Different Matrix Multiplication Routines in OpenCL

Kazuya Matsumoto, Naohito Nakasato, Stanislav G. Sedukhin
The University of Aizu, JAPAN
{kazuya-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).

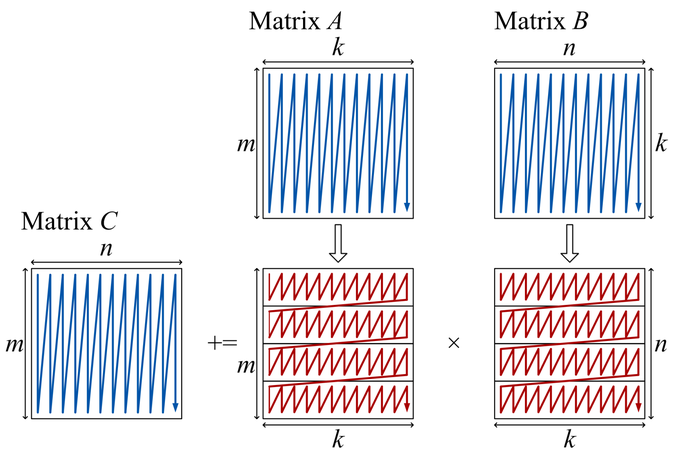
Implementation

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.

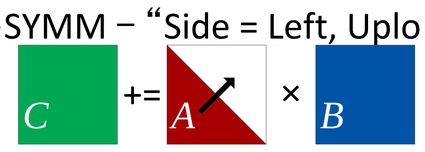
Using the Best GEMM Kernel

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

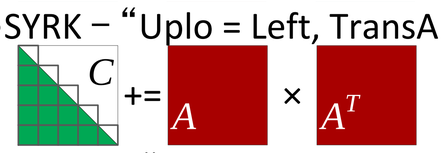


Other Routines

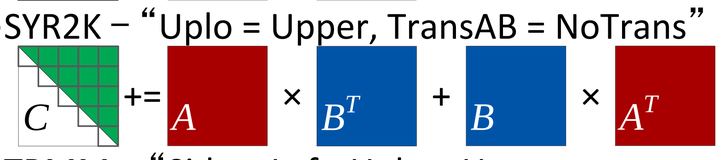
- SYMM – "Side = Left, Uplo = Lower"



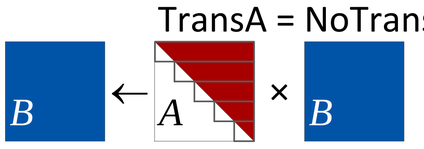
- SYRK – "Uplo = Left, TransA = NoTrans"



- SYR2K – "Uplo = Upper, TransAB = NoTrans"



- TRMM – "Side = Left, Uplo = Upper, TransA = NoTrans"

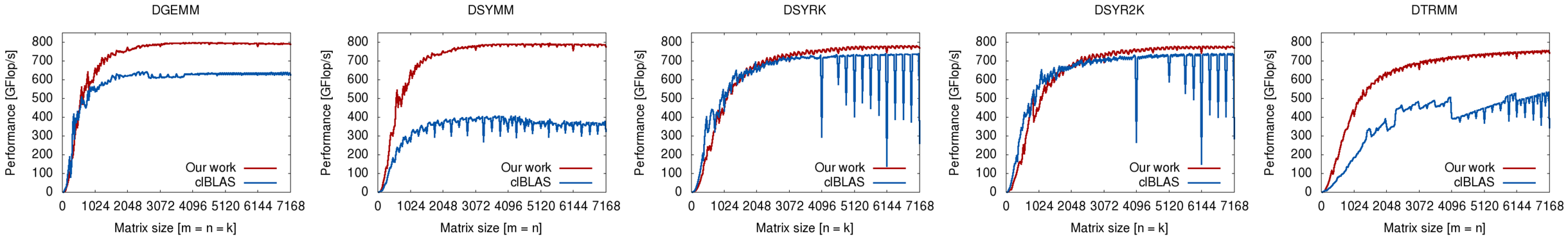


Performance

OpenCL Device Specifications

Device name	Radeon HD 7970	FirePro W8000	Radeon HD 6970	GeForce GTX Titan	Tesla K20c	Xeon Phi 5110P	Core i7 3960X
Vendor	AMD			NVIDIA		Intel	
Architecture	Southern Islands		Northern Islands	Kepler		MIC	Sandy Bridge
Codename	Tahiti		Cayman	GK110		Knights Corner	-
Core clock [MHz]	925	900	880	876	706	1053	3300
Maximum double-precision operations per clock	1024	896	768	1792	1664	944	48
Maximum single-precision operations per clock	4096	3584	3072	5376	4992	1888	96
Peak double-precision performance [GFlop/s]	947	806	676	1570	1175	994	158
Peak single-precision performance [GFlop/s]	3789	3224	2916	4709	3524	1988	317
Global memory size [GBytes]	3	4	1	6	5	8	-
Peak global memory bandwidth [GByte/s]	264	1	176	288	208	320	-

Performance on Radeon HD 7970



Maximum Performance on Various Devices