Outline

Evalutation & Feedback

- Help me to help you ...

https://docs.google.com/document/d/1Dxim2gU2zEMYGOpnTO2W_CptoKwa3IowM-v-1qIiXyU/edit?pli=1

NVIDIA Kepler - Overview

©NVIDIA Kepler Whitepaper

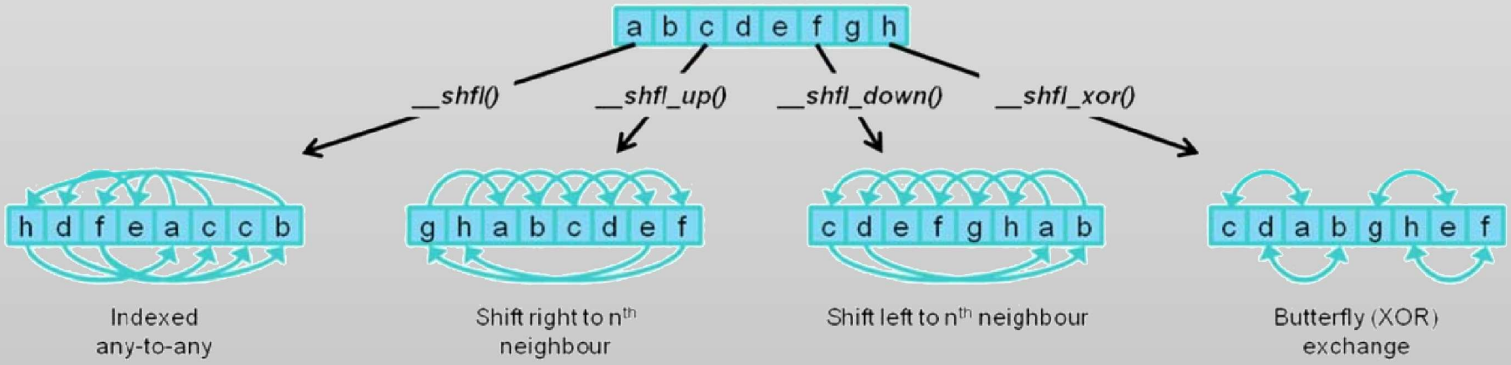
NVIDIA Kepler - Overview

- Compute capability 3.x
- 13-14 SMXs
- 2496-2688cores @ 0.73 Ghz
- Up to 3.95 TFLOPS SP
- Up to 1.31 TFLOPS DP
- Up to 6 GB global memory
  - Bandwidth 250 GB/sec
- Architectural features
  - Hyper-Q
  - Dynamic Parallelism

