CUDA (Compute Unified Device Architecture)

CUDA (Compute Unified Device Architecture) は、NVIDIA 社の「GPU」に対する「GPGPU(Greneral Purpose Graphics Processing Unit)」を目的とした「フレームワーク」または「統合開発環境」である。

具体的には、GPU を動作させるための｛プログラミング・モデル」および「プログラミング言語」と、その「コンパイラ」、「ライブラリ」で構成されており、NVIDIA の GPU を使って、プログラミングして実行するためのソフトウエア一式のことである。

本内容の作成時点（2019/04/18）では、Google colaboratory には CUDA10.0 がインストール済みである。

1. CUDA のスレッド階層構造

スレッドは GPU（のカーネル）で動作するプログラムの最小単位です。1 つのスレッドが GPU の 1 つのストーリミングプロセッサー SP（コア）で実行される。

図 1 に「Tesla K80」の GPU 情報を示す。図 1 の赤枠部分がスレッドに関する制約を表わしている。

```
Device 0: "Tesla K80"
  CUDA Driver Version / Runtime Version          10.0 / 10.0
  CUDA Capability Major/Minor version number:    3.7
  Total amount of global memory:                 11441 MBytes (11996954624 bytes)
  (13) Multiprocessors, (192) CUDA Cores/MP:     2496 CUDA Cores
  GPU Max Clock rate:                            824 MHz (0.82 GHz)
  Memory Clock rate:                             2505 Mhz
  Memory Bus Width:                              384-bit
  L2 Cache Size:                                 1572864 bytes
  Maximum Texture Dimension Size (x,y,z)         1D=(65536), 2D=(65536, 65536), 3D=(4096, 4096, 4096)
  Maximum Layered 1D Texture Size, (num) layers  1D=(16384), 2048 layers
  Maximum Layered 2D Texture Size, (num) layers  2D=(16384, 16384), 2048 layers
  Total amount of constant memory:               65536 bytes
  Total amount of shared memory per block:       49152 bytes
  Total number of registers available per block: 65536
  Warp size:                                     32
  Maximum number of threads per multiprocessor:  2048
  Maximum number of threads per block:           1024
  Max dimension size of a thread block (x,y,z): (1024, 1024, 64)
  Max dimension size of a grid size    (x,y,z): (2147483647, 65535, 65535)
  Maximum memory pitch:                          2147483647 bytes
  Texture alignment:                             512 bytes
  Concurrent copy and kernel execution:          Yes with 2 copy engine(s)
  Run time limit on kernels:                     No
  Integrated GPU sharing Host Memory:            No
  Support host page-locked memory mapping:       Yes
  Alignment requirement for Surfaces:            Yes
  Device has ECC support:                        Enabled
  Device supports Unified Addressing (UVA):      Yes
  Device supports Compute Preemption:            No
  Supports Cooperative Kernel Launch:            No
  Supports MultiDevice Co-op Kernel Launch:      No
  Device PCI Domain ID / Bus ID / location ID:   0 / 0 / 4
  Compute Mode:
     < Default (multiple host threads can use ::cudaSetDevice() with device simultaneously) >

deviceQuery, CUDA Driver = CUDART, CUDA Driver Version = 10.0, CUDA Runtime Version = 10.0, NumDevs = 1
Result = PASS
```

図 1　Tesla K80 の GPU 情報

CUDA では多数のスレッドを「グリッド」、「(スレッド）ブロック」という概念を導入し、そのなかで階層的に管理している。

図 2 にスレッド階層構造の例を示す。ここでは、ホストの CPU からのカーネル関数 Kernel の呼び出しにより、GPU にグリッドサイズ (3, 2, 1)、ブロックサイズ (5, 3, 1) の 1 つのグリッドが構成される様子を示している。

