CSCI-GA.3033-004

Graphics Processing Units (GPUs): Architecture and Programming

A Glimpse at the State-of-the-art

Mohamed Zahran (aka Z)
mzahran@cs.nyu.edu
http://www.mzahran.com
Volta Architecture
Quick Info

- Introduced in 2017
- Compute capability 7.0
- SM has tensor cores in addition to traditional ones
- 21.1 billion transistors
- 12 nm process technology
- NVLINK 2
  - V100 supports up to 6 NVLink links
  - 1 NVLINK provides 25 GB/s
- HBM2 global memory
  - 16GB
  - delivers 900 GB/sec peak memory bandwidth
- TDP (Thermal Design Power) level of 300W
Tensor Cores

- Tensor Cores are programmable matrix-multiply-and-accumulate units
- Each Tensor Core provides a matrix processing array which performs the operation $D = A \times B + C$, where $A$, $B$, $C$ and $D$ are $4 \times 4$ matrices
- The Tesla V100 GPU contains 640 Tensor Cores
- 8 per SM
Graphics Processing Clusters (GPCs) \(\rightarrow\) Texture Processing Clusters (TPCs) \(\rightarrow\) SMs
V100 in numbers

• Six GPCs
• 42 TPCs (each including two SMs)
• 84 Volta SMs, each SM contains:
  – 64 FP32 Cores
  – 64 INT32 Cores
  – 32 FP64 Cores
  – 8 Tensor Cores
• Eight 512-bit memory controllers (4096 bits total).
• 6144 KB of L2 cache
SM is partitioned into:

- **L1/shared combined:**
  - 128 KB/SM
  - 96 KB Shared Memory
  - All of it used as cache is no shared mem.

- Four processing blocks
- Each with:
  - 16 FP32 Cores
  - 8 FP64 Cores
  - 16 INT32 Cores
  - two Cores
  - a new L0 instruction cache
  - one warp scheduler
  - one dispatch unit
  - 64 KB Register File
Programming Wise: Cooperative Launch APIs

• What if you want to do synchronization with a smaller number of threads than a block? or bigger?

• The Cooperative Groups programming model describes synchronization patterns both within and across CUDA thread blocks.

• Also provides host-side APIs to launch grids whose threads are all guaranteed to be executing concurrently to enable synchronization across thread blocks.
Programming Wise: Independent Thread Scheduling

```c
if (threadIdx.x < 4) {
    A;
    B;
} else {
    X;
    Y;
}
Z;
```

Before

```
if (threadIdx.x < 4) {
    A;
    B;
} else {
    X;
    Y;
}
Z;
```

After
Programming Wise: Independent Thread Scheduling

Pre-Volta
- Program Counter (PC) and Stack (S)

32 thread warp

Volta
- Convergence Optimizer

32 thread warp with independent scheduling
if (threadIdx.x < 4) {
    A;
    B;
} else {
    X;
    Y;
}
Z;
__syncewarp()
Turing Architecture
Quick Info

• Introduced in 2018
• Introducing Ray-Tracing cores in addition to other core types
• Designed to handle real-time ray tracing.
## INTRODUCING TURING

**TU102 – FULL CONFIG**

**18.6 BILLION TRANSISTORS**

<table>
<thead>
<tr>
<th>Component</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>SM</td>
<td>72</td>
</tr>
<tr>
<td>CUDA CORES</td>
<td>4608</td>
</tr>
<tr>
<td>TENSOR CORES</td>
<td>576</td>
</tr>
<tr>
<td>RT CORES</td>
<td>72</td>
</tr>
<tr>
<td>GEOMETRY UNITS</td>
<td>36</td>
</tr>
<tr>
<td>TEXTURE UNITS</td>
<td>288</td>
</tr>
<tr>
<td>ROP UNITS</td>
<td>96</td>
</tr>
<tr>
<td>MEMORY</td>
<td>384-bit 7 GHz GDDR6</td>
</tr>
<tr>
<td>NVLINK CHANNELS</td>
<td>2</td>
</tr>
</tbody>
</table>

6 GPC → 36 TPC → 72 SMs
Each SM

- 1 Ray-Tracing core
- 64 CUDA Cores (i.e. SPs)
- 8 Tensor Cores
- 256 KB register file
- 96 KB of L1/shared memory
Conclusions

• More specialized cores are being added in new architectures.
• More fine grain synchronization are giving more control to the programmer.
• Still computation is the cheapest compared to memory access and communication.