High-performance computing hardware for high data rates

Agenda

Overview
- Performance
- Programming Models
- Tools and Libraries

OpenCL Programming Model

GPU Hardware

Optimization Strategies

Tutorial

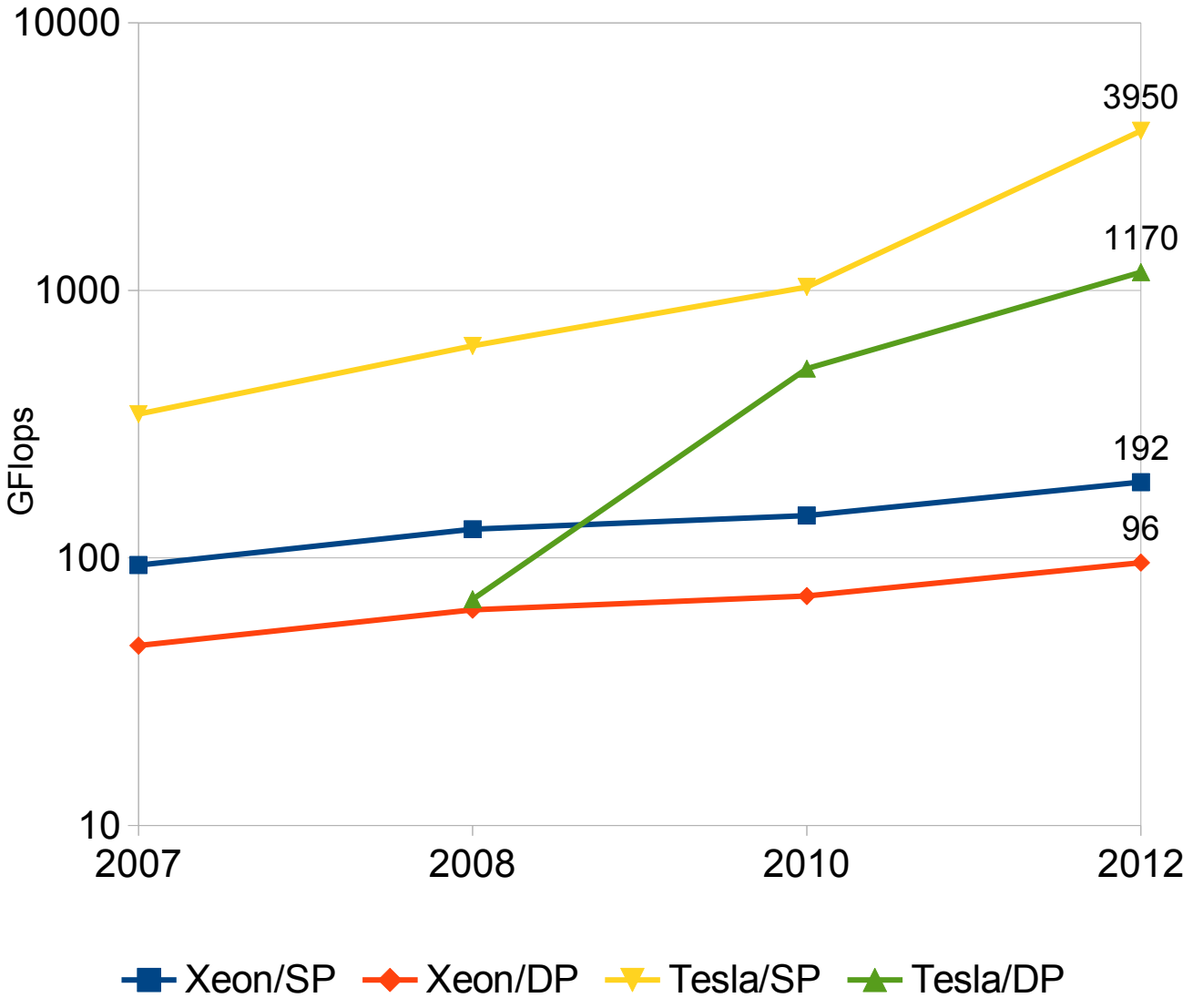
Why care about parallel programming?