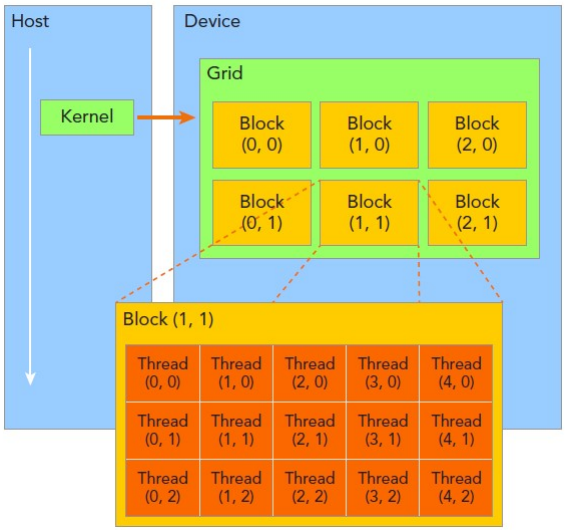

図 2　スレッド階層構造の例

図 1 のスレッドに関する制約は、それぞれ以下のようになっている。

(1) Maximum number of threads per block : 1024

⇒1 つのブロック内のスレッド数は 1024 以下でなければならないという制約がある。

(2) Max dimension size of a thread block (x, y, z) : (1024, 1024, 64)

⇒ ブロック内でスレッドは 3 次元的に管理されている。

また、各方向の数は、それぞれ (1024, 1024, 64) に制限される。

(1) の制約と合わせて (x, y, z) は

$$
\begin{aligned}
1 \leqq\ & x\ \leqq 1024 \\
1 \leqq\ & y\ \leqq 1024 \\
1 \leqq\ & z\ \leqq 64 \\
1 \leqq\ & xyz \leqq 1024
\end{aligned}
$$

をみたす整数に制限される。すなわち

$$
\begin{aligned}
(1024,\ 1,\ 1)\ &\bigcirc \\
(512,\ 2,\ 1)\ &\bigcirc \\
(2,\ 256,\ 2)\ &\bigcirc \\
(1024,\ 1024,\ 1)\ &\times \\
(1025,\ 1,\ 1)\ &\times
\end{aligned}
$$

となる。

(3) Max dimension size of a grid size (x, y, z) : (2147483647, 65535, 65535)

⇒1 グリッド内のブロックの数は 2147483647 × 65535 × 65535 個に制約されている。

(1)-(3) より、1 グリッド当たりの最大のスレッド数は

$$2147483647 \times 65535 \times 65535 \times 1024 \text{ 個}$$

となることが分かります。一方、SP 数は 2496 個でなので、32 スレッドを 1 ワープとした単位で順次 GPU のストーリミングマルチプロセッサー SMX に割り当てることで演算を実行する仕組みとなっている。

(4) Maximum number of threads per multiprocessor : 2048

⇒Kepler アーキテクチャの Tesla K80 の GPU では、ギガスレッドエンジンから SMX に割り付けられた（スレッド）ブロックのワープ（32 スレッド）群を 4 つのワープスケジューラに割り振

る。各ワープスケジューラは最大 16 個（表 1）のワープを分担し、これらのワープを並列に処理することができるようになっている。したがって、1 つの SMX には最大で

$$4 \times 16 \times 32 = 2048\,\text{個}$$

のスレッドが常駐できることを示している。

さらに、GPU にはハードウェアによる表 1 に示す制約（Compute Capability 3.7) がある。

	Tesla K80 (GK210)
Compute Capability	3.7
SM Version	sm_37
Threads / Warp	32
Warps / Multiprocessor	64
Threads / Multiprocessor	2048
Thread Blocks / Multiprocessor	16
Shared Memory / Multiprocessor (bytes)	114688
Max Shared Memory / Block (bytes)	49152
Register File Size / Multiprocessor (32-bit registers)	131072
Max Registers / Block	65536
Register Allocation Unit Size	256
Register Allocation Granularity	warp
Max Registers / Thread	255
Shared Memory Allocation Unit Size	256
Warp Allocation Granularity	4
Max Thread Block Size	1024
Shared Memory Size Configurations (bytes)	114688
[note: default at top of list]	98304
Warp register allocation granularities	256

表 1　GPU ハードウェアによる制約