Brook for GPUs
Ian Buck, Tim Foley, Daniel Horn, Jeremy Sugerman, Kayvon Fatahalian, Mike Houston, and Pat Hanrahan

Development
- Parallel effort to RStream compiler
- Developed at Stanford
- Brook for GPUs: http://brook.sourceforge.net
- Friendly programming environment
- Programmer need not know GL

Modern GPU Bandwidth Hierarchy
(Values provided are observed bandwidth on ATI Radeon 9800XT)

Future Merrimac Bandwidth Hierarchy

In the News…
- open source
  - http://sourceforge.net/projects/brook
- over 5,600 downloads in 5 months
- CIDS page hits
- GPGPU SIGGRAPH Course
  - in the news
- IEEE Computer
  - http://slashdot.org
  - http://opengl.org

Dense Matrix-Matrix Multiplication on the GPU and Pentium 4

<table>
<thead>
<tr>
<th>Multiplication of 1024 x 1024 Matrices</th>
<th>GPUs</th>
<th>Peak</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>NV GeForce FX 5900 Ultra</td>
<td>3.04</td>
<td>86%</td>
<td>8.07</td>
</tr>
<tr>
<td>ATI Radeon X800 X1</td>
<td>4.83</td>
<td>18%</td>
<td>12.06</td>
</tr>
<tr>
<td>Pentium 4 3GHz</td>
<td>7.78</td>
<td>68%</td>
<td>27.68</td>
</tr>
</tbody>
</table>

Maximum Observable FLOPS and Bandwidth

<table>
<thead>
<tr>
<th>Multiplication of 1024 x 1024 Matrices</th>
<th>GPUs</th>
<th>Cache BW</th>
<th>Seq Read BW</th>
</tr>
</thead>
<tbody>
<tr>
<td>NV GeForce FX 5900 Ultra</td>
<td>39.99</td>
<td>11.08</td>
<td>4.40</td>
</tr>
<tr>
<td>ATI Radeon X800 X1</td>
<td>26.4</td>
<td>12.20</td>
<td>7.55</td>
</tr>
<tr>
<td>Pentium 4 3GHz</td>
<td>41.98</td>
<td>14.20</td>
<td>7.55</td>
</tr>
</tbody>
</table>

GPU Results

- Compute
  - 32 2.4GHz P4 Xeons
  - 16GB DDR
  - 1.2GB disk
  - Intel E7505 chipset
  - Network
    - infiniband 4X
    - Gig6
  - Graphcs
    - ATI Radeon 9800
    - Pro 26M8

Future Work

GPU Clusters
- Matrix-Matrix multiplication on Merrimac Simulator achieves 80% of 128 GFLOP peak
- Fast cross-kernel communication
- Bllocking computation in SRF, Large block sizes
- Hand-tuned Merrimac implementation expected to achieve nearly 95% of peak.

Analyzing Matrix-Matrix Multiplication on Merrimac

- Richer memory hierarchy (faster/larger on-chip caches, SRF) required to keep GPU arithmetic units busy.

Future Work

- Richer memory hierarchy (faster/larger on-chip caches, SRF) required to keep GPU arithmetic units busy.

Applications
- ray-tracer
- BLAS
- linear algebra:
  - BLAS: SAXPY & SGEMV
  - Network
    - infiniband 4X
    - Gig6
  - Graphics
    - ATI Radeon 9800
    - Pro 26M8

Brook for GPUs
Ian Buck, Tim Foley, Daniel Horn, Jeremy Sugerman, Kayvon Fatahalian, Mike Houston, and Pat Hanrahan

Future Merrimac Bandwidth Hierarchy

In the News…
- open source
  - http://sourceforge.net/projects/brook
- over 5,600 downloads in 5 months
- CIDS page hits
- GPGPU SIGGRAPH Course
  - in the news
- IEEE Computer
  - http://slashdot.org
  - http://opengl.org

Dense Matrix-Matrix Multiplication on the GPU and Pentium 4

<table>
<thead>
<tr>
<th>Multiplication of 1024 x 1024 Matrices</th>
<th>GPUs</th>
<th>Peak</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>NV GeForce FX 5900 Ultra</td>
<td>3.04</td>
<td>86%</td>
<td>8.07</td>
</tr>
<tr>
<td>ATI Radeon X800 X1</td>
<td>4.83</td>
<td>18%</td>
<td>12.06</td>
</tr>
<tr>
<td>Pentium 4 3GHz</td>
<td>7.78</td>
<td>68%</td>
<td>27.68</td>
</tr>
</tbody>
</table>

Maximum Observable FLOPS and Bandwidth

<table>
<thead>
<tr>
<th>Multiplication of 1024 x 1024 Matrices</th>
<th>GPUs</th>
<th>Cache BW</th>
<th>Seq Read BW</th>
</tr>
</thead>
<tbody>
<tr>
<td>NV GeForce FX 5900 Ultra</td>
<td>39.99</td>
<td>11.08</td>
<td>4.40</td>
</tr>
<tr>
<td>ATI Radeon X800 X1</td>
<td>26.4</td>
<td>12.20</td>
<td>7.55</td>
</tr>
<tr>
<td>Pentium 4 3GHz</td>
<td>41.98</td>
<td>14.20</td>
<td>7.55</td>
</tr>
</tbody>
</table>

GPU Results

- Compute
  - 32 2.4GHz P4 Xeons
  - 16GB DDR
  - 1.2GB disk
  - Intel E7505 chipset
  - Network
    - infiniband 4X
    - Gig6
  - Graphcs
    - ATI Radeon 9800
    - Pro 26M8

Future Work

GPU Clusters
- Matrix-Matrix multiplication on Merrimac Simulator achieves 80% of 128 GFLOP peak
- Fast cross-kernel communication
- Bllocking computation in SRF, Large block sizes
- Hand-tuned Merrimac implementation expected to achieve nearly 95% of peak.

Analyzing Matrix-Matrix Multiplication on Merrimac

- Richer memory hierarchy (faster/larger on-chip caches, SRF) required to keep GPU arithmetic units busy.

Future Work

- Richer memory hierarchy (faster/larger on-chip caches, SRF) required to keep GPU arithmetic units busy.

Applications
- ray-tracer
- BLAS
- linear algebra:
  - BLAS: SAXPY & SGEMV
  - Network
    - infiniband 4X
    - Gig6
  - Graphics
    - ATI Radeon 9800
    - Pro 26M8