VUB VRIJE UNIVERSITEIT BRUSSEL ULB UNIVERSITÉ LIBRE DE BRUXELLES

GPU Computing

Lesson 1: The Power of GPUs

Gauthier Lafruit & Jan Lemeire

2022-2023

http://parallel.vub.ac.be/education/gpu

Course organisation

- 13/2: lessons 1 & 2
- 14/2: lessons 3 & 4
- 20/2: labo mini-project
- 27/2: labo mini-project
- 28/2: lessons 5 & 6
- 6/3: labo exercises blocks & shared memory
- 7/3: project subject: stereo matching
- 13/3: labo
- 14/3: lessons 7 (warps & coalescence)
- 20/3: labo coalescence
- 27/3: labo coalescence
- ...: project

Course evaluation

- Mini-project: 30%
- Project: 70%
  - Theory: 20%
  - Practice: 50%


ELITEGROUP
GeForce
8800 GTX
NVIDIA

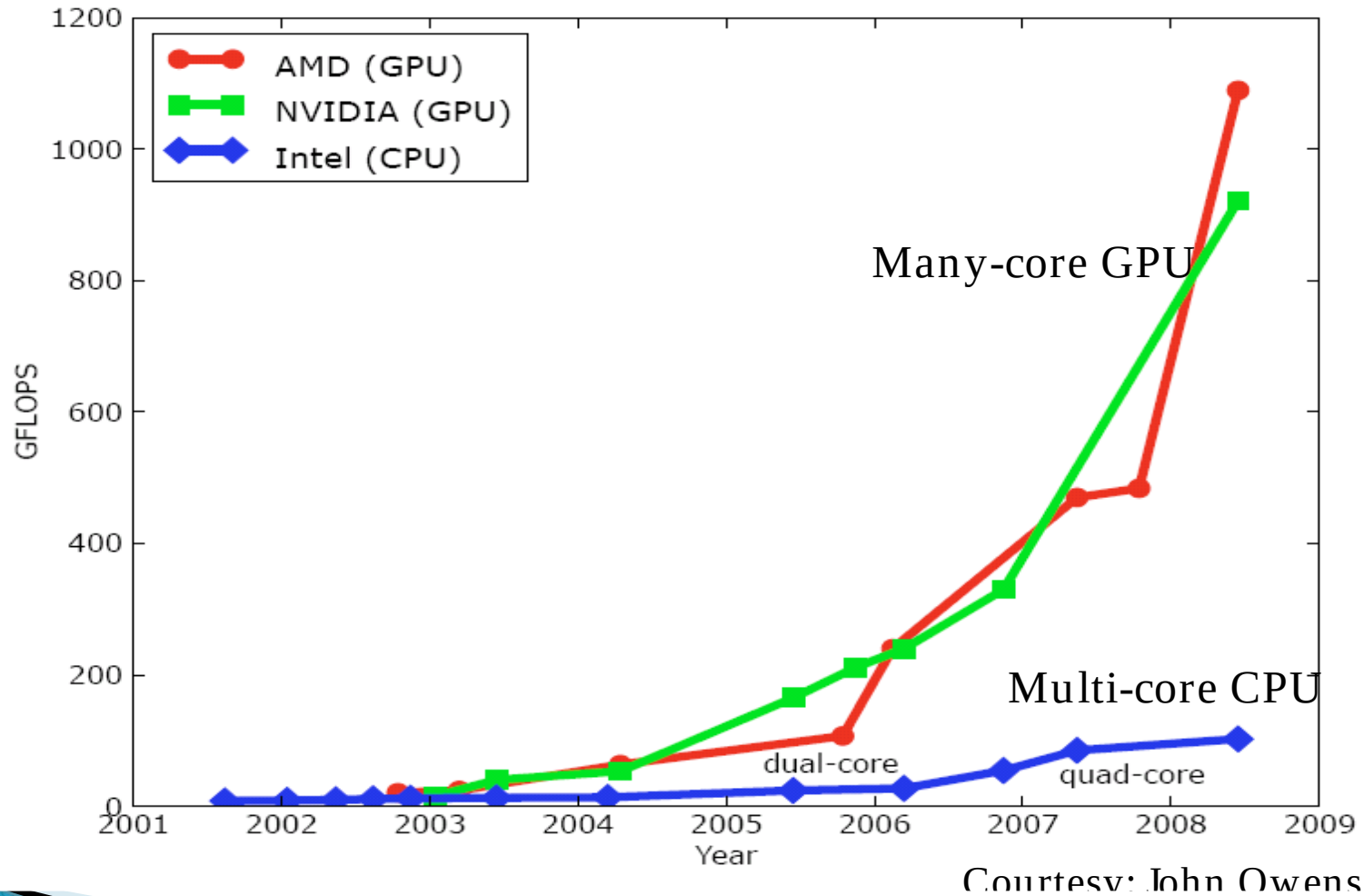
versus


intel
CORE i3
inside
intel


2010
350 Million triangles/second
3 Billion transistors GPU
1995
5.000 triangles/second
800.000 transistors GPU
2016
14.000 Million triangles/second
15 Billion transistors GPU
HELMET
BULLY DOG
OGIO

