Sesión 1 – Primeros pasos con CUDA

Antes de empezar

Todo lo necesario para la sesión 01 lo encontraréis en el fichero `Sesion01.tar` que hay que desempaquetar con el siguiente comando:

```
tar xvf Sesion01.tar
```

Recordad que previamente hay que hacer un `scp` de `Sesion01.tar` en boada.

ScanDevices

Este test sólo se ha de compilar y ejecutar. Para ejecutar un programa en CUDA en boada hay que utilizar el sistema de colas. En nuestro caso para compilar y ejecutar:

```
make
sbatch job.sh
```

En todas las sesiones incluiremos el `Makefile` correspondiente, así como el script para lanzar la ejecución a la cola `cuda` (esta cola está especificada en el fichero `job.sh`).

Al ejecutar este test sabremos qué GPUs tenemos instalada en el servidor y para cada una de ellas (pueden ser diferentes), información detallada de sus características. La información de una GPU es un listado parecido a éste:

```
Device 0: ---
  Major revision number:                         ---
  Minor revision number:                         ---
  Total amount of global memory:                 --- bytes
  Number of multiprocessors:                     ---
  Number of cores:                               ---
  Total amount of constant memory:               --- bytes
  Total amount of shared memory per block:       --- bytes
```

```
Total number of registers available per block: ---
Warp size:                                     ---
Maximum number of threads per block:           ---
Maximum sizes of each dimension of a block:    --- x --- x ---
Maximum sizes of each dimension of a grid:     --- x --- x ---
Maximum memory pitch:                          --- bytes
Texture alignment:                             --- bytes
Clock rate:                                    --- GHz
Concurrent copy and execution:                 ---
...
```

Algunos de estos valores (básicamente los marcados en **rojo**) los deberemos tener en mente al programar con CUDA en este servidor. Por ejemplo, si lanzamos un kernel con más threads de los que soporta la GPU, el kernel no funcionará.

Cuando instalamos una GPU + CUDA este es el primer test que hay que correr. Primero para comprobar que la GPU funciona, y luego para obtener estos parámetros.

Ahora, si miráis el `job.sh`, hay un parámetro interesante:

```
#SBATCH --gres=gpu:1
```

Este comando le dice al gestor cuantas GPUs queréis utilizar. Podéis cambiarlo por este valor:

```
#SBATCH --gres=gpu:4
```

Y volver a ejecutar el programa de otra vez. ¿Qué ha ocurrido?

Es un buen momento para consultar el manual online de CUDA. La información que ofrece este test se obtiene llamando a la función `cudaGetDeviceProperties()`. Si consultamos la siguiente página: `docs.nvidia.com/cuda/cuda-runtime-api/` y buscamos esta rutina, veremos la gran cantidad de información que podemos obtener con ella.

Para acceder a la documentación de NVIDIA, es necesario tener un usuario registrado. Os recomendamos que os déis de alta en `https://developer.nvidia.com` para poder acceder a todos los recursos de CUDA que ofrece NVIDIA.

SaxpyP

En este test se calcula la operación `Saxpy` que vimos como ejemplo en las clases de teoría. Para compilar y ejecutar usaremos los ficheros que hay en el directorio. En nuestro caso:

```
make
sbatch job.sh
```

Si miráis el fichero (`main.cu`), veréis que tiene una estructura similar a la que hemos visto en clase de teoría con algunos añadidos. Lo más novedoso es la forma en que medimos los tiempos de ejecución. Teniendo en cuenta que, la ejecución de determinadas operaciones en la GPU es asíncrona con respecto a la CPU, no es posible usar las mismas rutinas que usaríamos en un programa convencional para medir el tiempo de ejecución.

Hemos de usar los eventos de CUDA (`cudaEvent_t`). Lo que hay que hacer es registrar los eventos necesarios entre el código a medir:

```
cudaEventRecord(E0, 0); cudaEventSynchronize(E0);
// Código a medir
cudaEventRecord(E1, 0); cudaEventSynchronize(E1);
cudaEventElapsedTime(&tiempo, E0, E1);
```

La rutina `cudaEventSynchronize` provocará que el programa espere hasta que el evento quede registrado en la GPU. Esto es imprescindible cuando trabajamos con operaciones asíncronas. La rutina `cudaEventElapsedTime` calcula el tiempo transcurrido (en ms) entre el registro de los 2 eventos.

