VRIJE UNIVERSITEIT BRUSSEL

GPU Computing

Lesson 1: The Power of GPUs

Jan Lemeire
2020-2021

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

VRIJE UNIVERSITEIT BRUSSEL

Course Assignments

- Mini-project 30%
  - Small kernel with variations
  - Performance experiments
- Project 70%
  - With defense


ELITEGROUP
GeForce
8800 GTX
NVIDIA

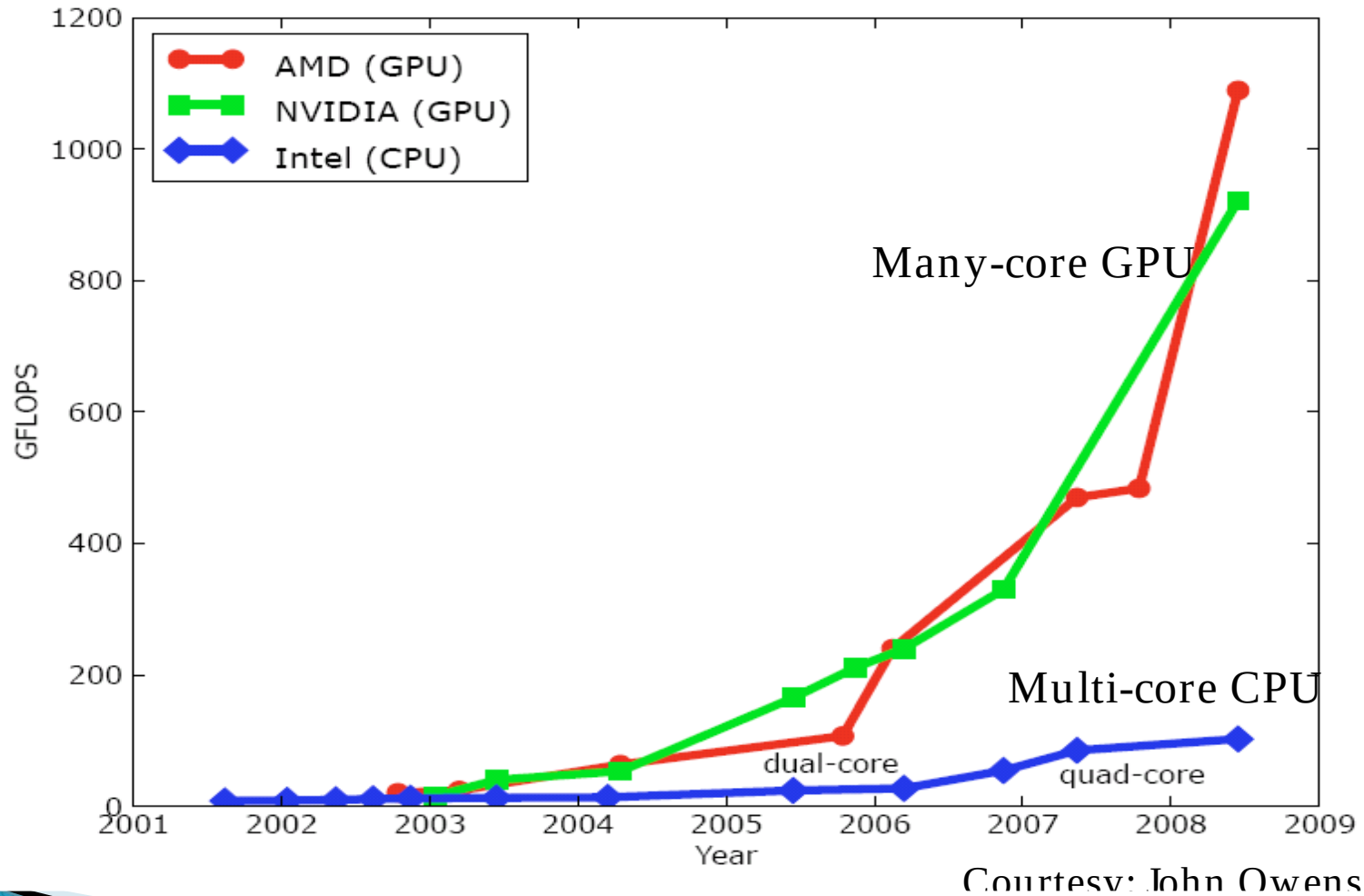
versus


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