VRIJE UNIVERSITEIT BRUSSEL

Graphical Processing Units (GPUs)



Courtesy: John Owens

Supercomputing for free

- FASTRA at university of Antwerp (2009)

http://fastra.ua.ac.be

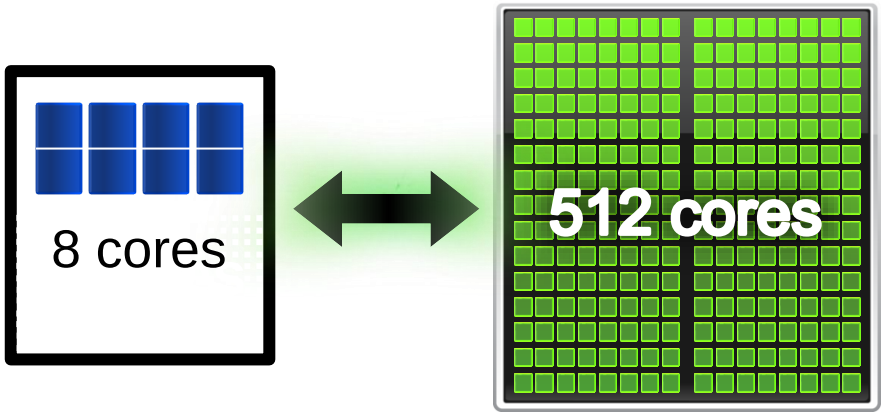
Collection of 8 graphical cards in PC

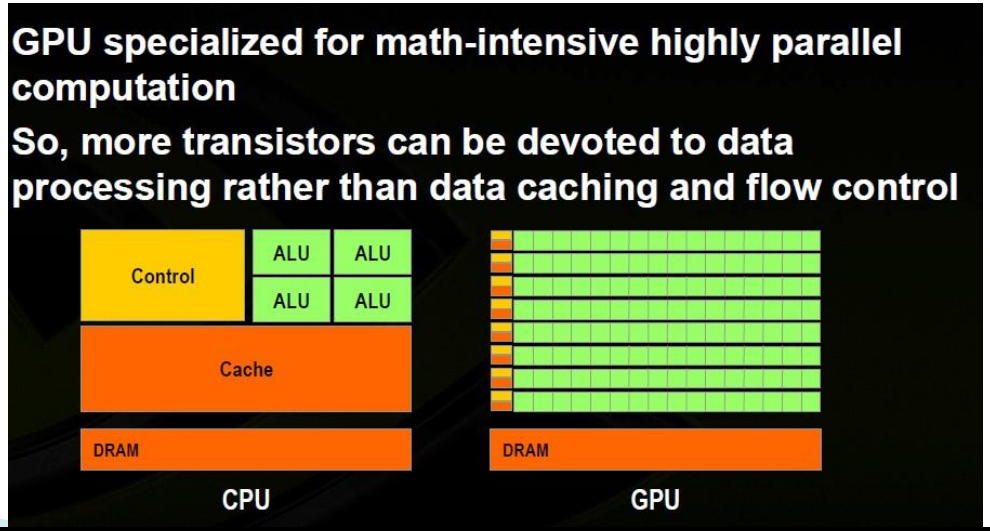
FASTRA 8 cards = 8x128 processors = 4000 euro

Similar performance as University's supercomputer (512 regular desktop PCs) that costed 3.5 million euro in 2005

"**Supercomputing in a box**": a high-end GPU cost 500 to 2500 euro and has equivalent power as 40 quadcore CPUs

Why are GPUs faster?





Both, about 12 billion transistors

No branch prediction, out-of-order execution, …

Devote transistors to… computation

VRIJE UNIVERSITEIT BRUSSEL

GPU vs CPU: NVIDIA 280 vs Intel i7 860

	GPU	CPU[1]
Registers	16,384 (32-bit) / multi-processor[3]	128 reservation stations
Peak memory bandwidth	141.7 Gb/sec	21 Gb/sec
Peak GFLOPs	562 (float)/ 77 (double)	50 (double)
Cores	240 (scalar processors)	4/8 (hyperthreaded)
Processor Clock (MHz)	1296	2800
Memory	1Gb	16Gb
Local/shared memory	16Kb/TPC[2]	N/A
Virtual memory	None	

[1]http://ark.intel.com/Product.aspx?id=41316
[2]TPC = Thread Processing Cluster (24 cores)
[3]30 multi-processors in a 280

Basic Architecture (more details in chapter 3)