Xeon E5 Series
Up to 12 cores
256 bit SIMD instructions

NVIDIA Titan
14 SMX cores
192 fp blocks each

Xeon Phi
~ 60 cores
512 bit SIMD instructions

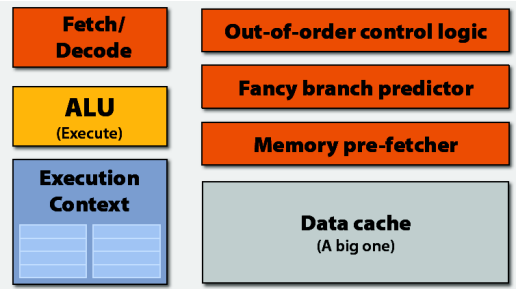
Samsung Galaxy Note
8 ARM cores

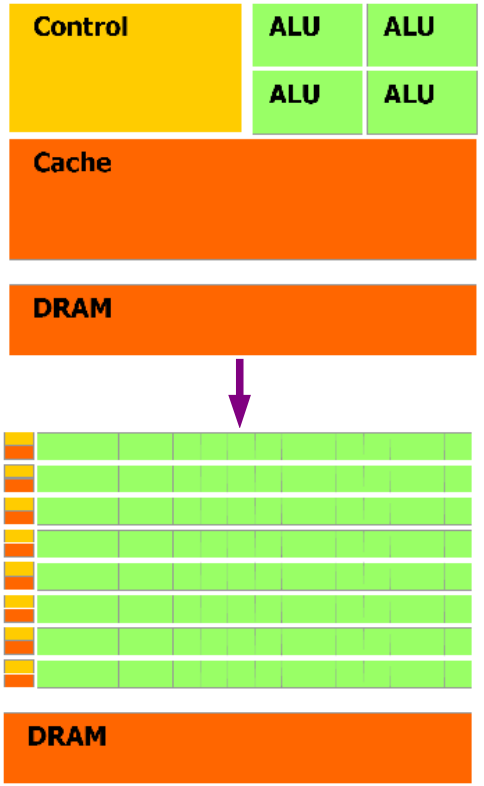


Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

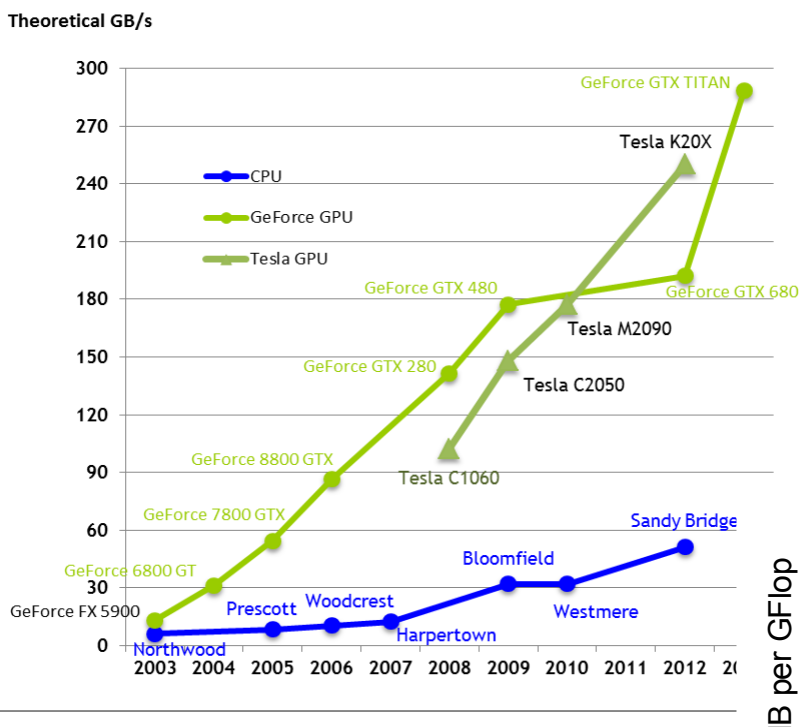
General-purpose processors vs. GPUs

- More computations, less logic
- Highly parallel architectures able to process up to several thousand floating point numbers in parallel
- SIMT architecture optimized to execute single command on multiple data items
- Varying hardware architectures requiring different optimization strategies
- Only certain operations are fast while other may execute slower than on general purpose processor
- Significantly worsened computation-to-memory ratio and smaller caches
- Many-TFlop GPUs are mass-market products and available under 500 EUR
- Easy scaling using PCI express bus. Standard desktop boards may handle up to 8 GPU-cores