http://fastra.ua.ac.be

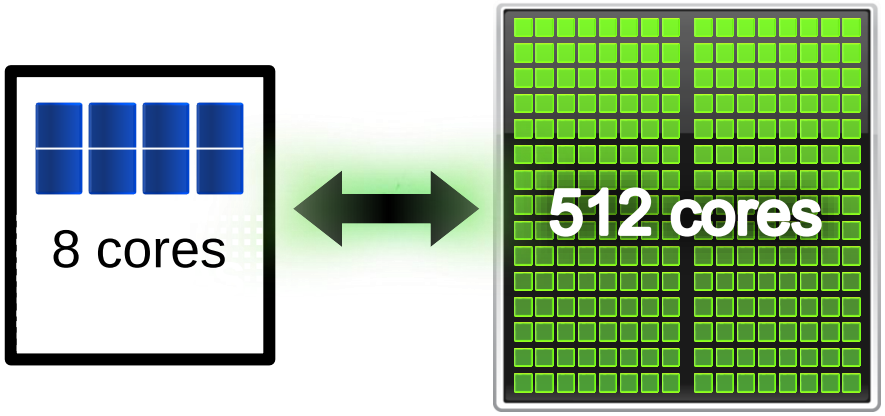
Collection of 8 graphical cards in PC

FASTRA 8 cards = 8x128 processors = 4000 euro

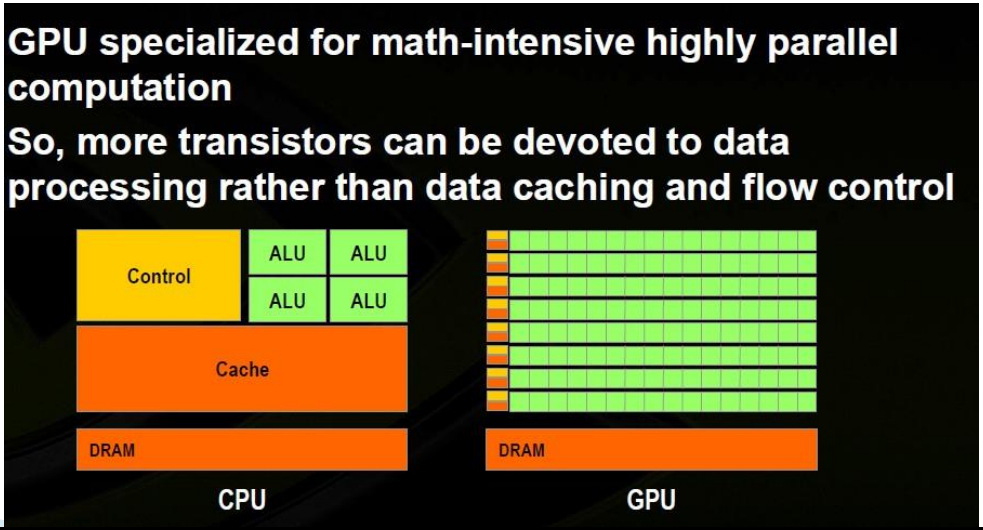
Similar performance as University's supercomputer (512 regular desktop PCs) that costed 3.5 million euro in 2005

"**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 5 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)