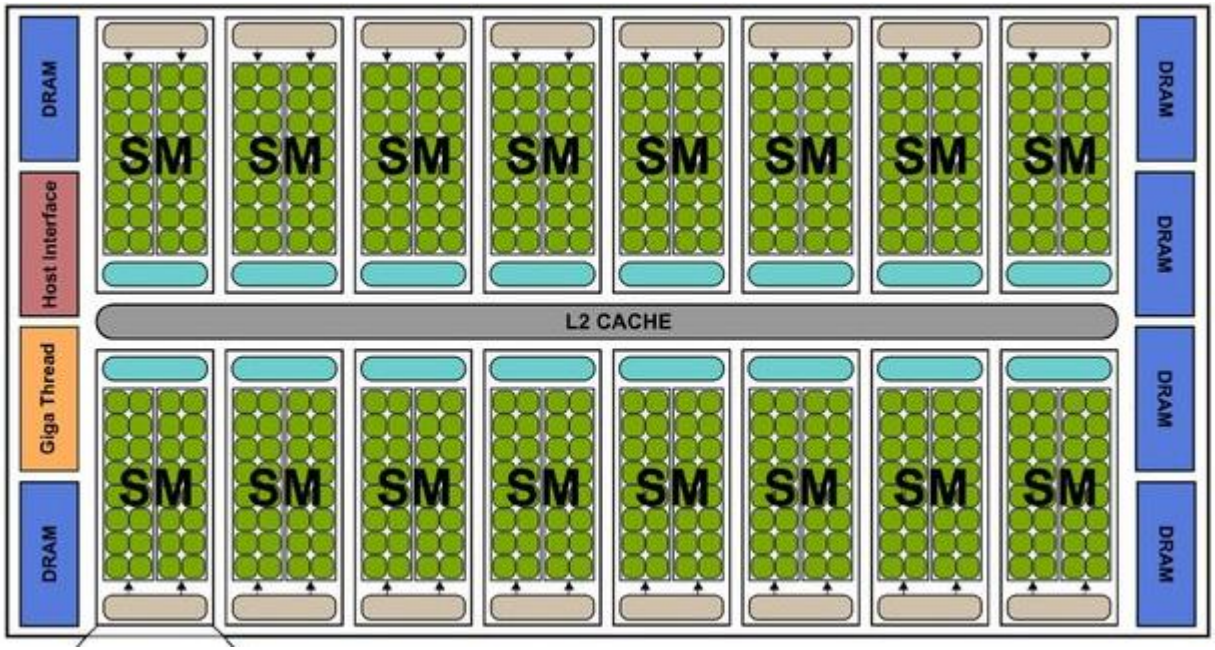
VRIJE UNIVERSITEIT BRUSSEL

Processing elements

- The processing elements of GPUs are called by Nvidia:
  - Scalar Processors (SPs): single-precision floating point operations
  - CUDA cores (in the data sheets)
- The SPs are grouped into Streaming Multiprocessors (MPs or SMs):
  - Is what we would call a *core* (see later)
- The number of SPs per MP is fixed for each GPU generation
- The number of MPs varies per GPU (determines power & price)

VRIJE UNIVERSITEIT BRUSSEL

A GPU consists of Streaming Multiprocessors (SMs/MPs)



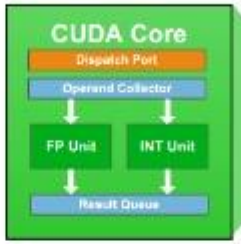
VRIJE UNIVERSITEIT BRUSSEL

Each Multiprocessor consists of Scalar Processors (called CUDA core by Nvidia)

CS 354

Streaming Multiprocessor (SM)

- Multi-processor execution unit
  - 32 scalar processor cores
  - Warp is a unit of thread execution of up to 32 threads
- Two workloads
  - Graphics
    - Vertex shader
    - Tessellation
    - Geometry shader
    - Fragment shader
  - Compute



45

VRIJE UNIVERSITEIT BRUSSEL

Nvidia GPU generations

	Tesla		Fermi		Kepler				Maxwell				Pascal	
Architecture	**G80**	**GT200**	**GF100**	**GF104**	**GK104 (K10)**	**GK110 (K20X)**	**GK110 (K40)**	**GK210 (K80)**	**GM107 (GTX750)**	**GM204 (GTX980)**	**GM200 (Titan X)**	**GM200 (Tesla M40)**	**GP104 (GeForce GTX 1080)**	**GP100 (Tesla P100)**
Time frame	**2006 /07**	**2008 /09**	**2010**	**2011**	**2012**	**2013**	**2013 /14**	**2014**	**2014 /15**	**2014 /15**	**2016**	**2016**	**2016**	**2017**
CUDA Compute Capability	**1.0**	**1.3**	**2.0**	**2.1**	**3.0**	**3.5**	**3.5**	**3.7**	**5.0**	**5.2**	**5.3**	**5.3**	**6.0**	**6.0**
N (multiprocs.)	16	30	16	7	8	14	15	30	5	16	24	24	40	56
M (cores/multip.)	8	8	32	48	192	192	192	192	128	128	128	128	64	64
Number of cores	128	240	512	336	1536	2688	2880	5760	640	2048	3072	3072	2560	3584

VRIJE UNIVERSITEIT BRUSSEL

Nvidia GPU generations

Nvidia Architecture	Clock freq MHz	SPs per MP	SFUs per MP	DPs per MP	RAM band-width (GBs)	latency Λ_{SP} (cycles)
Tesla		8	?	–	141	24
Fermi	1147	32	8	–	144	18
Kepler	1032	192	32	64	86	9
Maxwell	1058	128	32			6
Pascal	1506	128	32	64	192	6
Turing		64	8	?		