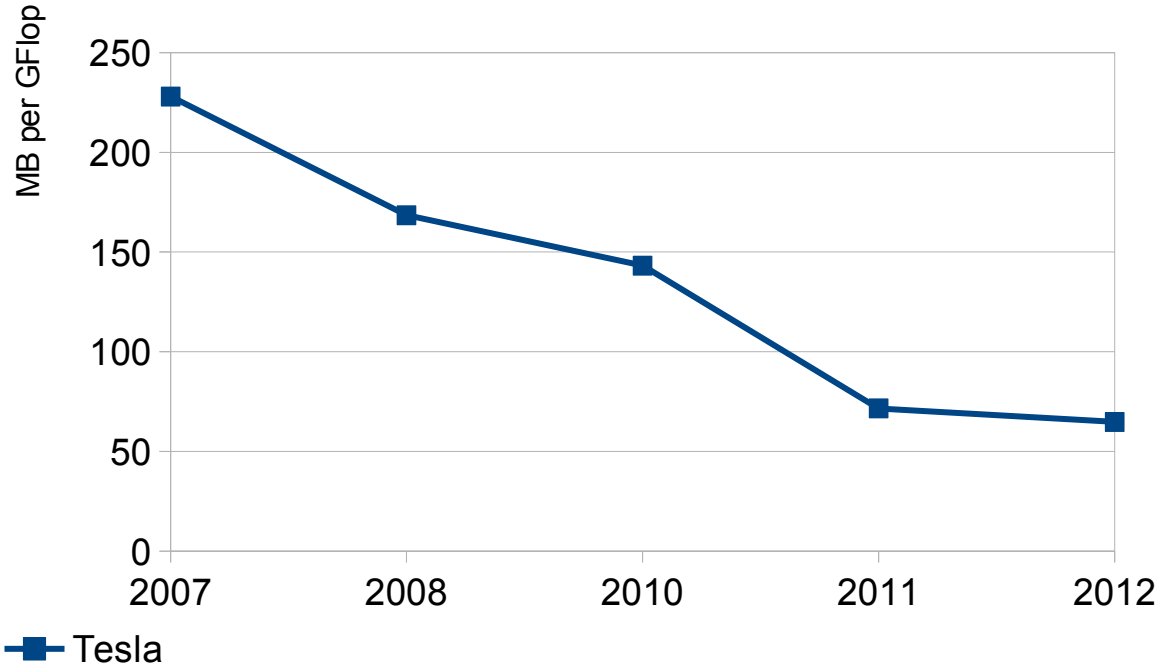


Computations vs. Memory



Memory bandwidth increased 10 times

But Bandwith/Flop ratio worsened 5 times

