How Have You Been Able to Gain Access to GPU Power?

There have been three ways:

1. Write a graphics display program (≥ 1985)

2. Write an application that looks like a graphics display program, but uses the fragment shader to do some per-node computation (≥ 2002)

3. Write in OpenCL or CUDA, which looks like C++ (≥ 2006)
Why do we care about GPU Programming?  
A History of GPU vs. CPU Performance

Note that the top of the graph on the previous page fits here.
The “Core-Score”. How can this be?

Due to the nature of graphics computations, GPU chips are customized to stream regular data. General CPU chips must be able to handle irregular data.

Another reason is that GPU chips do not need the significant amount of cache space that occupies much of the real estate on general-purpose CPU chips. The GPU die real estate can then be re-targeted to hold more cores and thus to produce more processing power.
Why have GPUs Been Outpacing CPUs in Performance?

Another reason is that general CPU chips contain on-chip logic to do **branch prediction** and **out-of-order execution**. This, too, takes up chip die space.

But, CPU chips can handle more general-purpose computing tasks.

So, which is better, a CPU or a GPU?

*It depends on what you are trying to do!*

Originally, GPU Devices were very task-specific
Today’s GPU Devices are not task-specific

Consider the architecture of the NVIDIA Tesla V100’s that we have in our DGX System

- 84 Streaming Multiprocessors (SMs) / chip
- 64 cores / SM
- Wow! 5,396 cores / chip? Really?
What is a “Core” in the GPU Sense?

Look closely, and you’ll see that NVIDIA really calls these “CUDA Cores”

Look even more closely and you’ll see that these CUDA Cores have no control logic – they are pure compute units. (The surrounding SM has the control logic.)

Other vendors refer to these as “Lanes”. You might also think of them as 64-way SIMD.

A Mechanical Equivalent…

“Streaming Multiprocessor”

“CUDA Cores”

“Data”
How Many Robots Do You See Here?

12? 72? Depends what you count as a "robot".

A Spec Sheet Example

<table>
<thead>
<tr>
<th>NVIDIA Card 4000 Series</th>
<th>Number of CUDA Cores</th>
<th>Size of Power Supply</th>
<th>Memory Type</th>
<th>Memory Interface Width</th>
<th>Memory Bandwidth (GB/sec)</th>
<th>Base Clock Speed</th>
<th>Boost Clock Speed</th>
<th>NOTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTX-4080</td>
<td>9920</td>
<td>750 watt</td>
<td>GDDR6</td>
<td>256 bit</td>
<td>710.8 GB/s</td>
<td>2.21 GHz</td>
<td>2.51 GHz</td>
<td>15 GB of Memory</td>
</tr>
<tr>
<td>RTX-4090</td>
<td>13848</td>
<td>350 watt</td>
<td>GDDR6</td>
<td>384 bit</td>
<td>1008 GB/s</td>
<td>2.25 GHz</td>
<td>2.52 GHz</td>
<td>24 GB of Memory</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NVIDIA Card 2000 Series</th>
<th>Number of CUDA Cores</th>
<th>Size of Power Supply</th>
<th>Memory Type</th>
<th>Memory Interface Width</th>
<th>Memory Bandwidth (GB/sec)</th>
<th>Base Clock Speed</th>
<th>Boost Clock Speed</th>
<th>NOTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTX-3070</td>
<td>5884</td>
<td>530 watt</td>
<td>GDDR6</td>
<td>256 bit</td>
<td>696 GB/s</td>
<td>1736 MHz</td>
<td>1770 MHz</td>
<td>Standard with 8 GB of Memory</td>
</tr>
<tr>
<td>RTX-3070 Ti</td>
<td>6764</td>
<td>530 watt</td>
<td>GDDR6</td>
<td>384 bit</td>
<td>1152 GB/s</td>
<td>1736 MHz</td>
<td>1770 MHz</td>
<td>Standard with 8 GB of Memory</td>
</tr>
<tr>
<td>RTX-3060</td>
<td>8104</td>
<td>750 watt</td>
<td>GDDR6</td>
<td>256 bit</td>
<td>760 GB/s</td>
<td>1440 MHz</td>
<td>1710 MHz</td>
<td>Standard with 16 GB of Memory</td>
</tr>
<tr>
<td>RTX-3060 Ti</td>
<td>10240</td>
<td>750 watt</td>
<td>GDDR6</td>
<td>384 bit</td>
<td>1152 GB/s</td>
<td>1736 MHz</td>
<td>1770 MHz</td>
<td>Standard with 16 GB of Memory</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NVIDIA Card 6000 Series</th>
<th>Number of CUDA Cores</th>
<th>Size of Power Supply</th>
<th>Memory Type</th>
<th>Memory Interface Width</th>
<th>Memory Bandwidth (GB/sec)</th>
<th>Base Clock Speed</th>
<th>Boost Clock Speed</th>
<th>NOTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTX-2060</td>
<td>1920</td>
<td>550 watt</td>
<td>GDDR6</td>
<td>256 bit</td>
<td>336 GB/s</td>
<td>1660 MHz</td>
<td>1660 MHz</td>
<td>Standard with 6 GB of Memory</td>
</tr>
<tr>
<td>RTX-2070 Super</td>
<td>2176</td>
<td>550 watt</td>
<td>GDDR6</td>
<td>256 bit</td>
<td>448 GB/s</td>
<td>1470 MHz</td>
<td>1660 MHz</td>
<td>Standard with 8 GB of Memory</td>
</tr>
<tr>
<td>RTX-2070</td>
<td>2564</td>
<td>550 watt</td>
<td>GDDR6</td>
<td>256 bit</td>
<td>448 GB/s</td>
<td>1470 MHz</td>
<td>1660 MHz</td>
<td>Standard with 8 GB of Memory</td>
</tr>
<tr>
<td>RTX-2080 Super</td>
<td>2944</td>
<td>550 watt</td>
<td>GDDR6</td>
<td>256 bit</td>
<td>448 GB/s</td>
<td>1515 MHz</td>
<td>1710 MHz</td>
<td>Standard with 8 GB of Memory</td>
</tr>
<tr>
<td>RTX-2080</td>
<td>3372</td>
<td>550 watt</td>
<td>GDDR6</td>
<td>256 bit</td>
<td>496 GB/s</td>
<td>1515 MHz</td>
<td>1815 MHz</td>
<td>Standard with 8 GB of Memory</td>
</tr>
<tr>
<td>RTX-2080 Ti</td>
<td>3852</td>
<td>550 watt</td>
<td>GDDR6</td>
<td>256 bit</td>
<td>512 GB/s</td>
<td>1515 MHz</td>
<td>1815 MHz</td>
<td>Standard with 11 GB of Memory</td>
</tr>
<tr>
<td>Titan RTX</td>
<td>4096</td>
<td>650 watt</td>
<td>GDDR6</td>
<td>384 bit</td>
<td>672 GB/s</td>
<td>1350 MHz</td>
<td>1770 MHz</td>
<td>Standard with 24 GB of Memory</td>
</tr>
</tbody>
</table>