PE: single-precision floating-point or integer
SFU: special function unit (cos, sin, …)
DP: double-precision unit (not present in old GPUs)

VRIJE UNIVERSITEIT BRUSSEL

GPU Peak Performance

- GPUs consist of or Streaming MultiProcessors (MPs) grouping a number of Scalar Processors (SPs)
- Nvidia GTX 280 (Tesla architecture):
  - 30MPs x 8 SPs/MP x 2FLOPs/instr/SP x 1 instr/clock x 1.3 GHz (clocks per second)

  = **624 GFlops**
- Nvidia Tesla C2050 (Fermi architecture):
  - 14 MPs x 32 SPs/MP x 2FLOPs/instr/SP x 1 instr/clock x 1.15 GHz = **1030 Gflops**
- Nvidia GTX 1080 (Pascal architecture):
  - 40 MPs x 64 SPs/MP x 2FLOPs/instr/SP x 1 instr/clock x 1.733 GHz = **8872 Gflops**

Memory bandwidth

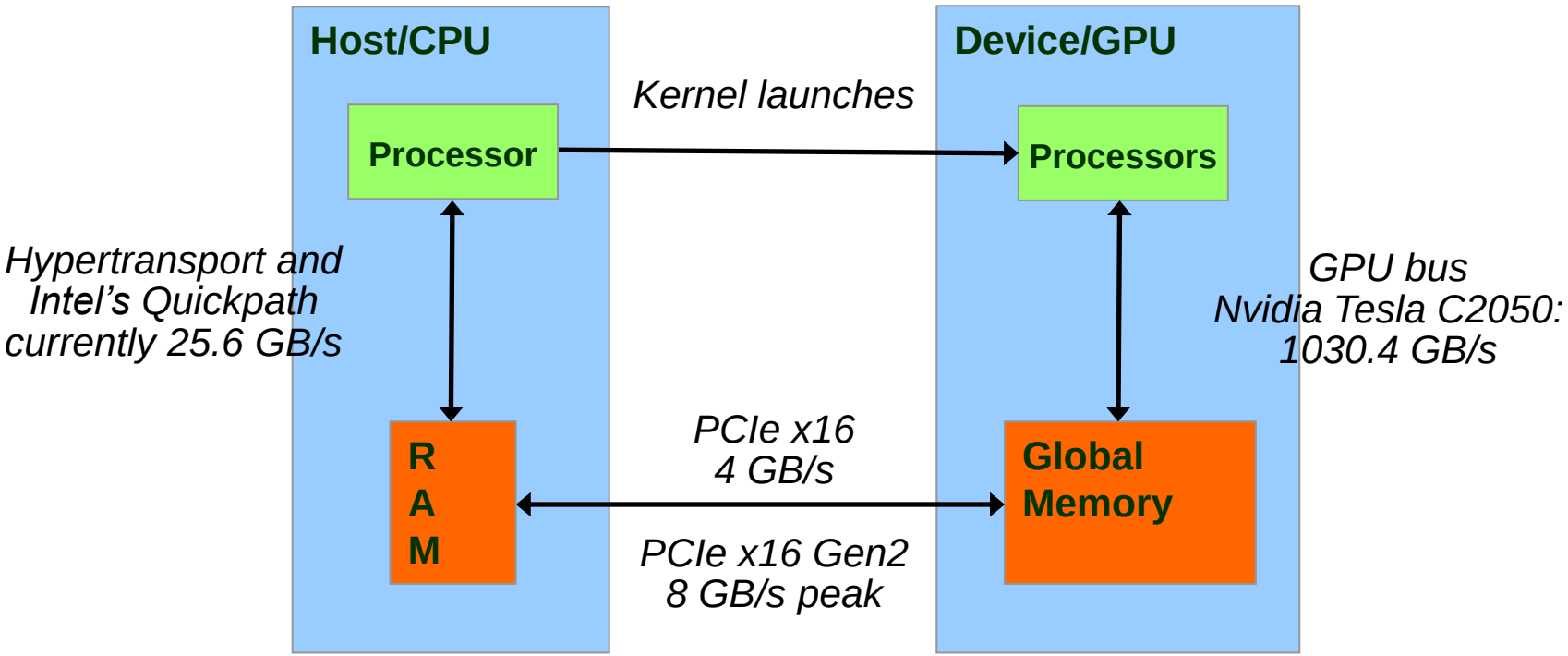
VRIJE UNIVERSITEIT BRUSSEL

Other limit: bandwidth

- Nvidia GTX 280:
  - 1.1 GHz memory clock
  - 141 GB/s
- Nvidia Tesla C2050:
  - 1.5 GHz memory clock
  - 144 GB/s
- Nvidia GTX 1080
  - 2.5 GHz memory clock
  - 320 GB/s