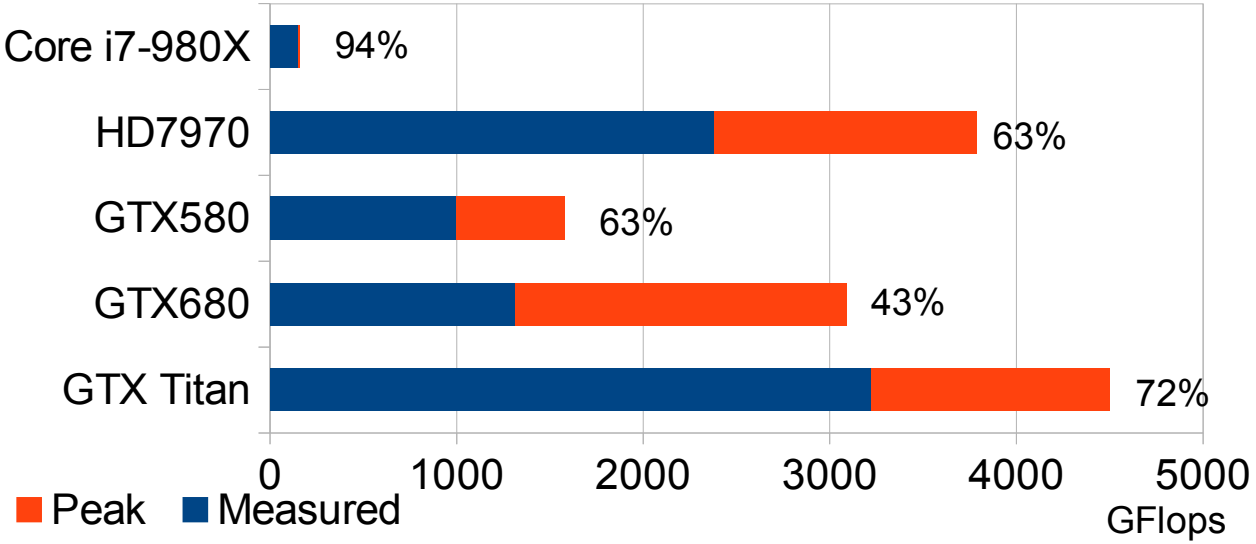


S. Chilingaryan et. all

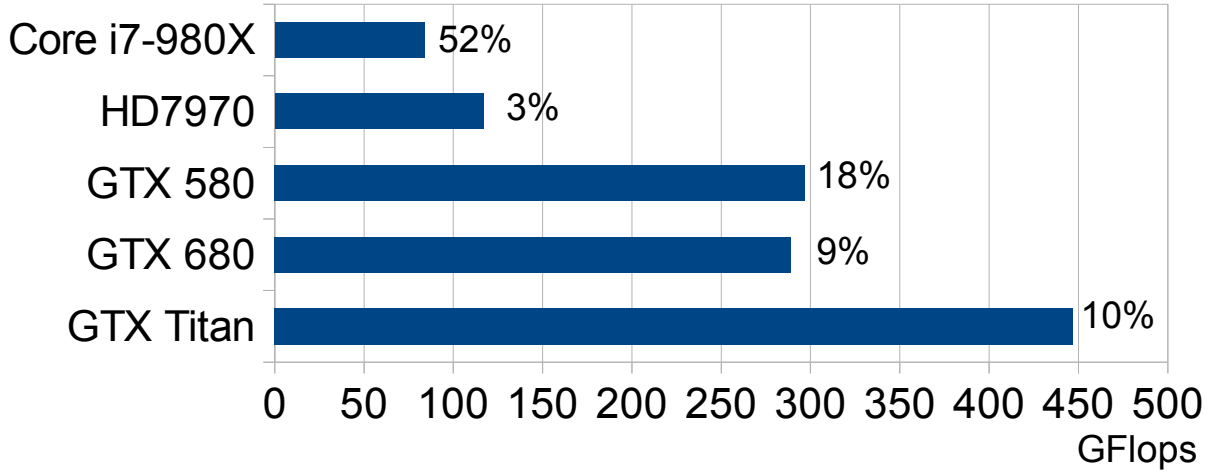
Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Actual Performance

