Master Educational Program
“Information technology in applications”

Mathematical computations with GPUs

Introduction

Alexey A. Romanenko
arom@ccfit.nsu.ru
Novosibirsk State University

How to..

* Process terabytes of seismic data on desktop PC.
* Estimate tsunami impact faster then wave reaches dry earth.
* Rasterize 3D scenes in real-time.

Topics of the course

* architecture of a GPU and its difference from general-purpose CPU;
* general notions related to CUDA;
* contents and functionality of GPU-optimized libraries;
* area of applicability of GPUs for solving scientific and applied problems;
* general principles of optimizing programs for GPUs;
* methods of debugging and profiling programs running on GPUs;
* tools and instruments for developing, debugging and profiling programs running on GPUs.

Computer power

«Titan» **2012**
299 008 Opteron Cores
18 688 k20 NVIDIA GPU
20+ Pflops

«K computer» **2011**
68 544 x 8-core SPARC64 VIIIfx.
8,162 Pflops

280 Tflops
212,992 CPUs

Performance

Time

Growth of performance

* Growth of clock rate
* Number of cores/CPUs
* CPU architecture complexity
  * Number of registers
  * Pipe-line
  * Digital capacity (4bit CPU, 16, 32, 64)
  * etc

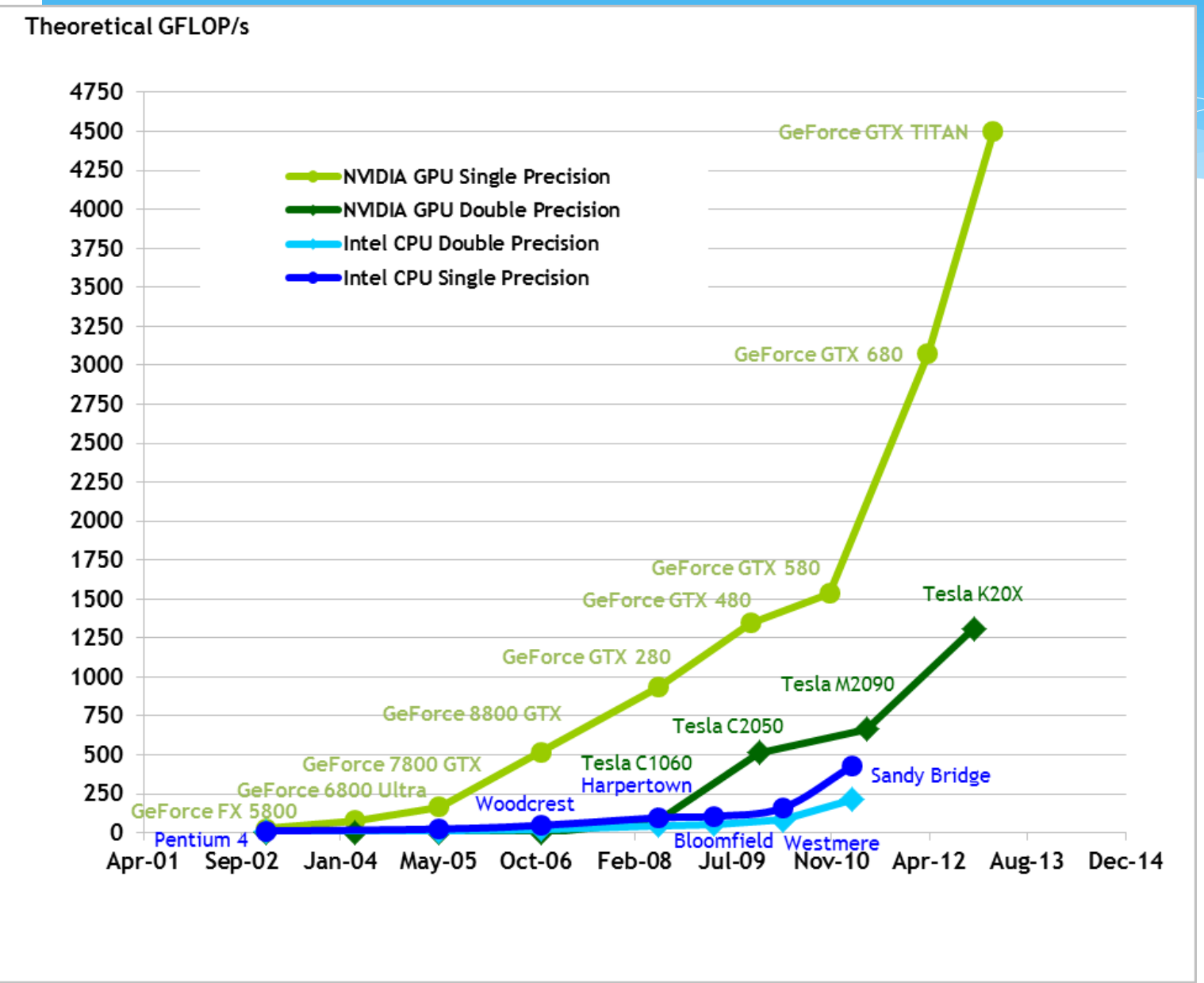
Computer graphics

Computer graphics

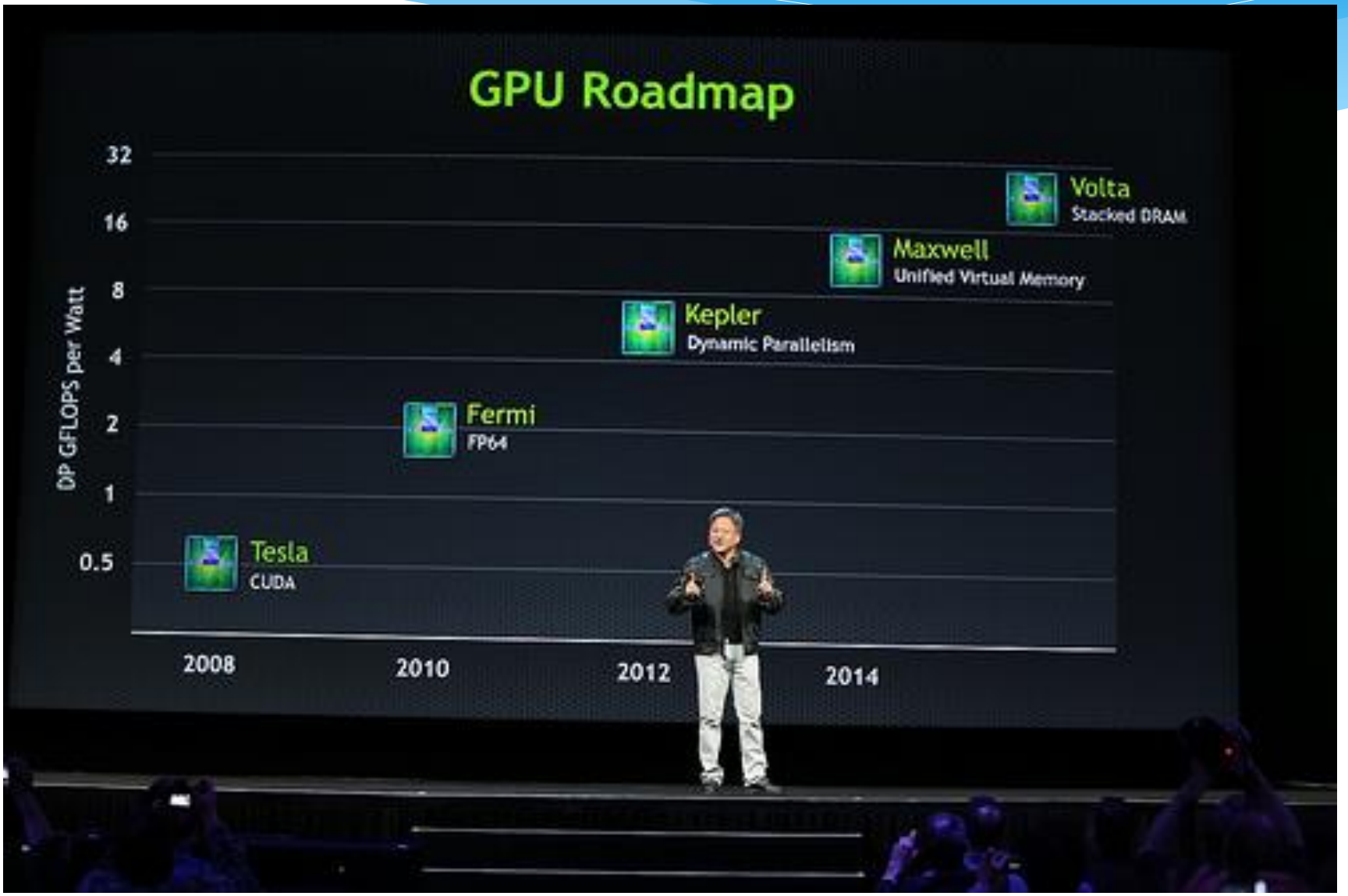
* Small vectors
* Small matrixes
* filters/post-processing
* Calculation of projections
* etc.

* Huge mumber of identical small operations.