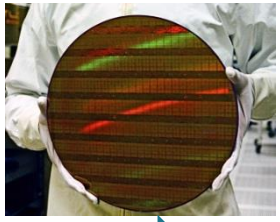
VRIJE UNIVERSITEIT BRUSSEL

Host (CPU) – Device (GPU)

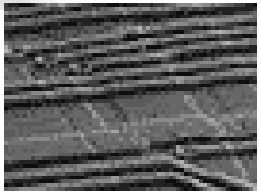


Example: real-time image processing

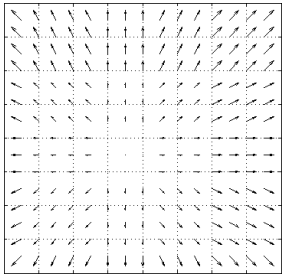
KLA-Tencor
Accelerating Yield



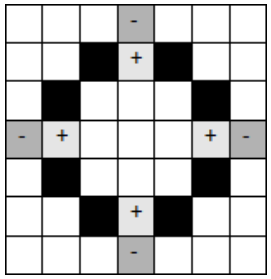
Images of 20MegaPixels



Pixel rescaling



lens correction



pattern detection

CPU gives only 4 fps
next generation machines need 50 fps
GPUs deliver 70 fps

VRIJE UNIVERSITEIT BRUSSEL

Example: pixel transformation (FPN)

```
byte transform(byte pixel, byte gain, byte gain_divide, byte
offset)
{
    int x = (pixel * gain / gain_divide) + offset;
    if (x < 0)
        x = 0;
    if (x > 255)
        x = 255;
    return x;
 }
```

Pixel transformation

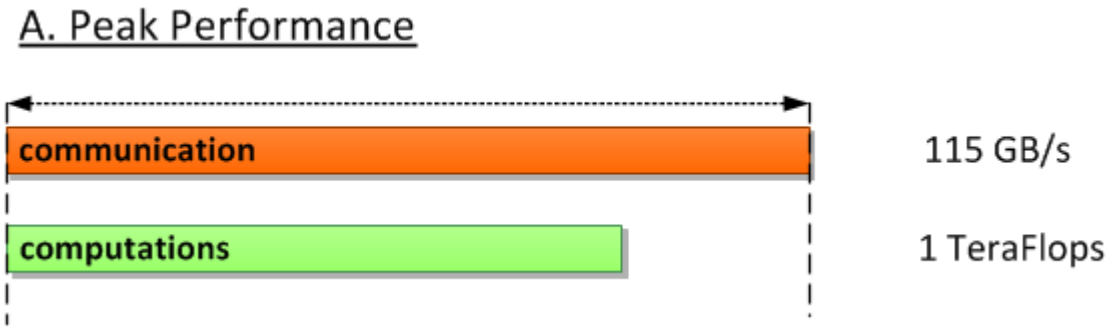
- Performance on Tesla C2050
- 1 pixel is represented by 1 byte [0-255]
  - Per pixel: read 4 bytes (pixel & gain & divide & offset) and write 1 byte
- Integer operations: performance is half of floating point operations
- **Pixel transformation**: around 6 operations (1 index calculation, 3 integer calculations and 2 comparisons)

P_{mem} *(bytes/s)*	115 GB/s	P_{ops} *(ops/s)*	500 Gops/s	
bytes/pixel	5	*Ops/pix*	6	CI=1,2
P_{mem}xCI *(pix/s)*	**23 Gpix/s**	**Pix/s**	**83 Gpix/s**	

Memory-bound (minimum of both)

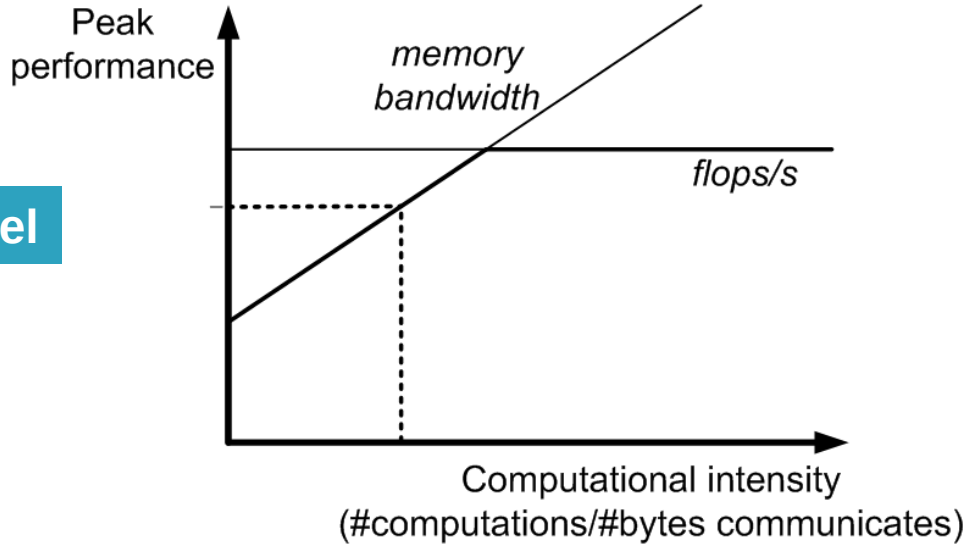
CI = Computational Intensity

What is taking longer: memory transfer or the computations?