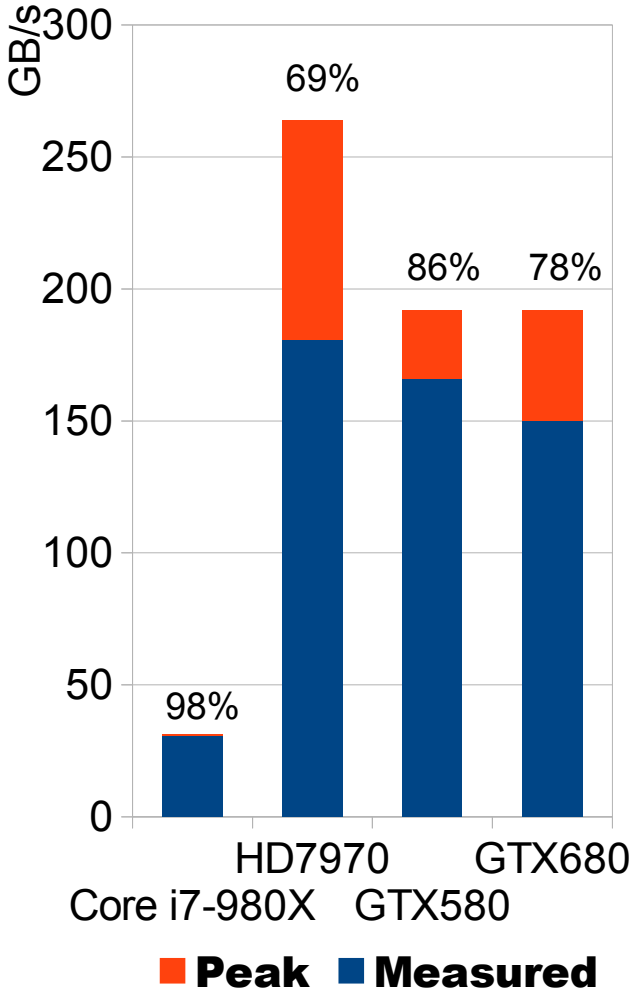
Matrix Multiplication



1D Fast Fourier Transform



Memory Bandwidth



GTX Titan performance is taken from Anandtech

Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Programming Models

OpenCL

Open industry standard supporting multiple hardware platforms by major vendors including CPUs and GPUs. OpenCL is defines C programming interface and allows full control over device operation.

PyOpenCL | JOCL | WebCL

CUDA

NVIDIA CUDA defines a set of extensions over C language and support some advanced features of NVIDIA hardware which are not available in OpenCL.

PyCUDA | JCUDA | MATLAB

OpenAC

Declarative programming model similar to OpenMP. Only commercial solutions are available at the moment.

Vendor

RenderScript (Android), DirectCompute (DirectX), ...