Con este test se pueden realizar muchas pruebas. Os enumeramos algunas:

1. Compilad, ejecutad y comprobad que funciona correctamente.
2. Calculad los MFLOPS contando sólo el kernel.
3. Calculad los MFLOPS contando también las transferencias de datos.
4. Calculad el ancho de banda de las transferencias CPU → GPU, y GPU → CPU.
5. Calculad los mismos anchos de banda, pero ahora utilizando memoria “pinned”. El código necesario ya está escrito en un comentario.
6. La rutina que comprueba que el resultado es correcto, hace este test de error: `(abs(a-b)/a > Error)`. Cambiad el test de error por uno de igualdad `(a != b)`. ¿Qué ocurre en este caso?

7. Modificad job.sh para cambiar el tamaño del problema y el número de threads que se utiliza. No hay que modificar el código, sólo hay que invocar la ejecución así:

```
./SaxpyP.exe 16777216 1024
```

   Siendo 16777216 el tamaño del problema (N) y 1024 en número de threads (nThreads).

   Jugad con el número de threads a ver qué pasa.

Profiling

Como en muchos otros entornos, tenemos herramientas para analizar el rendimiento de una aplicación CUDA. La primera herramienta que utilizaremos es `nvprof`. La forma de invocarla es muy simple:

```
nsys nvprof --print-gpu-summary ./SaxpyP.exe
```

El resultado del profiling se escribirá en el fichero que nos devuelve la cola cuda. Parte de la información que ofrece `nvprof` es la siguiente:

- Estadísticas de ejecución de los kernels
- Estadísticas de las operaciones de comunicación CPU – GPU
- El tiempo que hemos gastado en cada una de las rutinas de CUDA
- …

Para más detalles se puede consultar `docs.nvidia.com/cuda/profiler-users-guide`.

Control de Errores en CUDA

En este último test hemos incluido el código necesario para comprobar los errores en CUDA. En CUDA, todas las llamadas devuelven un código de error (a excepción de la ejecución de un kernel). El código de error es un tipo predefinido de CUDA: `cudaError_t`. También existe una función de CUDA que devuelve el código del último error:

```
cudaError_t cudaGetLastError(void)
```

Para saber de que error se trata, disponemos de una función CUDA que devuelve un string con la descripción del error:

```
char* cudaGetErrorString(cudaError_t)
```

La rutina de error que podéis encontrar en el código es la siguiente:

```
void CheckCudaError(char sms[]) {
  cudaError_t error;

  error = cudaGetLastError();
  if (error) printf("%s:%s\n", sms, cudaGetErrorString(error));
}
```

Esta rutina se puede usar después de cada invocación a una rutina CUDA, como por ejemplo:

```
cudaMemcpy(H_y, d_y, numBytes, cudaMemcpyDeviceToHost);
CheckCudaError((char *) "Copiar Datos Device --> Host");
```

En caso de que la rutina **cudaMemcpy** produzca un error, nos aparecerá el mensaje:

```
 Copiar Datos Device --> Host: descripción del error
```

Para ver cómo funciona el control de errores podéis probar las siguientes cosas:

1. Modificad el número de threads, siempre con valores múltiplo de 32, hasta que el código deje de funcionar. Podéis probar con valores entre 32 y 1024. ¿Por qué no funciona? ¿Se puede arreglar?
2. Probad con 1024 + 32 threads. No funciona. ¿Porqué?

Recursos Adicionales

Junto con la distribución gratuita de CUDA que se puede obtener en la página web de NVIDIA podemos encontrar algunos recursos muy útiles. Estos recursos adicionales los podéis encontrar en: `/Soft/cuda/VER`[1]. En particular nos interesa lo siguiente:

- **Ejemplos**. En el directorio `/Soft/cuda/VER/samples` encontraréis numerosos ejemplos de cuda. Desde aplicaciones muy simples, hasta ejemplos avanzados. Estos ejemplos se pueden compilar fácilmente modificando los `Makefiles` de la sesión actual.
  Estos ejemplos actualizados, con la última versión, se pueden obtener de forma gratuita en: https://github.com/NVIDIA/cuda-samples

[1] **VER** depende de la versión de cuda instalada, nosotros estamos usando la 12.0.1.

- **Documentación**. En el directorio `/Soft/cuda/VER/doc/pdfs` se podían encontrar todos los manuales de cuda de la versión VER: manuales de consulta, de buenas prácticas, de optimización, … Este directorio existía hasta la versión 9. Para encontrar información más actualizada hay que visitar: `https://docs.nvidia.com/cuda` (ahí está la información de todas las versiones del compilador).
- **Herramientas**. En el directorio `/Soft/cuda/VER/tools` encontraréis el fichero Excel `CUDA_Occupancy_Calculator.xls`. La utilidad de esta herramienta la tendréis que investigar vosotros.