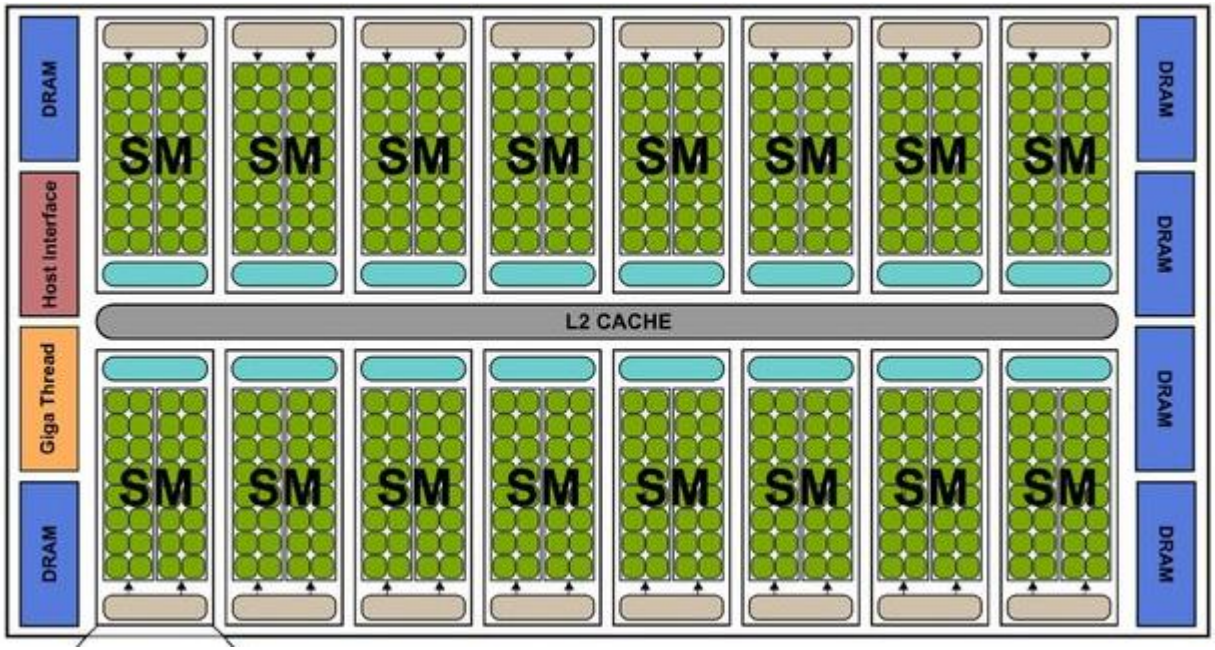
VRIJE UNIVERSITEIT BRUSSEL

Processing elements

- The Processing Elements (PEs) of GPUs are called by Nvidia:
  - Scalar Processors (SPs): single-precision floating point operations
  - CUDA cores (in the data sheets)
- The PEs are grouped into Compute Units (CUs):
  - called by Nvidia as Streaming MultiProcessors (MPs or SMs)
  - Is what we would call a *core* (see later)
- The number of PEs per CU are fixed for each GPU generation
- The number of CUs varies per GPU (determines power & price)

VRIJE UNIVERSITEIT BRUSSEL

A GPU consists of Compute Units /Streaming Multiprocessors (SMs)



VRIJE UNIVERSITEIT BRUSSEL

Each Compute Unit consists of Processsing Elements

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

Nvidia Architecture	Clock freq MHz	PEs per CU	SFUs per CU	DPs per CU	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 Compute Units (CUs) or Streaming MultiProcessors (MPs) grouping a number of Processing Elements (PEs) or 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
    = **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 (clocks per second)
    = **1030 GFlops**

Memory bandwidth

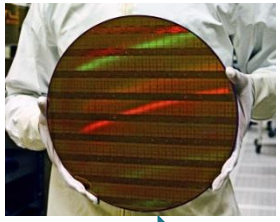
VRIJE UNIVERSITEIT BRUSSEL

Other limit: bandwidth

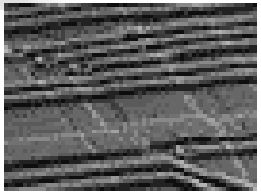
- Nvidia GTX 280:
  - 1.1 GHz memory clock
  - 141 GB/s
- Nvidia Tesla C2050:
  - 1.5 GHz memory clock
  - 144 GB/s