OpenCL Platforms

Karlsruhe Institute of Technology

Processors

X86 by AMD/Intel
IBM Power

Accelerators

IBM CELL
Intel Xeon Phi

GPUs

AMD GPUs & APUs
NVIDIA GPUs
Intel HD & Iris
ARM Mali

Apple

OpenCL 1.2

- x86
- OSX
- NVIDIA
- ATI

Intel

OpenCL 1.2

- x86
- Win
- Phi
- Linux
- Intel HD — Windows only

AMD

OpenCL 1.2

- x86
- Win
- ATI
- Linux

NVIDIA

OpenCL 1.1

- NVIDIA
- Win
- Linux

FPGA

Altera Stratix V

Other

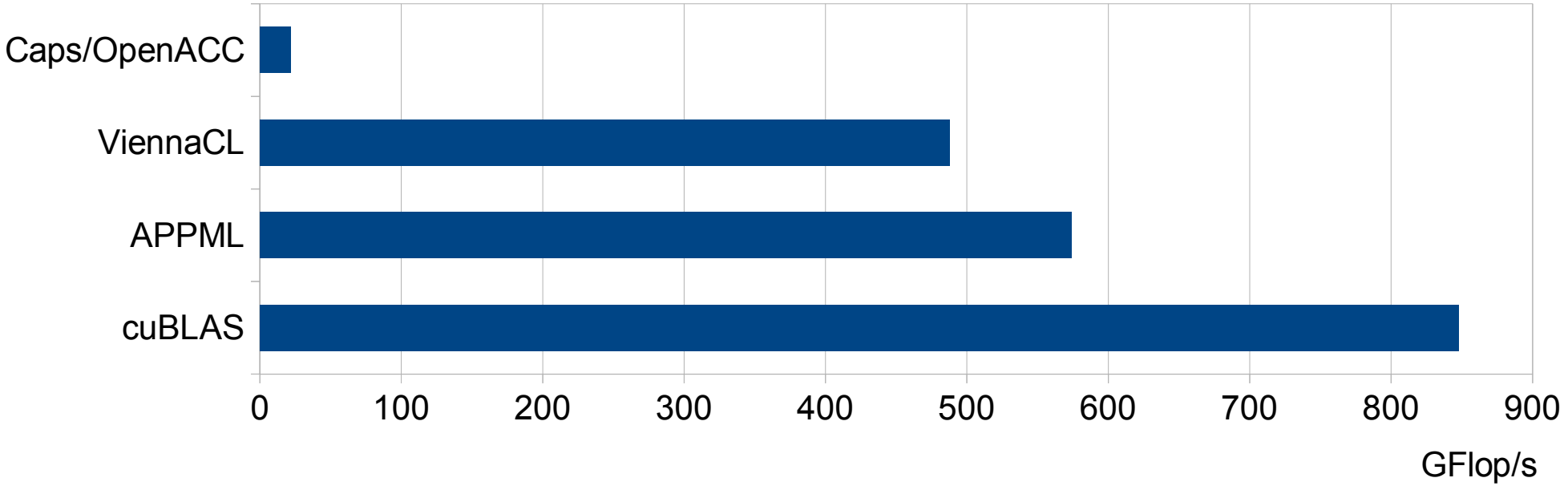
Tilera

Libraries

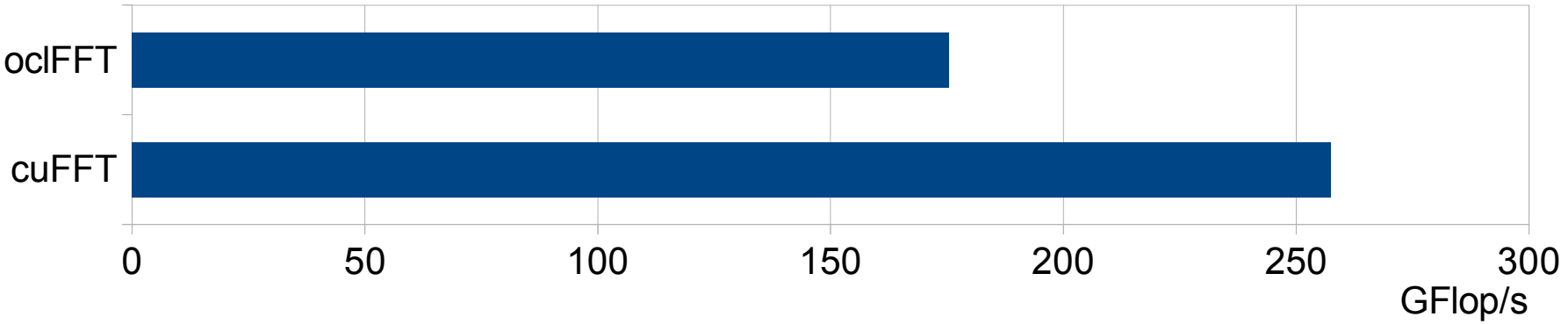
	CUDA	OpenCL
BLAS	cuBLAS	ViennaCL / APPML
Sparse BLAS	cuSPARSE	ViennaCL
Lapack	CULA / Magma	clMagma
Sparse Solvers	cuSP	ViennaCL
FFT	cuFFT	oclFFT / APPML
RNG	cuRAND	MTGP, Random123
STL	NVIDIA Thrust	AMD Bolt