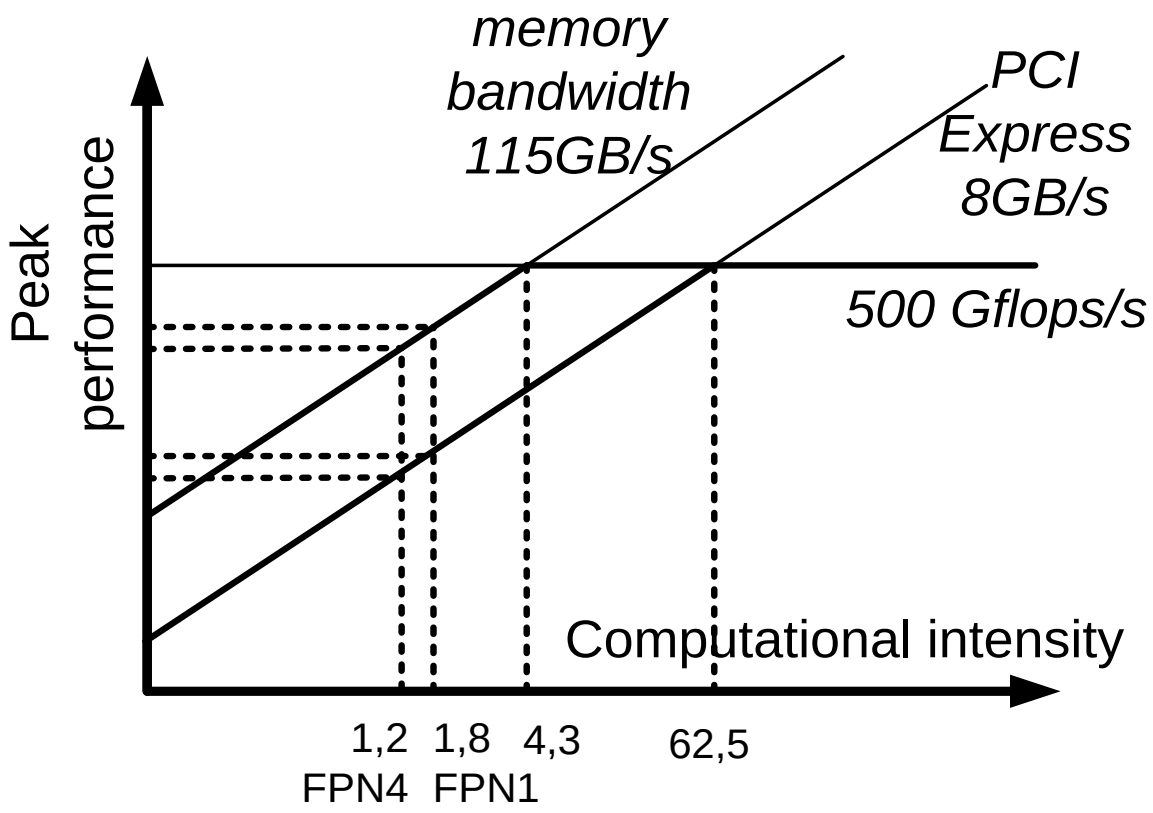
Depends on Computational Intensity (CI)

