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.

Theoretical GFLOPs:
- NVIDIA GPU Single Precision
- NVIDIA GPU Double Precision
- Intel CPU Single Precision
- Intel CPU Double Precision

Years:
- 2000
- 2005
- 2010
- 2015

NVIDIA GPUs:
- GeForce GTX
- Tesla

Note that the top of the graph on the previous page fits here.
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.

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, parts of GPU chips were very task-specific.

Today’s GPU devices are not task-specific – they can be dynamically re-purposed for any GPU function.

Consider the architecture of the NVIDIA 4090.

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"

A Spec Sheet Example

Another Spec Sheet Example

NVIDIA’s Ampere Chip
The Bottom Line is This

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.

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 CPUs 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!
Something New – Tensor Cores

Tensor Cores Accelerate Fused-Multiply-Add Arithmetic

What is Fused Multiply-Add?

Many scientific and engineering computations take the form:
\[ D = A + (B^C); \]

A “normal” multiply-add would likely handle this as:
\[ \text{tmp} = B^C; \]
\[ D = A + \text{tmp}; \]

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

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

\[
\begin{align*}
123 \\
x456 \\
738 \\
615 \\
492 \\
+789 \\
56,877
\end{align*}
\]

Can start adding the 9 the moment the 8 is produced!

Note: “Normal” \(A+(B^C)\) ≠ “FMA” \(A+(B^C)\)

Something Even Newer – Ray-Trace Cores

https://www.youtube.com/watch?v=QL7sXc2INJ8

NVIDIA
There are Two Approaches to Combining CPU and GPU Programs

1. Combine both the CPU and GPU code in the same code file. Somehow mark what part is CPU code and what part is GPU code. The CPU compiler compiles just 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 by a CPU compiler and a GPU compiler.

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

Looking ahead:

If threads all execute the same program, what happens on flow divergence?

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

1. On a GPU, the line “if (a > b)” creates a vector of 0/1 Boolean values giving the results of the if-statement for each thread. This becomes a “bitmask”.

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.

Dismantling a Graphics Card

This is an Nvidia 1080 ti card – one that died on us. It willed its body to education.

• GPUs were originally designed for the streaming-ness of computer graphics
• That same streaming-ness can now also be applied to data-parallel computing
• GPUs are better for some things. CPUs are better for others.
Dismantling a Graphics Card

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:

Dismantling a Graphics Card

Power half of the board:

Power distribution
Power input
Graphics half of the board:

This one contains 7.2 billion transistors!
The newer cards contain 70+ billion transistors.
(Thank you, Moore’s Law)

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=dSCN9DIBGE