิติ** ใน CUDA

```
|thread 1| |thread 2| |thread 3|
    |x=0, y=0| |x=1, y=0| |x=2, y=0|
    -----
    |thread 4| |thread 5| |thread 6|
    |x=0, y=1| |x=1, y=1| |x=2, y=1|
    _____
     -----
|u | |thread 7| |thread 8| |thread 9|
|n| |x=0, y=2| |x=1, y=2| |x=2, y=2|
                          ----> แกน X |
threadIdx.x ตัวที่ 1 หรือ thread 1 (threadIdx.x = 1), threadIdx.x = 0 , threadIdx.y =
threadIdx.Z มีค่า 0
threadIdx.x ตัวที่ 6 หรือ thread 6 (threadIdx.x = 2), threadIdx.y = 1 threadIdx.Z มีค่า
threadIdx.x ตัวที่ 4 หรือ thread 4 (threadIdx.x = 0), threadIdx.y = 1 threadIdx.Z มีค่า
threadIdx.x ตัวที่ 9 หรือ thread 9 (threadIdx.x = 2), threadIdx.y = 2 threadIdx.Z มีค่า
_blockDim.x = 3, _blockDim.y = 3, blockDim.z = 0 ก็คือมี Thread ทั้งหมด 3*3 = 9 thread
```

มาเริ่มที่ข้อ 1. กันเถอะ

1. Use CUDA to assign a value to each element of the array of integers A[256] using 256 threads. Each A[i] is assigned with the value of 2*i, for i = 0 to 255.

โจทย์ข้อนี้ให้เราใช้ CUDA เพื่อกำหนดค่าในอาเรย์ตำแหน่งโดยใช้ อาเรย์ขนาด 256 และใช้เธรดขนาด 256 เธรด โด ยอาเรย์ A ที่มีขนาด 256 นั้นในแต่ละตำแหน่งจะถูกกำหนดค่าโดย i * 2 โดยที่ i เริ่มตั้งแต่ 0 จนถึง 255

```
#include <stdio.h>
#define T 256 // As Threads

__global___ void vecMultiply(int *A) {
    int i = threadIdx.x;
    A[i] = A[i] * 2; // ตำแหน่งจะถูกกำหนดค่าโดย i * 2 โดยที่ i เริ่มตั้งแต่ 0 จนถึง 255
}

int main (int argc, char *argv[])
```

```
int i;
        int size = T*sizeof(int);
        int a[T], *devA;
        for (i=0; i < T; i++)
                a[i] = i + 1;
        cudaMalloc( (void**)&devA,size);
        cudaMemcpy( devA, a, size, cudaMemcpyHostToDevice);
        vecMultiply<<<1, T>>>(devA); // ตรงนี้เรียกว่า Kernel launch
        printf("Before\n");
        for (i=0; i < T; i++)
                printf("%d ", a[i]);
        printf("\n");
        cudaMemcpy(a, devA, size, cudaMemcpyDeviceToHost);
        cudaFree(devA);
        printf("After\n");
        for (i=0; i < T; i++)
    {
                printf("%d ",a[i]);
        printf("\n");
}
```

ผลลัพธ์ที่ออกมาจะประมาณนี้

vecMultiply<<<1, T>>>(devA); ตรงนี้ใน **Nvidia CUDA** จะเรียกว่า **kernel Launch** ซึ่งค่า **1** ตรงนี้หมายถึง จำนวนกริด ของเธรดบล็อก (Grid number of Thread Blocks) และ ค่า **T** หมายถึง เธรดบล็อกที่มีขนาดเป็น T ตัวในการทำงานแบบคู่ ขนาน (thread block has T parallel threads)ถ้าเราจะมองให้เห็นภาพกว่านี้อีกนิ๊ดด

```
mykernel<<<blooks, threads, shared_mem, stream>>>(args);
vecMultiply<<<1, T>>>(devA);
```

คือตรง mykernel อ่ะมันเป็นคำสั่งหรือ **kernel launch** เพื่อส่งข้อมูลไปหา Device (CUDA) ประมวลผลลัพธ์มาให้โดย เทียบกับ **vecMultiply** มันก็คือ Kernel launch เช่นกันโดยค่า 1 คือ **blocks** และ T คือ **Threads Blocks** นั่นเอง

ตัวอย่าง

Function ที่ CUDA สามารถติดต่อกับ Devices ได้

```
__host__ // เรียกใช้ได้เฉพาะบน Host และรันได้แค่ Host เท่านั้น
__global__ // เรียกใช้จาก Host ไปรันบน Device (CUDA)
__device__ // เรียกใช้จาก Device ไปรันบน Device เท่านั้น (CUDA)
```

