Different Matrix Multiplication Routines in OpenCL
Kazuya Matsumoto, Naohito Nakasato, Stanislav G. Sedukhin
The University of Aizu, JAPAN
{kaizya-m, nakasato, sedukhin}@u-aizu.ac.jp

This poster presents our implementation of different matrix-matrix multiplication routines written in OpenCL (Open Computing Language). The routines are GEMM (General Matrix-Matrix Multiply), SYMM (Symmetric Matrix-Matrix Multiply), SYRK (Symmetric Rank-K Update), SYR2K (Symmetric Rank-2K Update), and TRMM (Triangular Matrix-Matrix Multiply) in Level-3 BLAS (Basic Linear Algebra Subprograms). We evaluated the performance on various GPUs (AMD Radeon HD 7970, FirePro W8000, Radeon HD 6970, NVIDIA GeForce GTX Titan, and Tesla K20c), an accelerator (Intel Xeon Phi 5110P), and a CPU (Intel Core i7 3960X).

Implementation

Auto-tuning System for GEMM
- GEMM code generator
- Input: Parameter set
- List of major parameters:
  - Blocking factors related to work-group size
  - Blocking factors related to work-item size
  - Vector variable width
  - Usage of local memory (shared memory)
  - GEMM algorithm
  - Matrix storage layout
- Output: GEMM kernel in OpenCL
- Search Engine
  - Heuristically finds the best (fastest) GEMM kernel by measuring performance of many kernel patterns.

Using the Best GEMM Kernel
- Copying matrix data into storage layouts required to utilize the best GEMM kernel
- Example: $C \leftarrow AB + C$ when $C \leftarrow AB^T + C$
  - GEMM kernel is the best
- Matrix Multiply

Other Routines
- SYMM = “Side = Left, Uplo = Lower”
- SYRK = “Uplo = Left, TransA = NoTrans”
- SYR2K = “Uplo = Upper, TransAB = NoTrans”
- TRMM = “Side = Left, Uplo = Upper, TransA = NoTrans”

OpenCL Device Specifications

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Radeon HD 7970</td>
<td>AMD</td>
<td>Southern Islands</td>
<td>Tahiti</td>
<td>925</td>
<td>1024</td>
<td>4096</td>
<td>947</td>
<td>5789</td>
<td>3</td>
<td>264</td>
</tr>
<tr>
<td>FirePro W8000</td>
<td>AMD</td>
<td>Northern Islands</td>
<td>Cayman</td>
<td>900</td>
<td>896</td>
<td>3584</td>
<td>806</td>
<td>3224</td>
<td>4</td>
<td>288</td>
</tr>
<tr>
<td>Radeon HD 6970</td>
<td>NVIDIA</td>
<td>Kepler</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeForce GTX Titan</td>
<td>NVIDIA</td>
<td>GK110</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tesla K20c</td>
<td>NVIDIA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xeon Phi 5110P</td>
<td>Intel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Core i7 3960X</td>
<td>Intel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Performance on Radeon HD 7970

Maximum Performance on Various Devices