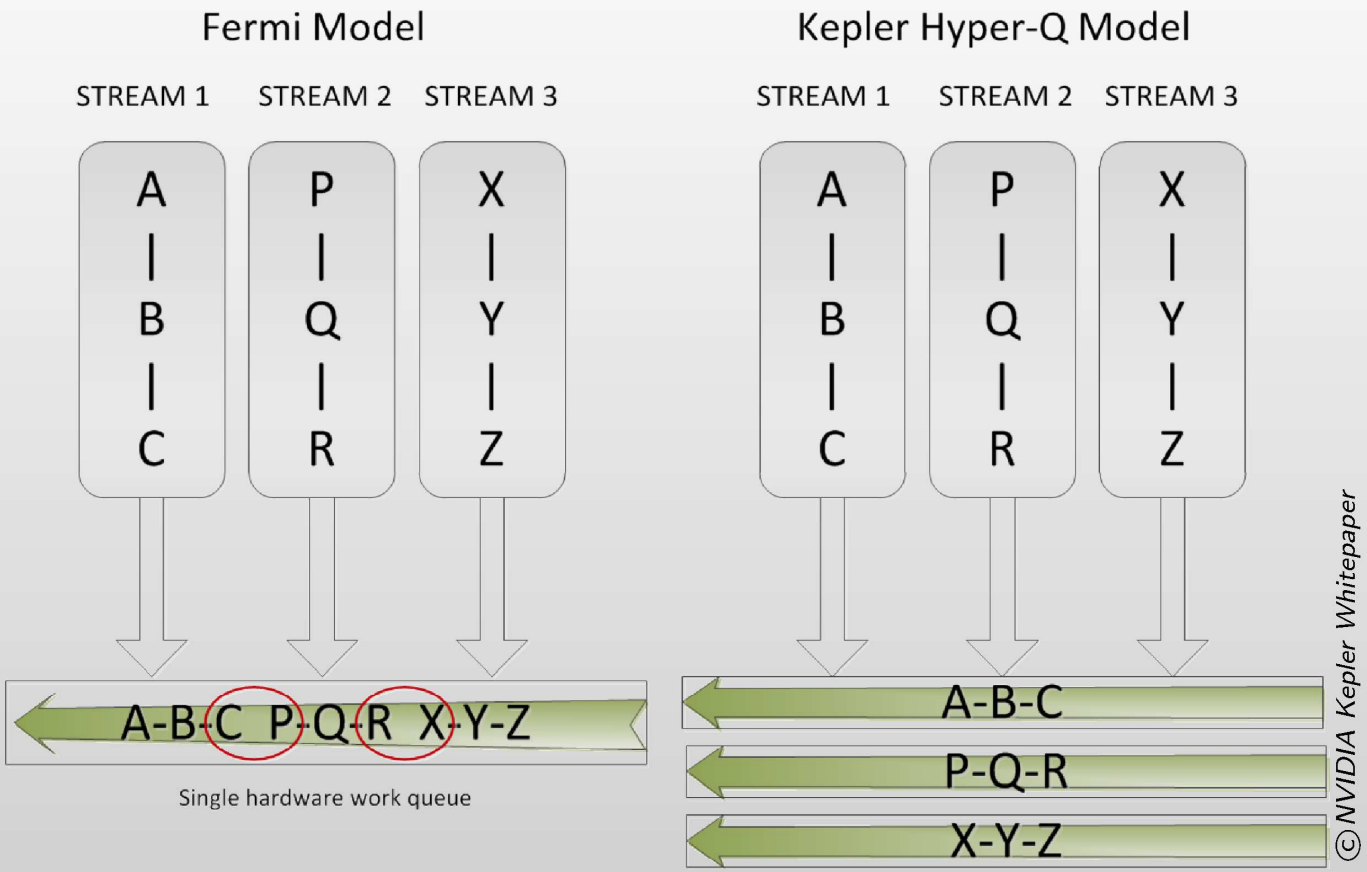
©NVIDIA Kepler Whitepaper

- Faster atomics
- 48KB read-only data cache
- Improved L2 cache
- PCIe 3.0
- Shuffle instructions