วิธีส่งโค้ดไปคำนวณบน CUDA Device

วิธีมันจะพิสดารกว่าปกติหน่อยเพราะมี **Memory** ที่แยกกันออกมาเลยต้องกำหนดขนาด Memory แล้วค่อยโยนโค้ดไปคิด แล้วโยนผลลัพธ์กลับมาหาเครื่อง วิธีก็ตามนี้

```
    Allocate Memory บน Device # cudaMalloc( (void**)&devA,size);
    Transfer ข้อมูลจาก Host ไปยัง Device # cudaMemcpy( devA, a, size, cudaMemcpyHostToDevice);
    Kernel Launch # vecMultiply<<<1, T>>>(devA);
    Transfer ข้อมูลกลับจาก Device ไปยัง Host # cudaMemcpy(a, devA, size, cudaMemcpyDeviceToHost);
    Free Memory บน Device # cudaFree(devA);
```

วิธีคอมไพล์ไฟล์ CUDA (.cu)

ให้คำสั่งตามนี้

```
nvcc hello.cu -o hello # เพื่อคอมไพล์
nvcc -o hello hello.cu # หรือแบบนี้ก็ย่อมได้
```

วิธีรัน CUDA

วิธีรันคล้ายๆกับการรันภาษาซีบน Unix เลยแค่

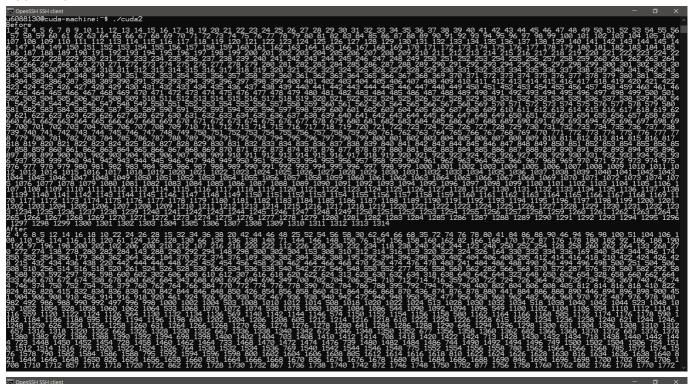
```
. / ชื่อไฟล์ที่จะรัน
```

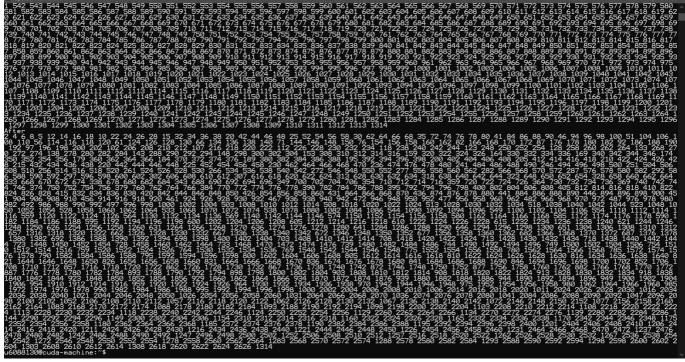
ข้อ 2.

Repeat Question 1 with the array A[1314], using only 256 threads. ก็คือเอาโค้ดจากข้อ 1 มาแก้เพิ่มโดยกำหนดขนาดอาเรย์ A ให้มีขนาดเป็น 1314 และใช้เธรดแค่ 256 ตัวเท่านั้น

```
#include <stdio.h>
#define T 256 // As Threads
#define ArraySize 1314
__global__ void vecMultiply(int *A)
        int i;
        int threadID = threadIdx.x;
        int start = (threadID * ArraySize) / 256;
        int end = ((threadID + 1) * ArraySize) / 256) - 1;
        for(i = start ; i < end ; i++)</pre>
                A[i] = A[i] * 2;
        }
int main (int argc, char *argv[])
{
        int i;
        int size = ArraySize*sizeof(int);
        int a[size], *devA; // ตรงนี้ a[size] จะกำหนดขนาดด้วย 1314 เรียบร้อยแล้ว
        for (i=0; i< ArraySize; i++)
                a[i] = i + 1;
        cudaMalloc( (void**)&devA,size);
        cudaMemcpy( devA, a, size, cudaMemcpyHostToDevice);
        vecMultiply<<<1, 256>>>(devA); // 1 , 256 mean send each data with total
256 thread blocks
        printf("Before\n");
        for (i=0; i< ArraySize; i++)
                printf("%d ", a[i]);
        printf("\n");
        cudaMemcpy(a, devA, size, cudaMemcpyDeviceToHost);
        cudaFree(devA);
        printf("After\n");
        for (i=0; i < ArraySize; i++)
    {
                printf("%d ",a[i]);
        printf("\n");
```

ผลลัพธ์ที่ได้





ข้อ 3.