Example: real-time image processing

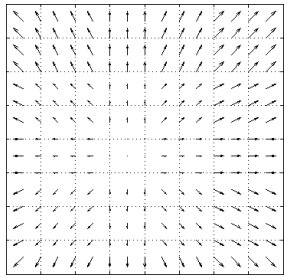
KLA-Tencor
Accelerating Yield



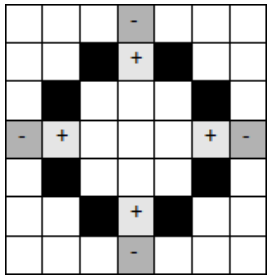
Images of
20MegaPixels



Pixel rescaling



lens correction



pattern detection

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

VRIJE UNIVERSITEIT BRUSSEL

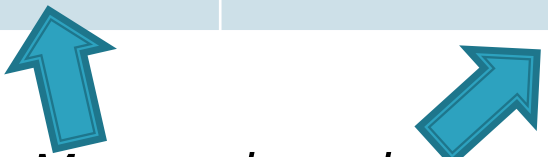
Example: pixel transformation (FPN)

```
usgn_8  transform(usgn_8 in, sgn_16 gain, sgn_16 gain_divide,
sgn_8 offset)
{
    sgn_32 x;

    x = (in * 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 & offset) and write 1 byte
- Integer operations: performance is half of floating point operations
- **Pixel transformation**: typically 6 operations (1 index calculation, 3 integer calculations and 2 comparisons)