Roofline model



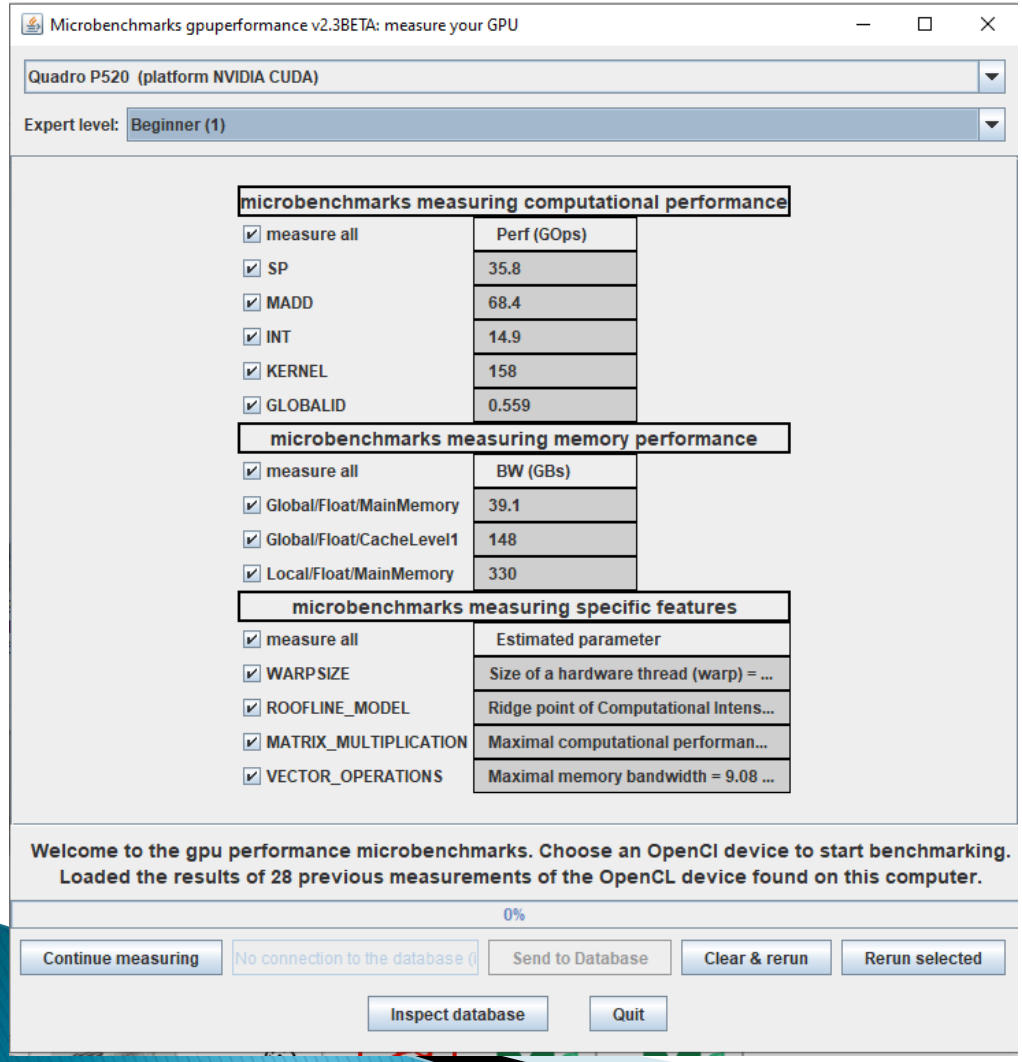
Equation of Memory line: peak performance = CI * BW