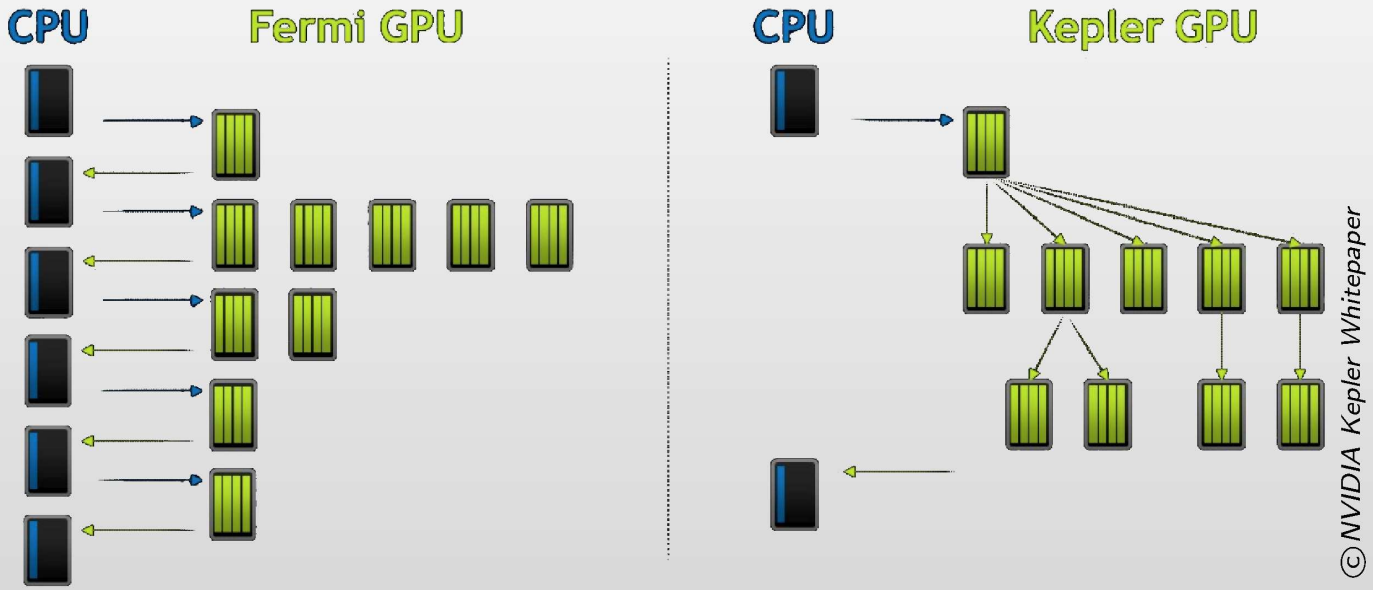
© NVIDIA Kepler Whitepaper



© NVIDIA Kepler Whitepaper



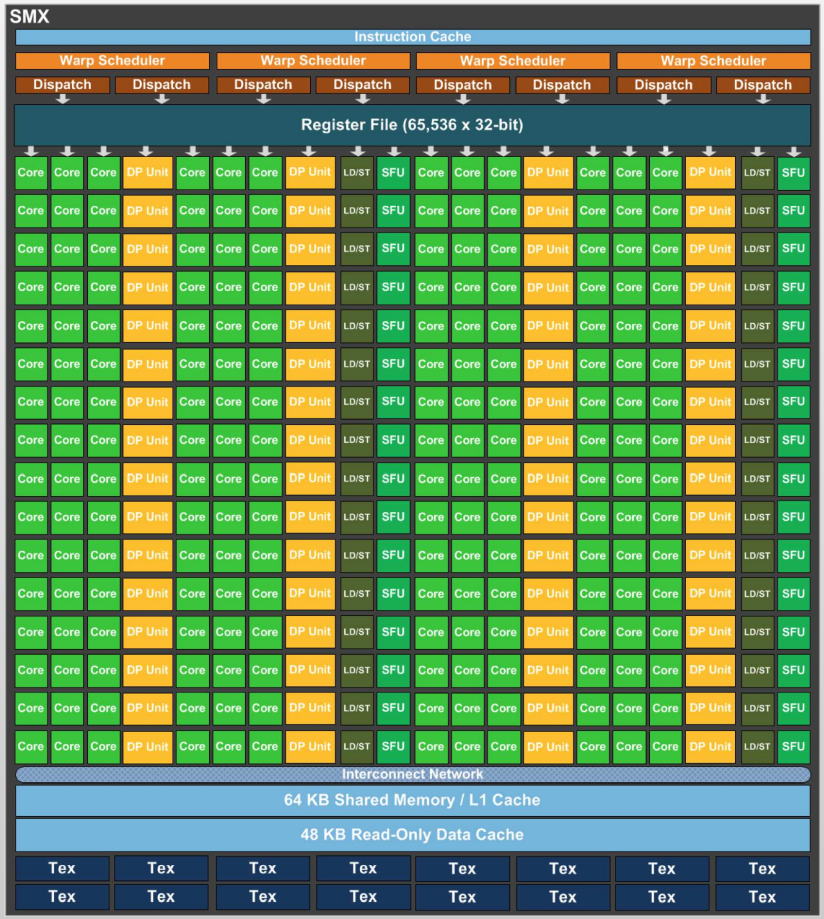
- 32 concurrent kernels
- 32 independent hardware queues
- Only for CC $\geq$ 3.5



- Recursions supported
- Avoids unnecessary CPU $\leftrightarrow$ GPU communications
- Dynamically adapt grid size as needed

