LOW-LATENCY GPGPU

A 5-minute intro and investigation
Disclaimer

- These findings reflect the point of view of someone who’s been courting only CUDA in a hobbyist setting since 2010, and in a (lightweight) professional setting since 2017

- I’d love to hear the viewpoints of AMD, Intel, Direct3D, Metal, and Vulkan folks on this - hit me up afterwards!
  - And maybe give me some hardware to play with...?
So, GPGPUs have latency issues...

- Calling GPU functions takes time
- Moving memory around takes a lot of time
- *The GPU Driver* takes its sweet, sweet time

- ...right?
TIME TO FINISH
SILLY RACE: + 20

EVERYTHING YOU KNOW IS WRONG
10+ years of GPGPU

- APIs are slimmer!
  - (D3D12 / Vulkan / Metal vs. OpenGL/D3D9)
- GPUs are faster!
- PCI Express is faster!

- Most of all, **Drivers are faster!**
Literally 2 weeks ago: NVIDIA introduces “Ultra Low Latency Mode”
“Latency” is relative

<table>
<thead>
<tr>
<th>Domain</th>
<th>Acceptable Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Folding Simulation</td>
<td>Days</td>
</tr>
<tr>
<td>Weather Simulation</td>
<td>Hours</td>
</tr>
<tr>
<td>Radar Signal Convolution</td>
<td>200~300ms</td>
</tr>
<tr>
<td>Videogame</td>
<td>10~30ms</td>
</tr>
<tr>
<td>Audio Processing</td>
<td>5~10ms</td>
</tr>
<tr>
<td>High Frequency Stock Trading</td>
<td>&lt;1ms</td>
</tr>
</tbody>
</table>
“Latency” is relative

<table>
<thead>
<tr>
<th>Domain</th>
<th>Acceptable Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Folding Simulation</td>
<td>Days</td>
</tr>
<tr>
<td>Weather Simulation</td>
<td>Hours</td>
</tr>
<tr>
<td>Radar Signal Convolution</td>
<td>200~300ms</td>
</tr>
<tr>
<td>Videogame</td>
<td>10~30ms</td>
</tr>
<tr>
<td>Audio Processing</td>
<td>5~10ms</td>
</tr>
<tr>
<td>High Frequency Stock Trading</td>
<td>&lt;1ms</td>
</tr>
</tbody>
</table>
Experiment: Real-time FM Synth

Solar Modulation - Savaged Regime
Test subjects

**GeForce 640M**
*(this computer)*
- Kepler Architecture, SM 3.0, 2012
- 2GB GDDR5 / 128bit / 900MHz
- ~390 Gflops
  - 2x PlayStation 3
  - Intel UHD 620
- ~25 Giops

**Quadro P400**
*(office workstation)*
- Pascal Architecture, SM 6.1, 2017
- 2GB GDDR5 / 64bit / 2GHz
- ~630 Gflops
  - ½ Xbox One
  - 2x Intel Iris 5100
- ~200 Giops
DEMO TIME!
THANK YOU!

And go do something awesome with that GPU of yours!
Image Sources

- Screenshot: Marble Madness, c. Atari 1984
- FM Music Video: Savaged Regime