Roofline model applied to pixel transformation



VRIJE UNIVERSITEIT BRUSSEL

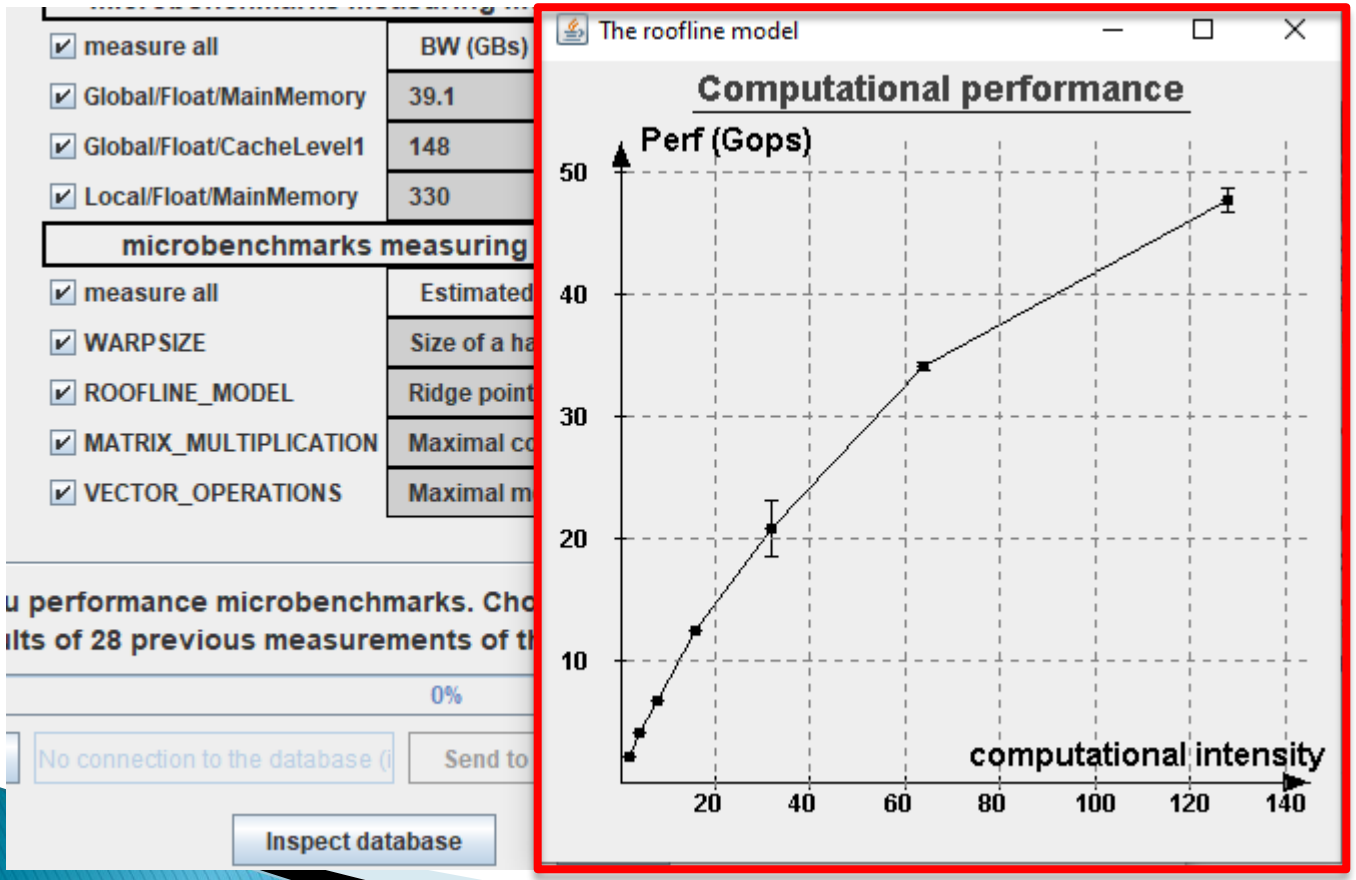
Measure your GPU's performance



- www.gpuperformance.org
- Java app
- Uses OpenCL, so can be used for every GPU
- The GPUs are automatically identified
- Uses small programs (microbenchmarks) to experimentally measure performance
- Computational and memory performance.

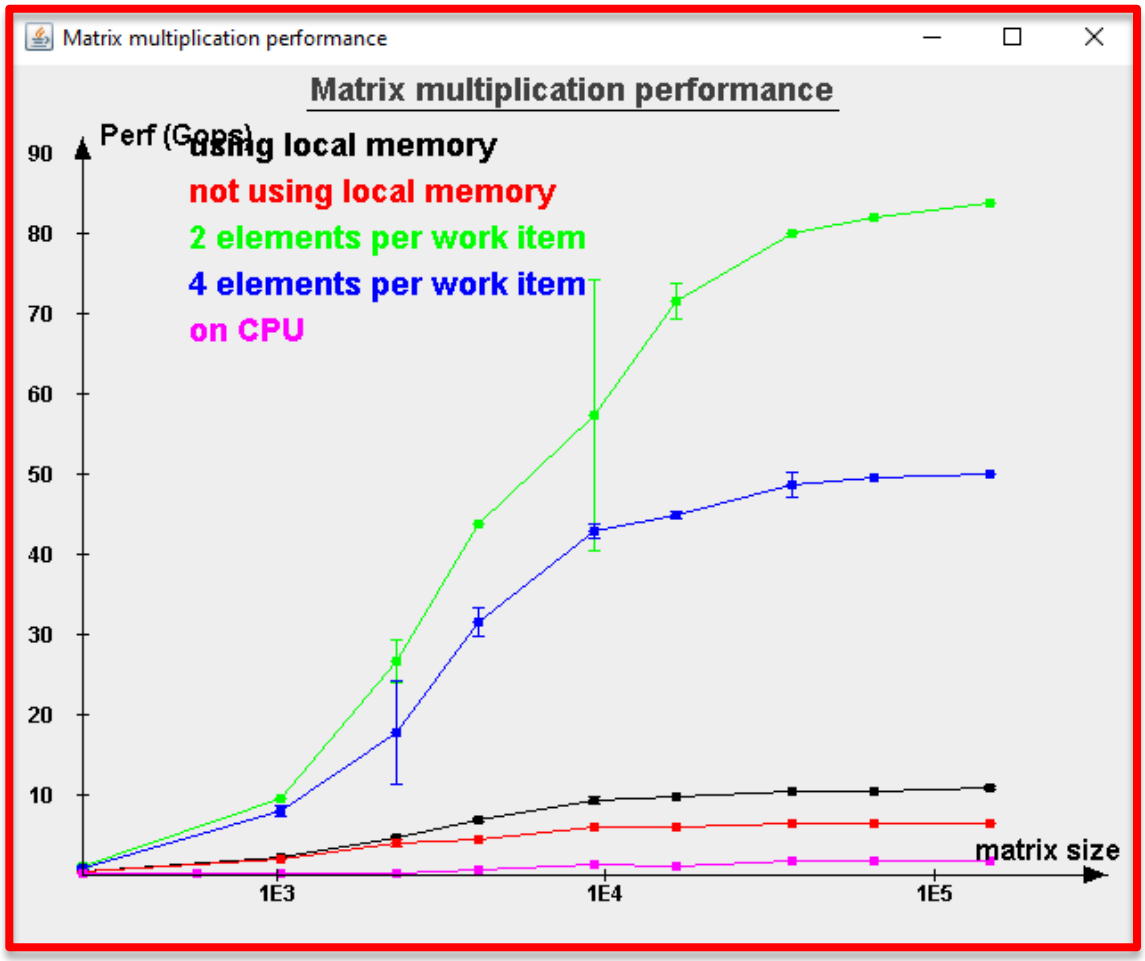
Experimental Roofline Model

- Double click on dark gray fields to see the data



VRIJE UNIVERSITEIT BRUSSEL

Matrix Multiplication



- Optimal version reaches peak performance of 83Gflops!