Performance of video cards

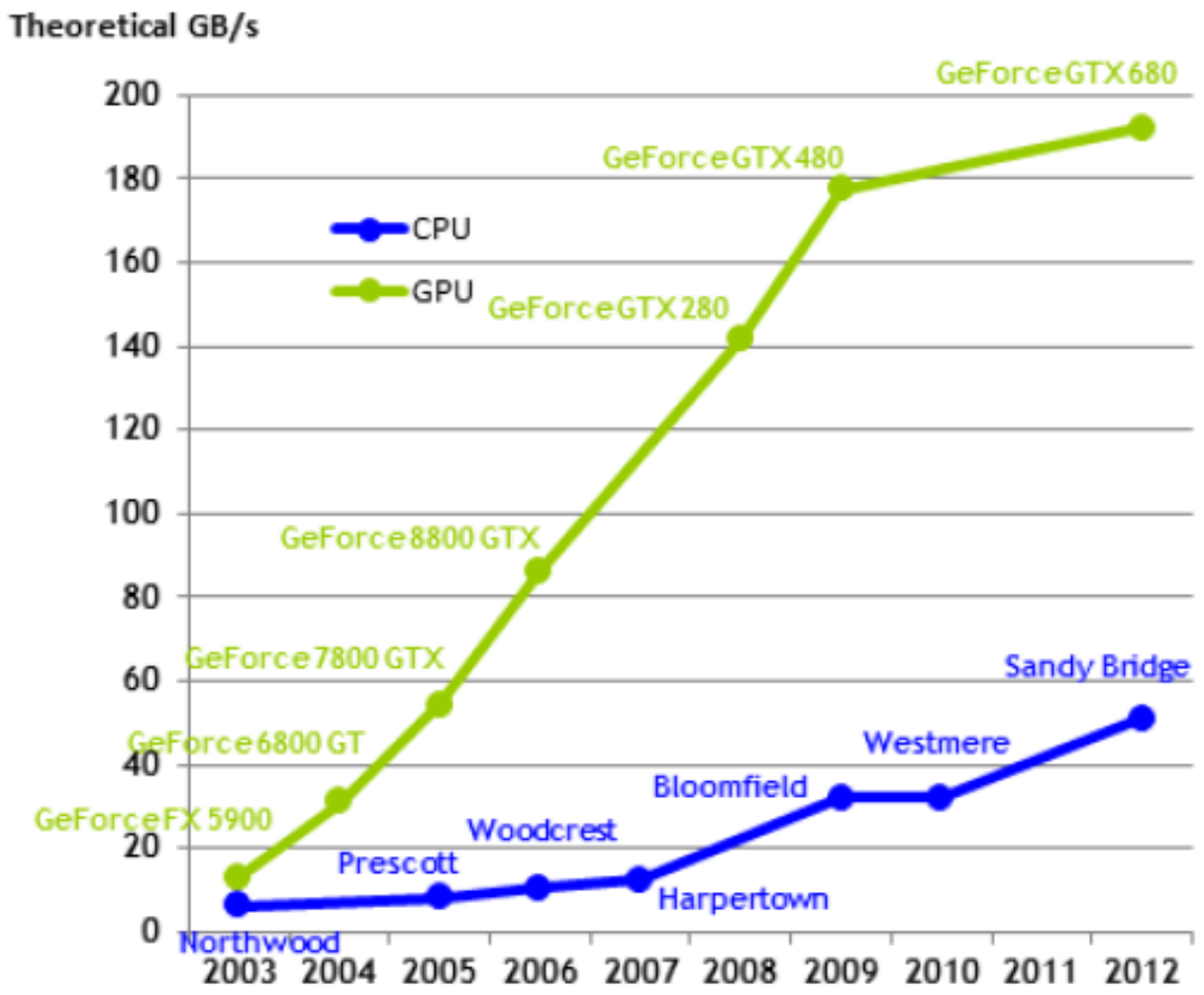


March 2013



* Kepler: 28nm, 1.4 Tflops DP
* Maxwell: 22nm, 4 Tflops DP

Memory bandwidth



Heterogeneous systems

+

3 of 5 leaders in top500 list!

Heterogeneous vs. Homogeneous computing systems (06.2012)

* **Tianhe-1A** (2 place)
  * Xeon X5670 6C 2.93 GHz, NVIDIA 2050

  * 4,7 Pflops (peak)
  * 4040 kW

* **Jaguar** (3 place)
  * Opteron 6-core 2.6 GHz

  * 2,3 Pflops (peak)
  * 6950 kW

Toward exaflop

Direct scaling…

* **Tianhe-1A**
* 4,7 Pflops (peak) * 213 = 1001,1 Pflops
* 4040 kW * 213 = 860 MW