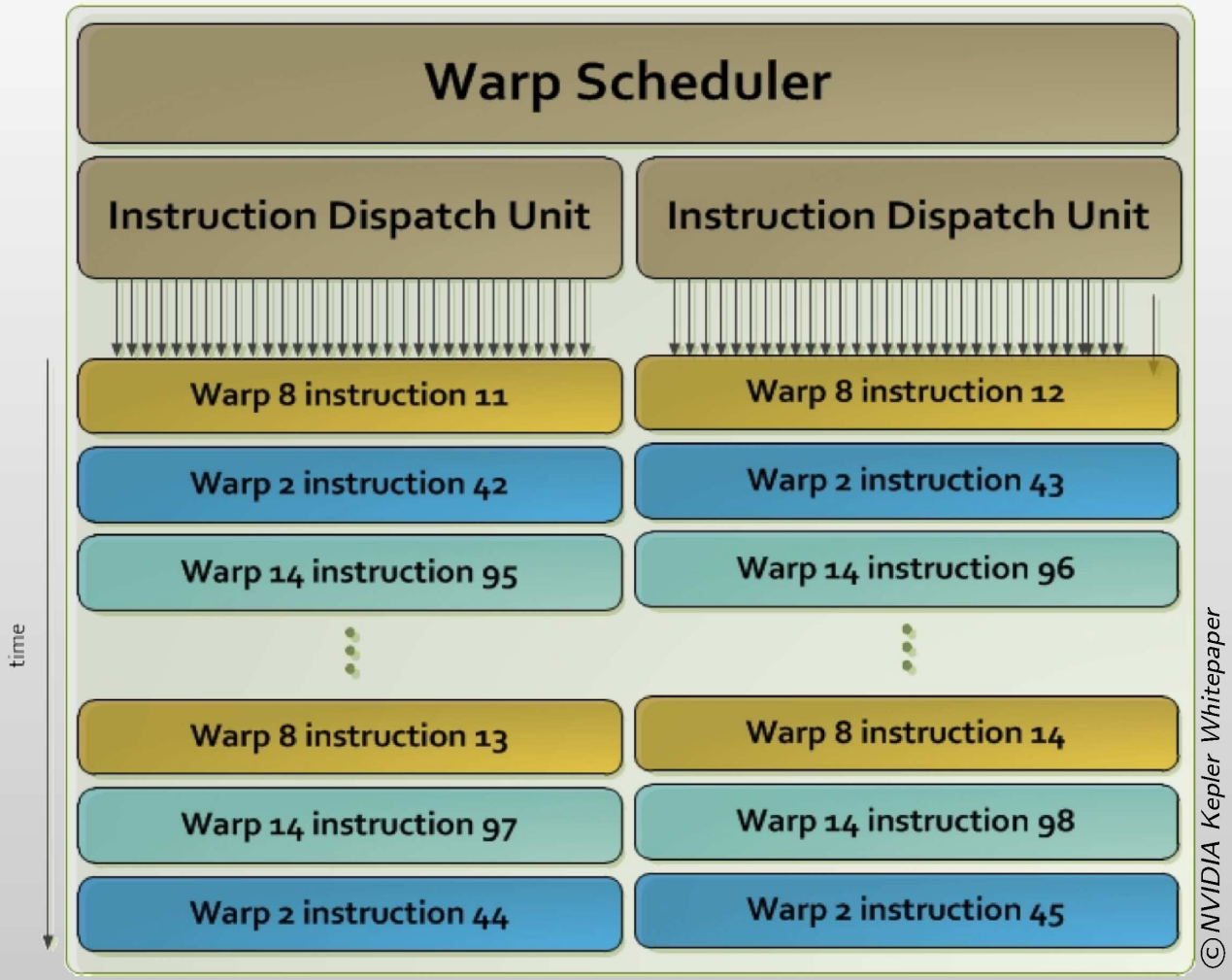
NVIDIA Kepler - SMX

- 48KB/32KB/16 KB *shared memory*
  - Bank size 4byte or 8byte
  - 2× throughput
  - 128bit load instruction
- 65536 4byte registers
  - Max 255 registers per thread
- Up to . . .
  - . . . 16 threadblocks
  - . . . 64 warps
  - . . . 2048 threads



© NVIDIA Kepler Whitepaper

NVIDIA Kepler - SMX



You have to use Instruction-Level Parallelism (ILP)

NVIDIA Kepler - Further Information

- NVIDIA Keper Tuning Guide
- NVIDIAs Next Generation CUDA Compute Architecture: Kepler GK110
- NVIDIA C Programming Guide

Outline

- CUDA implementation of BLAS
- Included in CUDA SDK
- Does not auto-parallelize across multiple GPUs
- Uses column-major layout

