Resource Sharing for GPUs

Vishwesh Jatala (IITK), Jayvant Anantpur (IISc), Amey Karkare (IITK)

Problem

Resource allocation at thread block granularity in Graphics Processing Units (GPUs) has the following disadvantages [1]:

1. Resource underutilization
   - Registers and Scratchpad Memory get underutilized
2. Reduction in thread level parallelism (TLP)
   - Limited number of resident threads and blocks in streaming multiprocessors (SMs)
3. Potential reduction in throughput

Solution: Resource Sharing

Idea: Share the resources between thread blocks

Strategy:
- Increase the TLP by launching additional thread blocks in each SM
- Minimize the resource wastage with the help of addition thread blocks that:
  1. Use wasted resource
  2. Share the resources with other resident blocks
- Access resources effectively to avoid deadlocks and to guarantee minimum number of blocks that always make progress.

Motivating Example

Number of resources per SM: 35K Units
Resource requirement per block: 10K Units
Thread block size: 10 Warps

Resource Access Mechanism

A shared warp can access unshared register directly, but it can access shared register only after acquiring an exclusive lock.

Optimizations

Type of warps in the SM:
- Unshared warps (warps from unshared thread block)
- Owner warps (warps that have exclusive lock)
- Non-owner warps (warps without lock)

1. Owner Warp First (OWF):
   - Schedule the warps according to the priority: owner warp, unshared warp, and non-owner warp

2. Unroll Register Declarations:
   - Unroll and re-order register declarations to delay access to shared registers

3. Dynamic Warp Execution (Dyn):
   - Control the execution of long latency instructions from non-owner warps to reduce cache misses.

Example (Resource Sharing)

Resource Sharing: 50%

Example (Resource Sharing)

Number of resources per SM: 35K Units
Resource requirement per block: 10K Units
Thread block size: 10 Warps

Results

Note: First row of the results corresponds to register sharing, second row corresponds to scratchpad sharing

Experimental Setup

<table>
<thead>
<tr>
<th>Resource</th>
<th>GPU Configuration [2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of SMs</td>
<td>14</td>
</tr>
<tr>
<td>Max Num of TBs</td>
<td>8</td>
</tr>
<tr>
<td>Max Num of Threads</td>
<td>1536</td>
</tr>
<tr>
<td>Number of Registers</td>
<td>32768</td>
</tr>
<tr>
<td>Scratchpad Memory</td>
<td>16KB</td>
</tr>
<tr>
<td>Warp Scheling</td>
<td>1KB</td>
</tr>
</tbody>
</table>

References