* **Jaguar**
* 2,3 Pflops (peak) * 435 = 1000,5 Pflops
* 6950 kW * 435 = 3023 MW

exa	10^{18}
peta	10^{15}
tera	10^{12}
giga	10^{9}
mega	10^{6}
kilo	10^{3}

Sayano–Shushenskaya Dam

Turbines	10 × 640 MW (initial) 6 × 640 MW (current)
Installed capacity	5,120 MW (current)
Maximum capacity	6,400 MW
Annual generation	23.5 TW

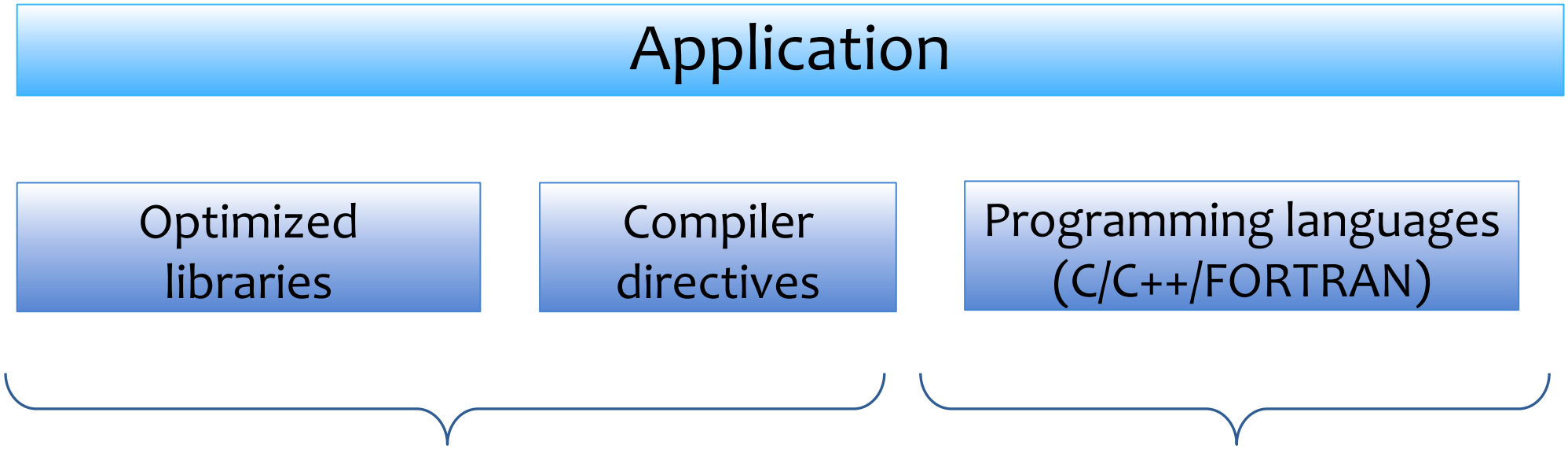
GFLOPs per Watt

* Green500 (November 2013)
  * www.green500.org/lists
  * List of most power effective supercomputer
  * Ten of top have NVIDIA GPUs. Two – 6 month ago.
* Performance / Watt
  * 4,5 – now
  * 0,3 – November, 2007
* Reaching exascale means boosting speeds by 50-100 times, while keeping power relatively static.

Application speedup

Example Applications	URL	Application Speedup
Seismic Database	http://www.headwave.com	66x to 100x
Mobile Phone Antenna Simulation	http://www.acceleware.com	45x
Molecular Dynamics	http://www.ks.uiuc.edu/Research/vmd	21x to 100x
Neuron Simulation	http://www.evolvedmachines.com	100x
MRI processing	http://bic-test.beckman.uiuc.edu	245x to 415x
Atmospheric Cloud Simulation	http://www.cs.clemson.edu/~jesteel/clouds.html	50x

Approaches to GPU programming



CUDA Roadmap

2014 – 7 years!

* CUDA 1.0 – 2007
* CUDA 2.0 – 2008
* CUDA 3.0 – 2009
* CUDA 4.0 – 2011
* CUDA 5.0 – 2012
* CUDA 5.5 – 2013
* CUDA 6.0 – 2014

CUDA vs. OpenCL

* CUDA
  * NVIDIA GPU (Cray, HP, IBM, T-Platforms, NextIO…)
  * Close to peak performance
  * Functionality
  * Comfort for developers (debugger, performance analyzer, etc.)
  * Support
  * Teaching materials, libraries
* OpenCL
  * Architecture is not fixed, universality
  * Performance is not high priority

CUDA vs. OpenCL performance

* CUDA applications have up to 30% higher performance than OpenCL application.
  * http://arxiv.org/ftp/arxiv/papers/1005/1005.2581.pdf
  * http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6047190

Useful links

* http://www.nvidia.com/object/cuda_home.html
* http://www.gputechconf.com/page/home.html
* http://docs.nvidia.com/

Questions