Getting started

- CUDA matrixMulCUBLAS sample
- NVIDIA CUBLAS Users Guide

cublasSgemm example

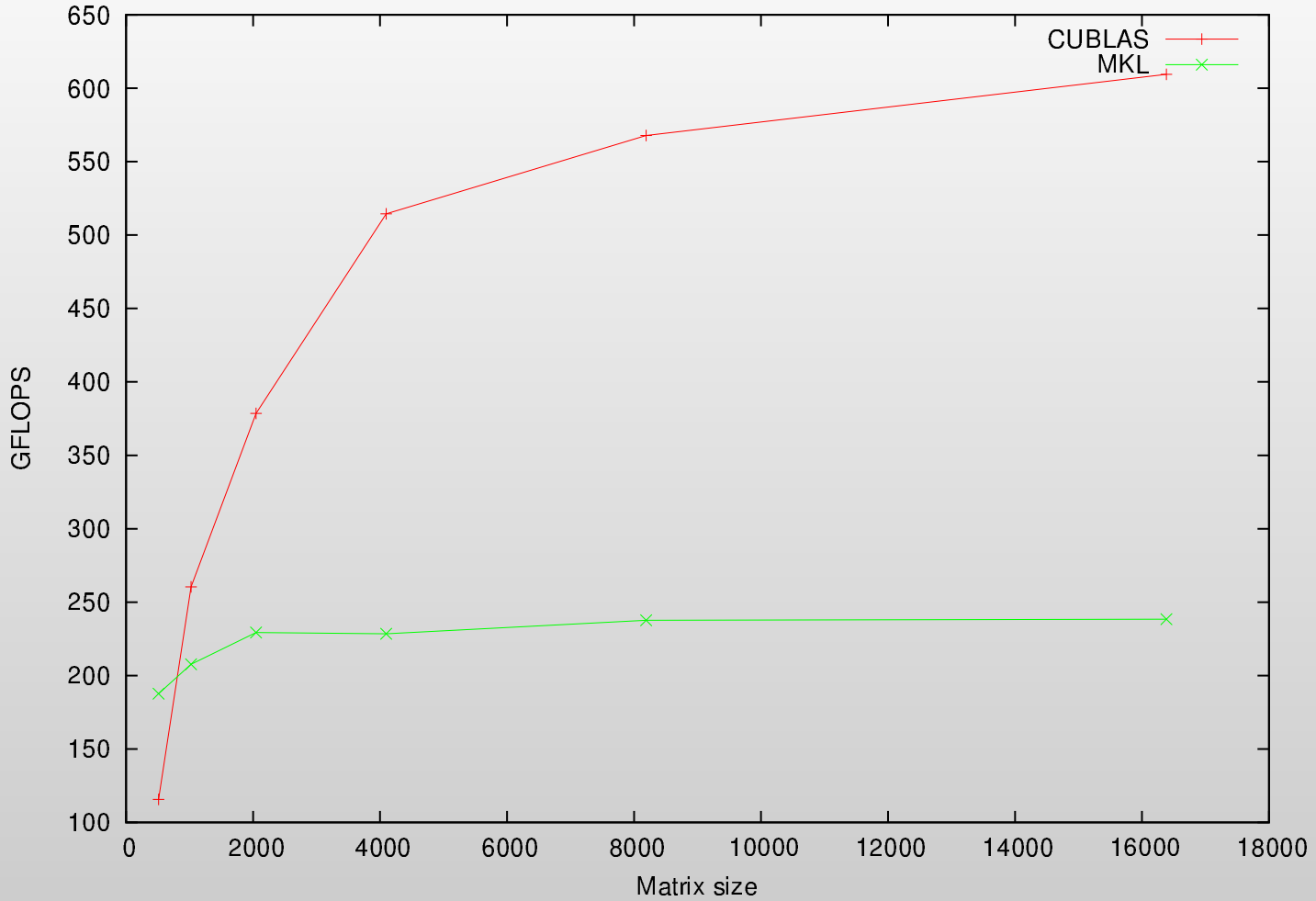


Figure: SGEMM performance (incl. HtD and DtH). MKL @ 12 core Intel X5650. CUBLAS @ NVIDIA Quadro 6000.

- DLA library comparable to LAPACK
- Offers subset of CUBLAS
- Available for ...
  - ...NVIDIA GPUs
  - ...AMD GPUs
  - ...Intel Xeon Phi

Getting started

- `http://icl.cs.utk.edu/magma/index.html`
- MAGMA samples provided in the *testing/* directory

- Requirements
  - CUDA
  - CPU BLAS and LAPACK
- Auto-parallelizes across multiple GPUs and CPUs[1]
- No knowledge of CUDA necessary
- Similar interface as LAPACK

[1] Fengguang Song and Jack Dongarra. "A scalable framework for heterogeneous GPU-based clusters". In: *Proceedinbgs of the 24th ACM symposium on Parallelism in algorithms and architectures*. ACM. 2012, pp. 91–100.

