Probabilistic Memory Consistency Specifications

Reese Levine and Tyler Sorensen
Motivation

Can we apply *approximate computing* techniques to *relaxed memory consistency specifications*?

- Parallel Application
- weakened memory operations
- Device A: implementing a *relaxed MCS*
  - Traditional MCS
    - Accuracy: ??
  - Probabilistic MCS
    - Accuracy: 98.5%
Methodology

Overview of data collection and PMCS evaluation against shared memory programs

GPUHarbor

Weak behavior collection framework

GPU Framework

e.g. WebGPU, Vulkan

Collect data on devices

PMCS

Synthesize probabilistic MCS

Accuracy prediction

Shared Memory Program

Run on device

Validate results
Initial Results

https://gpuharbor.ucsc.edu/webgpu-mem-testing/

Test Parameters
- Testing Workgroups: 610
- Maximum Workgroups: 1034
- Workgroup Shuffle Percentage: 100
- Barrier Percentage: 100
- Scratch Memory Size: 3048
- Memory Stride: 4

Litmus Test

<table>
<thead>
<tr>
<th>Device</th>
<th>MP</th>
<th>S</th>
<th>R</th>
<th>LB</th>
<th>SB</th>
<th>2+2W</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARM Mali - G71</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ARM Mali - G78</td>
<td>0.23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Qualcomm Adreno 610</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Qualcomm Adreno 640</td>
<td>0</td>
<td>0.13</td>
<td>0</td>
<td>0.09</td>
<td>0.13</td>
<td>0.17</td>
</tr>
<tr>
<td>Qualcomm Adreno 642L</td>
<td>0</td>
<td>0.22</td>
<td>0</td>
<td>0.15</td>
<td>0.2</td>
<td>0.27</td>
</tr>
<tr>
<td>Qualcomm Adreno 660</td>
<td>0.01</td>
<td>0.59</td>
<td>0</td>
<td>0.28</td>
<td>0.64</td>
<td>0.58</td>
</tr>
<tr>
<td>PowerVR GE8320</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NVIDIA Tegra X1</td>
<td>0.0006</td>
<td>0.0007</td>
<td>0.0009</td>
<td>0.0009</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>