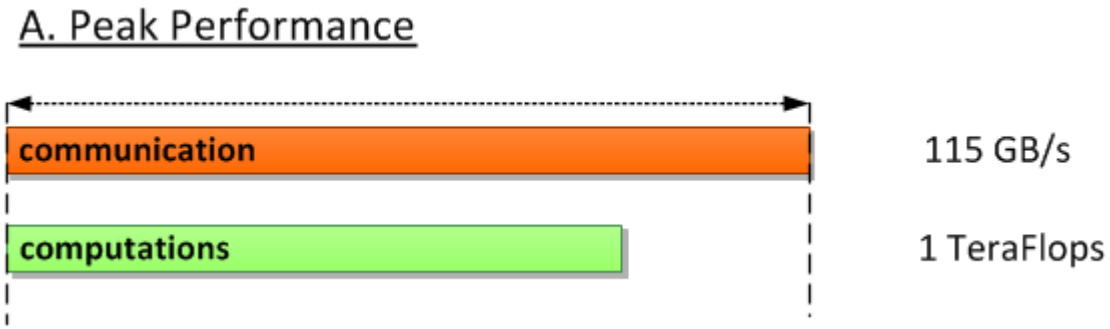
P_{mem} *(bytes/s)*	115 GB/s	P_{ops} *(ops/s)*	500 Gops/s	
bytes/pixel	5	*Ops/pix*	6	CI=1,2
P_{mem}xCI *(pix/s)*	**23 Gpix/s**	**Pix/s**	**83 Gpix/s**	



Memory-bound (minimum of both)

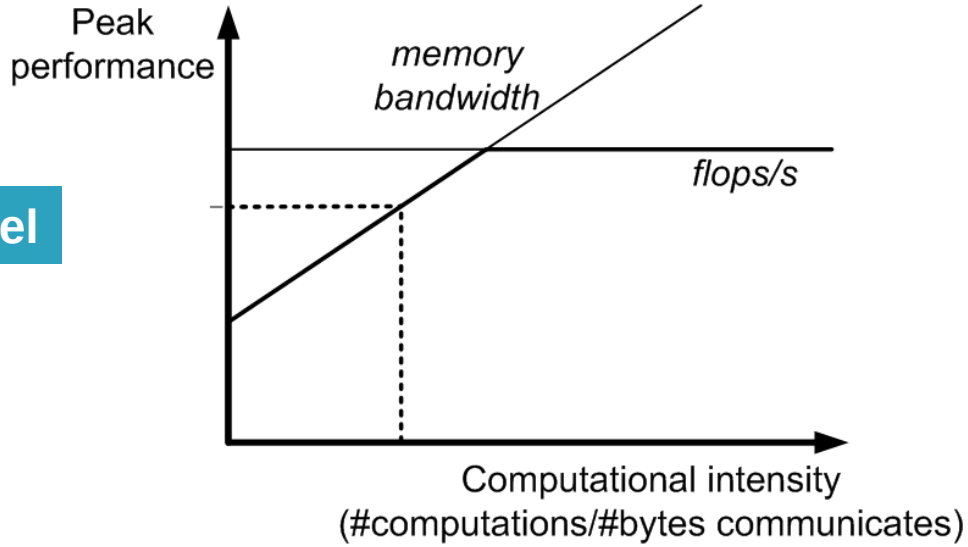
CI = Computational Intensity

What is taking longer: memory transfer or the computations?



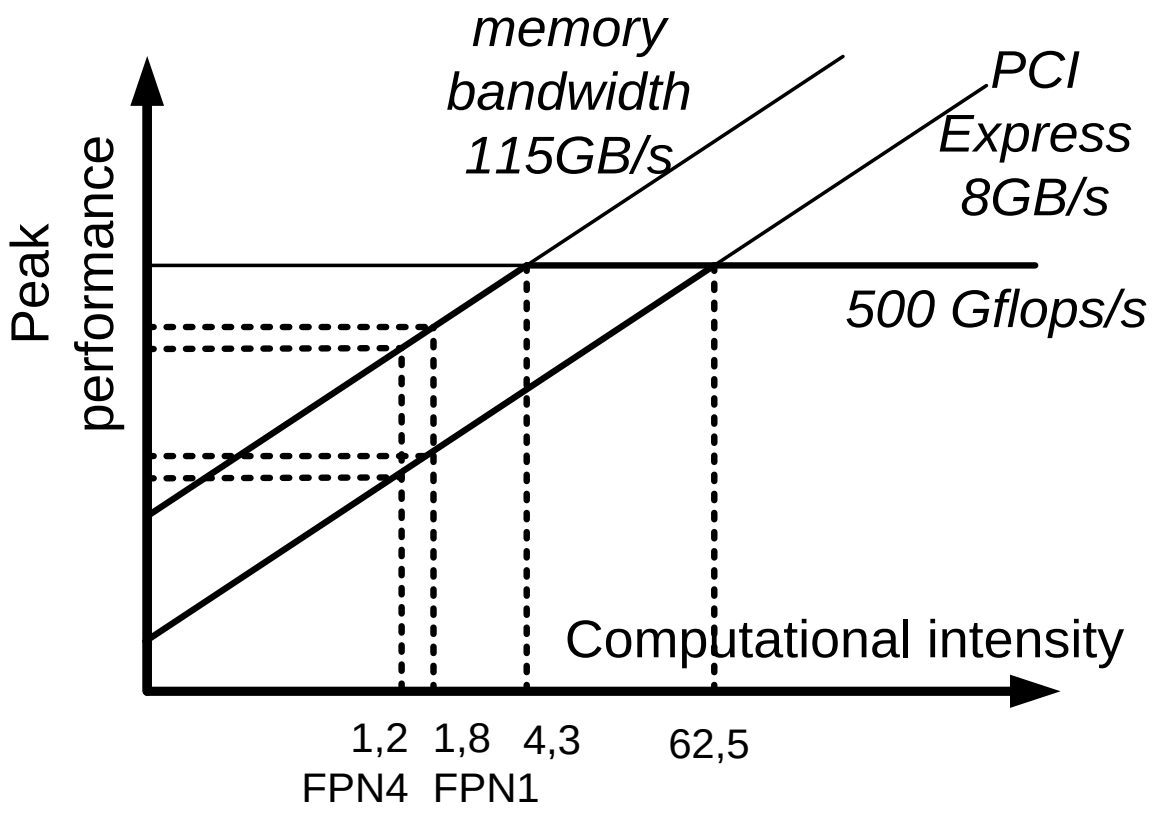
Depends on Computational Intensity (CI)

Roofline model



Equation of Memory line: peak performance = CI * BW

Roofline model applied to pixel transformation



Computational microbenchmarks

www.gpuperformance.org

See paper Lemeire 2016:

Jan Lemeire, Jan G. Cornelis, Laurent Segers, Microbenchmarks for GPU characteristics: the occupancy roofline and the pipeline model, Procs of 24th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP), Heraklion, Greece, 2016