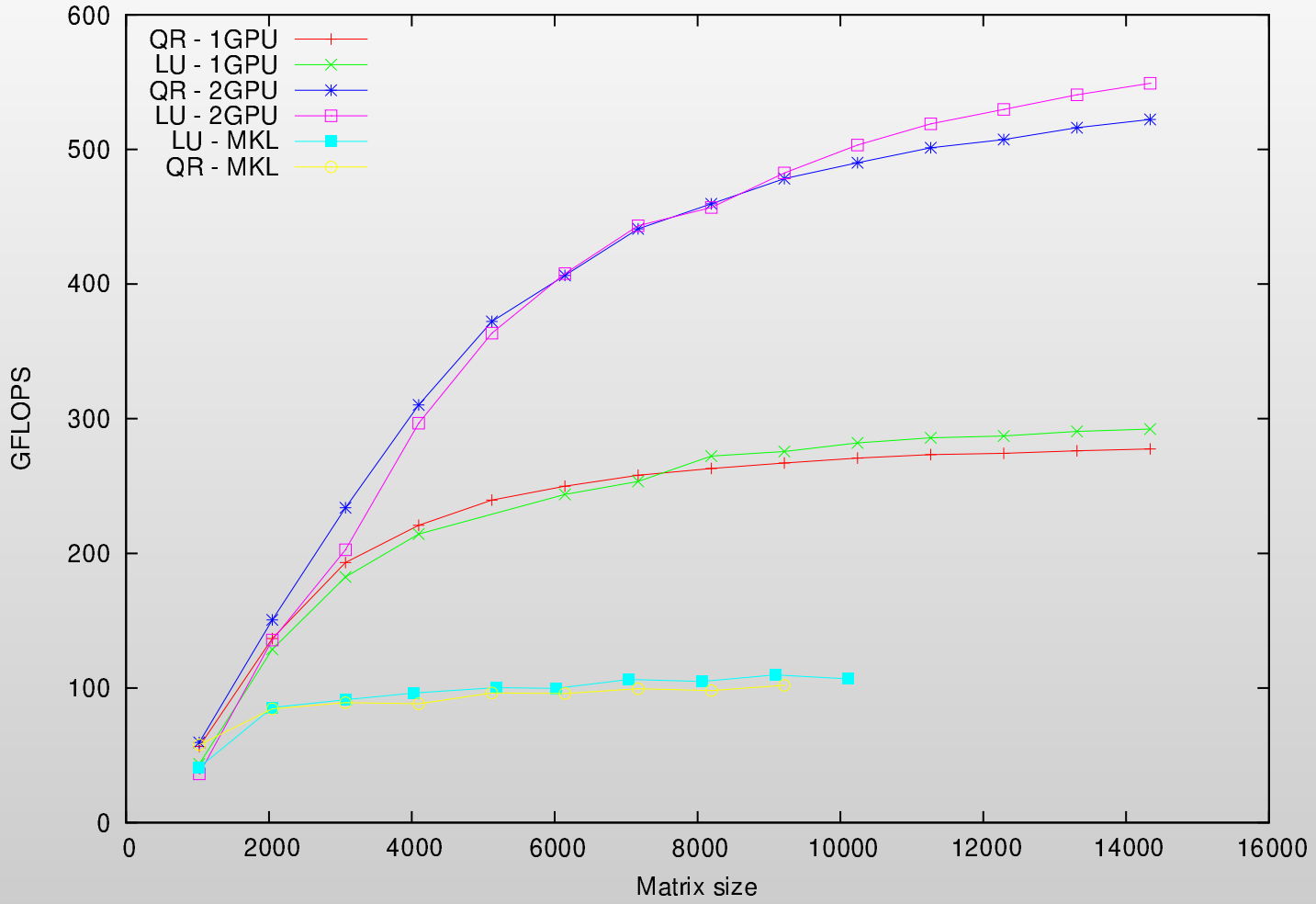
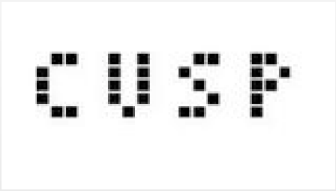
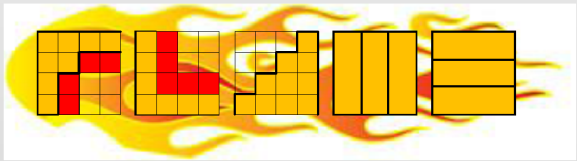


Figure: MAGMA @ 1-2 NVIDIA Quadro 6000 vs. MKL @ 12 core Intel X5650 using DP.





CUPARSE