NVIDIA
NVIDIA 4090 Spec Sheet

<table>
<thead>
<tr>
<th>Graphics Card</th>
<th>Relative Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphics Core</td>
<td></td>
</tr>
<tr>
<td>Memory Core</td>
<td></td>
</tr>
<tr>
<td>Processor</td>
<td></td>
</tr>
<tr>
<td>Main System</td>
<td></td>
</tr>
<tr>
<td>Memory</td>
<td></td>
</tr>
<tr>
<td>Graphics</td>
<td></td>
</tr>
</tbody>
</table>

NVIDIA's Ampere Chip

- **Memory**
  - Memory Size: 24 GB GDDR6X
  - Memory Type: GDDR6X
  - Memory Bus: 192-bit
  - Memory Bandwidth: 192 GB/s

- **Graphics Features**
  - Clocks: 1,280 MHz (CSP, 1,400 MHz (DP, Mac))
The Bottom Line is This

It is obvious that it is difficult to directly compare a CPU with a GPU. They are optimized to do different things.

So, let’s use the information about the architecture as a way to consider what CPUs should be good at and what GPUs should be good at

<table>
<thead>
<tr>
<th>CPU</th>
<th>GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>General purpose programming</td>
<td>Data parallel programming</td>
</tr>
<tr>
<td>Multi-core under user control</td>
<td>Little user control</td>
</tr>
<tr>
<td>Irregular data structures</td>
<td>Regular data structures</td>
</tr>
<tr>
<td>Irregular flow control</td>
<td>Regular Flow Control</td>
</tr>
</tbody>
</table>

BTW,
The general term in the OpenCL world for an SM is a **Compute Unit**.
The general term in the OpenCL world for a CUDA Core is a **Processing Element**.

Compute Units and Processing Elements are Arranged in Grids

A GPU Platform can have one or more **Devices**.

A GPU **Device** is organized as a grid of **Compute Units**.

Each Compute Unit is organized as a grid of **Processing Elements**.

So in NVIDIA terms, their new V100 GPU has 84 Compute Units, each of which has 64 Processing Elements, for a grand total of 5,396 Processing Elements.
Thinking ahead to CUDA and OpenCL...

How can GPUs execute General C Code Efficiently?

• Ask them to do what they do best. Unless you have a very intense Data Parallel application, don't even think about using GPUs for computing.

• GPU programs expect you to not just have a few threads, but to have thousands of them!