Performance

Multiplication of real single-precision matrices (1024x1024) on GTX590



1D FFT of a real single-precision vector on GTX680

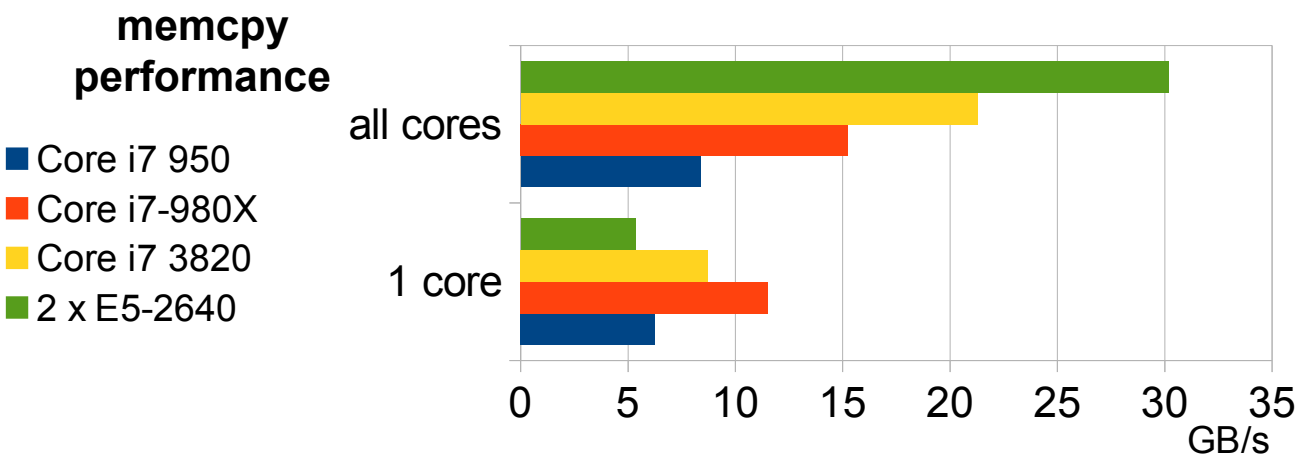


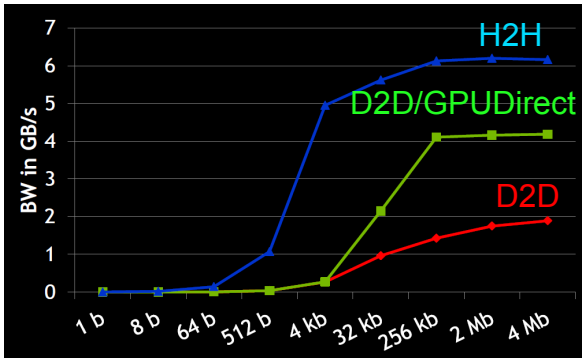
Institute for Data Processing and Electronics
Karlsruhe Institute of Technology