Given two array of integers A and B, each having size 256 elements, write a CUDA program with the following kernel to copy the elements of array A to array B in reverse order using 256 threads. For example, if input array $A = \{1, 2, 3, ..., 256\}$, the output array $B = \{256, ..., 3, 2, 1\}$.

```
__global__ void reverseArray (int *A, int *B)
{
```

```
/* Code to reverse array is here */
}
```

โจทย์ข้อนี้เราจะต้องใช้อาเรย์สองตัวแต่ละตัวมีขนาดเก็บข้อมูลได้ 256 ช่อง โดยให้ใช้การเขียนโดย CUDA เพื่อทำการ Copy ค่าจากอาเรย์ A ไปอาเรย์ B โดยค่าที่ Copy นั้นจะต้องเป็นค่าที่กลับกัน จากท้ายสุดมาหน้าสุด (Reversed)

```
#include <stdio.h>
#define T 256 // As Threads
__global__ void reverseArray(int *A, int *B)
        int threadID = threadIdx.x;
        int Reverse = (T - 1) - threadID; // ตรงนี้เอาไว้ทำ Reverse
        B[Reverse] = A[threadID]; // ตรงนี้ก็เช่นกัน
}
int main (int argc, char *argv[])
        int i;
        int size = T*sizeof(int);
        int a[T],b[T], *devA,*devB;
        for (i=0; i < T; i++)
                a[i] = i + 1;
        cudaMalloc( (void**)&devA,size);
        cudaMalloc( (void**)&devB,size);
        cudaMemcpy( devA, a, size, cudaMemcpyHostToDevice);
        cudaMemcpy( devB, b, size, cudaMemcpyHostToDevice);
        reverseArray<<<1, T>>>(devA,devB); // 1 , T mean send 1 until total 256
thread blocks
        printf("Before\n");
        for (i=0; i < T; i++)
        {
                printf("%d ", a[i]);
        printf("\n");
        cudaMemcpy(a, devA, size, cudaMemcpyDeviceToHost);
        cudaMemcpy(b, devB, size, cudaMemcpyDeviceToHost);
        cudaFree(devA);
        cudaFree(devB);
        printf("After\n");
        for (i=0; i < T; i++)
    {
                printf("%d ",b[i]);
        printf("\n");
```

```
}
```

ผลลัพธ์ที่ได้



สวยงาม

ข้อ 4.

Repeat Question 3 with the array A and B, each having 1314 elements, using only 256 threads. ก็เหมือนเดิมเพิ่ม เดิมคือกำหนดขนาดไว้เป็น 1314 และใช้เธรดแค่ 256 ตัวนะจ๊ะ

```
#include <stdio.h>
#define T 256 // As Threads
#define ArraySize 1314
__global__ void reverseArray(int *A, int *B)
        int threadID = threadIdx.x;
        int start = (threadID * ArraySize) / 256;
        int end = ((threadID + 1) * ArraySize) / 256) - 1;
        while(end > ∅)
                B[end] = A[start];
                end--;
                start++;
        }
int main (int argc, char *argv[])
{
        int i;
        int size = ArraySize*sizeof(int);
        int a[ArraySize],b[ArraySize], *devA,*devB;
        for (i=0; i< ArraySize; i++)
                a[i] = i + 1;
```

```
cudaMalloc( (void**)&devA, size);
        cudaMalloc( (void**)&devB,size);
        cudaMemcpy( devA, a, size, cudaMemcpyHostToDevice);
        cudaMemcpy( devB, b, size, cudaMemcpyHostToDevice);
        reverseArray<<<1, 256>>>(devA,devB); // 1 , 256 mean send each data with
total 256 thread blocks
        printf("Before\n");
        for (i=0; i< ArraySize; i++)
                printf("%d ", a[i]);
        printf("\n");
        cudaMemcpy(a, devA, size, cudaMemcpyDeviceToHost);
        cudaMemcpy(b, devB, size, cudaMemcpyDeviceToHost);
        cudaFree(devA);
        cudaFree(devB);
        printf("After\n");
        for (i=0; i < ArraySize; i++)
    {
                printf("%d ",b[i]);
        printf("\n");
}
```

ผลลัพธ์ที่ได้