• Each thread executes the same program (called the kernel), but operates on a different small piece of the overall data.

• Thus, you have many, many threads, all waking up at about the same time, all executing the same kernel program, all hoping to work on a small piece of the overall problem.

• CUDA and OpenCL have built-in functions so that each thread can figure out which thread number it is, and thus can figure out what part of the overall job it's supposed to do.

• When a thread gets blocked somehow (a memory access, waiting for information from another thread, etc.), the processor switches to executing another thread to work on.

So, the Trick is to Break your Problem into Many, Many Small Pieces

Particle Systems are a great example.

1. Have one thread per each particle.
2. Put all of the initial parameters into an array in GPU memory.
3. Tell each thread what the current Time is.
4. Each thread then computes its particle’s position, color, etc. and writes it into arrays in GPU memory.
5. The CPU program then initiates OpenGL drawing of the information in those arrays.

Note: once setup, the data never leaves GPU memory!
Tensor Cores Accelerate Fused-Multiply-Add Arithmetic

\[ D = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \\ a_5 & a_6 & a_7 & a_8 \\ a_9 & a_{10} & a_{11} & a_{12} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} + \begin{bmatrix} c_1 & c_2 & c_3 & c_4 \\ c_5 & c_6 & c_7 & c_8 \\ c_9 & c_{10} & c_{11} & c_{12} \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{bmatrix} \]

cuBLAS Mixed-Precision GEMM
(FP16 Input, FP32 Compute)

Relative Performance

Matrix Size (M=N=K)

- Tesla P100
- Tesla V100
- Tensor Cores
What is Fused Multiply-Add?

Many scientific and engineering computations take the form:

\[ D = A + (B \times C); \]

A “normal” multiply-add would likely handle this as:

\[ \text{tmp} = B \times C; \]
\[ D = A + \text{tmp}; \]

A “fused” multiply-add does it all at once, that is, when the low-order bits of \( B \times C \) are ready, they are immediately added into the low-order bits of \( A \) at the same time the higher-order bits of \( B \times C \) are being multiplied.

Consider a Base 10 example: \( 789 + (123 \times 456) \)

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>x 456</td>
</tr>
<tr>
<td>738</td>
<td></td>
</tr>
<tr>
<td>615</td>
<td></td>
</tr>
<tr>
<td>492</td>
<td></td>
</tr>
<tr>
<td>+789</td>
<td></td>
</tr>
<tr>
<td>56,877</td>
<td>Can start adding the 9 the moment the 8 is produced!</td>
</tr>
</tbody>
</table>

Note: “Normal” \( A + (B \times C) \neq “FMA” A + (B \times C) \)

There are Two Approaches to Combining CPU and GPU Programs

1. Combine both the CPU and GPU code in the same file. The CPU compiler compiles its part of that file. The GPU compiler compiles just its part of that file.

2. Have two separate programs: a .cpp and a .somethingelse that get compiled separately.

Advantages of Each

1. The CPU and GPU sections of the code know about each others’ intents. Also, they can share common structs, #define’s, etc.

2. It’s potentially cleaner to look at each section by itself. Also, the GPU code can be easily used in combination with other CPU programs.

Who are we Talking About Here?

\( 1 = \) NVIDIA’s CUDA
\( 2 = \) Khronos’s OpenCL

We will talk about each of these separately – stay tuned!
Looking ahead:
If threads all execute the same program, what happens on flow divergence?

```plaintext
if( a > b )
  Do This;
else
  Do That;
```

1. The line “if( a > b )” creates a vector of Boolean values giving the results of the if-statement for each thread. This becomes a “mask”.
2. Then, the GPU executes all parts of the divergence:
   Do This;
   Do That;
3. During that execution, anytime a value wants to be stored, the mask is consulted and the storage only happens if that thread's location in the mask is the right value.

• GPUs were originally designed for the streaming-ness of computer graphics
• Now, GPUs are also used for the streaming-ness of data-parallel computing
• GPUs are better for some things. CPUs are better for others.
This is an Nvidia 1080 ti card – one that died on us. It willed its body to education.

Removing the covers:
Dismantling a Graphics Card

Removing the heat sink:

This transfers heat from the GPU Chip to the cooling fins.

Dismantling a Graphics Card

Removing the fan assembly reveals the board:

GPU Chip  Memory
Dismantling a Graphics Card

Power half of the board:

Dismantling a Graphics Card

Graphics half of the board:

This one contains 7.2 billion transistors!
The newer cards contain 70+ billion transistors.
(Thank you, Moore’s Law)
Dismantling a Graphics Card
Underside of the board:

Here is a fun video of someone explaining the different parts of this same card:
https://www.youtube.com/watch?v=dSCNf9DIBGE