Development Tools

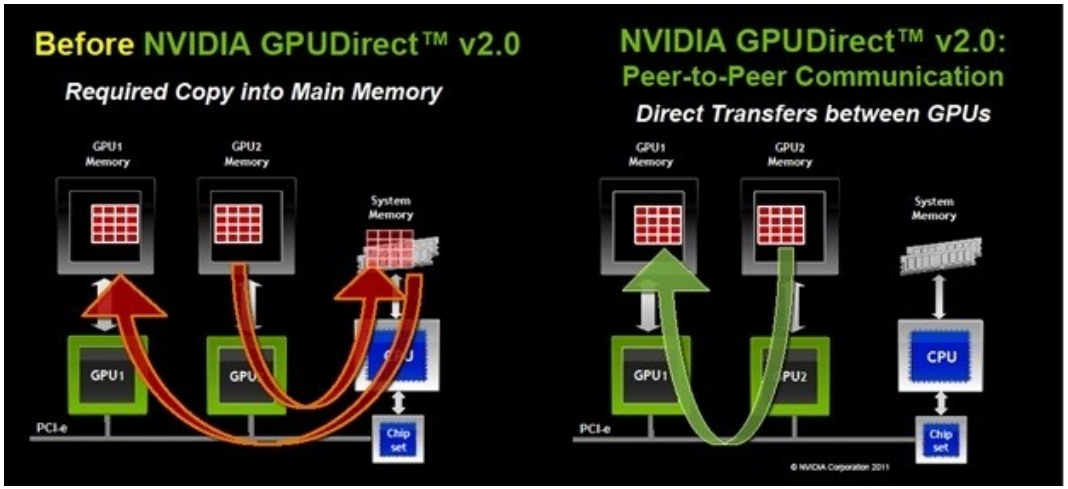
	Platforms	CUDA	Mode	Includes
NVIDIA ComputeProf	NVIDIA	CUDA 4.0 / OpenCL	GUI / Console	Debugger / Profiler
NVIDIA Nsight	NVIDIA	CUDA	GUI / Console	IDE / Debugger / Profiler
AMD CodeXL	AMD	OpenCL	GUI	Debugger / Profiler
gDebugger	NVIDIA, AMD	OpenCL	GUI	Debugger

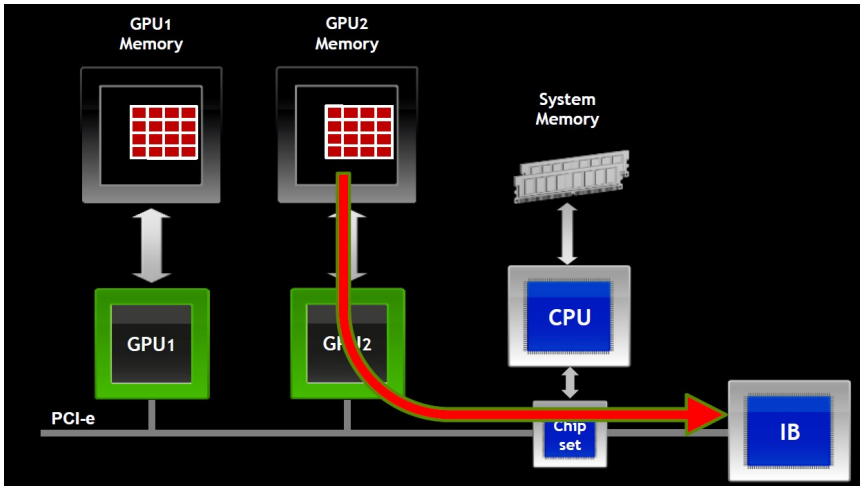
NVIDIA GPUDirect





MVAPICH2 1.9b throughput





Direct communication between GPUs, Network, and other devices on PCI express bus

S. Chilingaryan et. all

OpenCL vs. CUDA

- ✔ Advanced features of NVIDIA Cards, specially GPUDirect
- ✔ Faster introduction of new features
- ✔ Better Libraries
- ✔ Shorter support code

- ✔ Open standard supported by many vendors
- ✔ Hybrid computing using both GPUs and CPUs
- ✔ Better synchronization across multiple devices
- ✔ Easier run-time compilation support
- ✔ Full C99 compatibility
- ✔ WebCL – ready for Web