VRIJE UNIVERSITEIT BRUSSEL

Java app

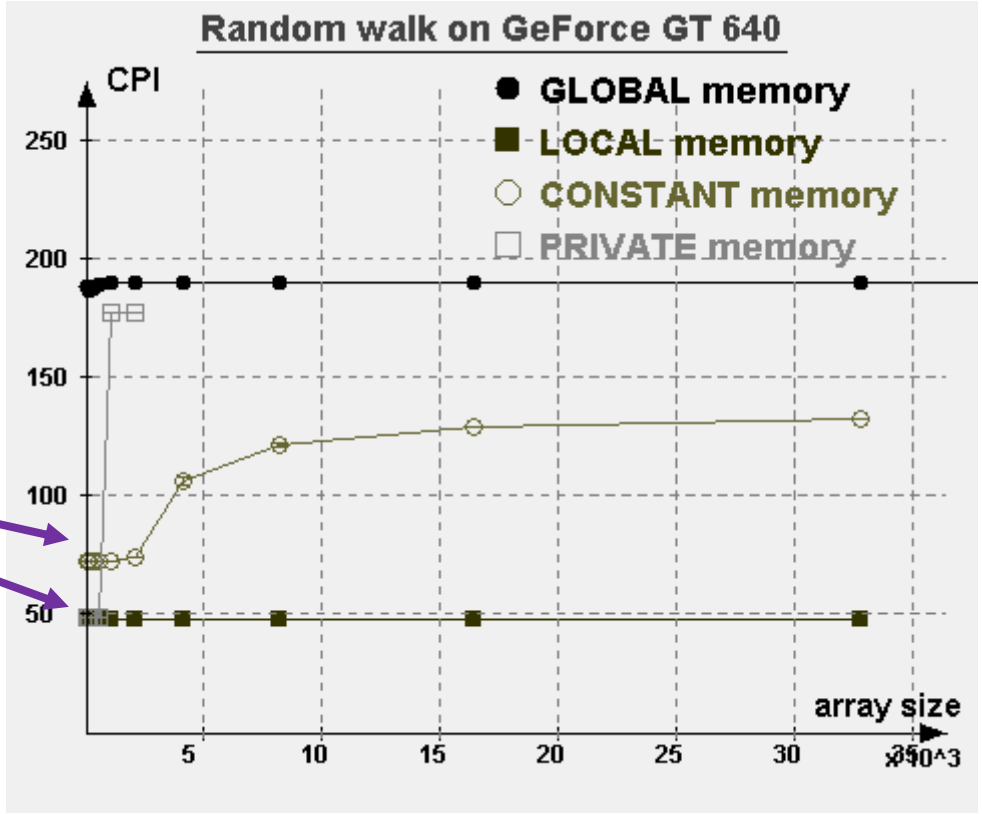
Microbenchmarks: measure your GPU

platform 1, device 1: GeForce GTX 650 Ti

			Computational performance	
iType	Peak	lambda	Lambda	Ridge point
SP	467.2	0.28	10.80	11264
MADD	1001.9	0.13	5.89	11264
INT	308.2	0.43	5.73	11264
SF	nope	nope	nope	nope
DP	nope	nope	nope	nope
			Memory performance	
Global	11.3	11.18	...	...
Char	34.0	3.84	80.98	1024
Char2	60.7	4.31	79.68	1024
Float	69.0	7.56	81.93	512
Float2	72.7	14.45	90.58	256
Float4	74.0	28.24	173.09	256
Local	200.0	0.66	...	...
Constant	83.7	1.57	...	...
Private	1820.0	0.07	...	...

Memory benchmarks

Latencies for different memory types



caching