GPU Computing with NVIDIA CUDA

WAGENINGEN
UNIVERSITY & RESEARCH

Introduction to parallel computing

Wageningen on 21.06.2019

heiko.loewe@dell.com
christian.schramm@dell.com

most images courtesy of NVIDIA

DELL EMC

Introduction to GPU computing with CUDA

A little bit of background

How could you increase the speed of a computing process?

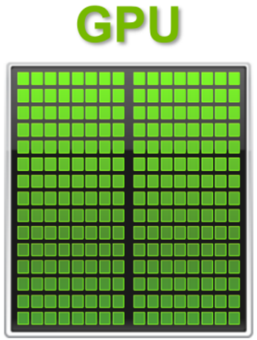
-higher clock speed
-more work per clock cycle
-more processors

Introduction to GPU computing with CUDA

A little bit of background

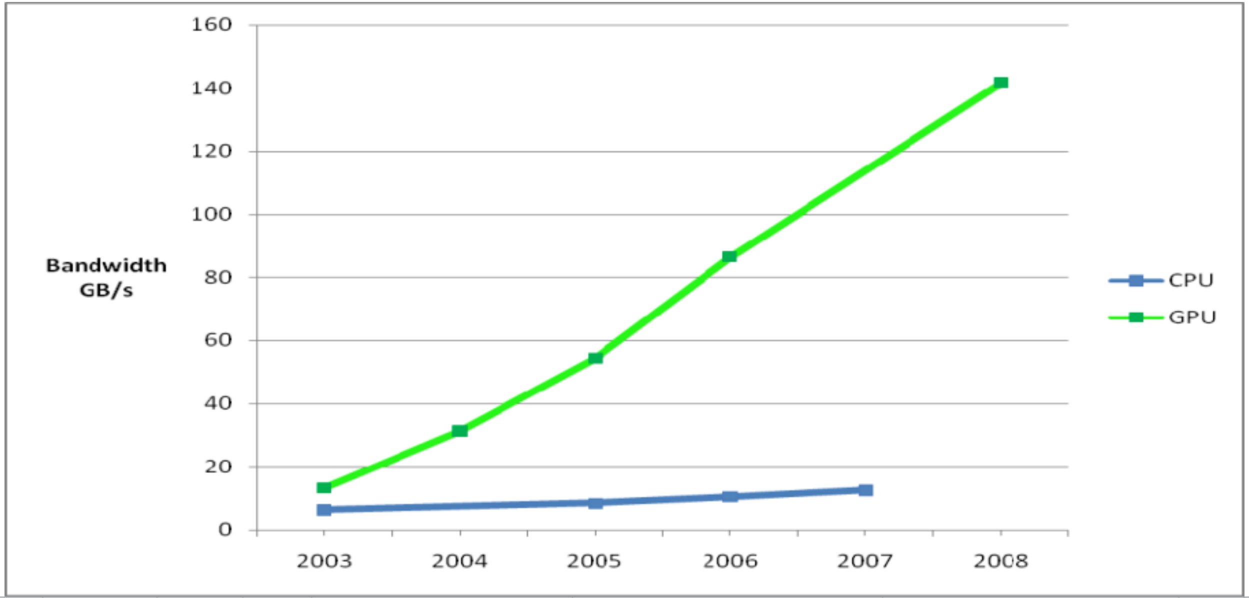
Central Processing Unit (CPU)
-consists of a few cores
-each one is powerful and optimized for **sequential** processing.

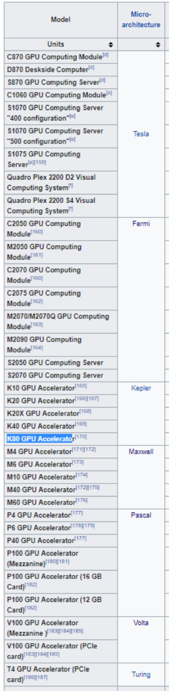
Graphic Processing Unit (GPU)
-consists of hundreds and thousands of smaller, less powerful cores
-the architecture is designed for handling multiple tasks **simultaneously**.



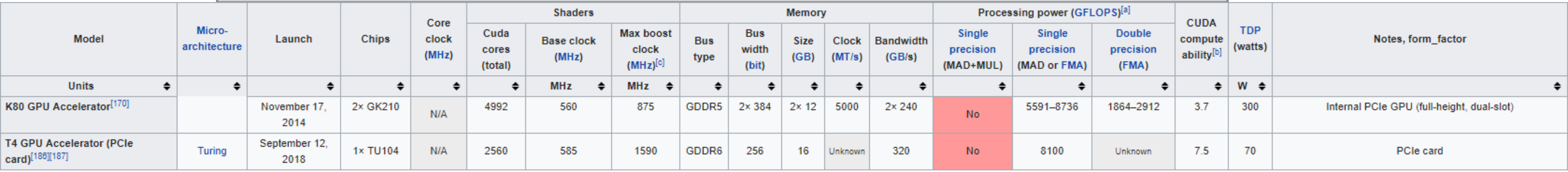
CPU versus GPU

Supercomputing revolution





Model	Micro-architecture
Units	
C870 GPU Computing Module[d]	Tesla
D870 Deskside Computer[d]	
S870 GPU Computing Server[d]	
C1060 GPU Computing Module[e]	
S1070 GPU Computing Server "400 configuration"[e]	
S1070 GPU Computing Server "500 configuration"[e]	
S1075 GPU Computing Server[e][159]	
Quadro Plex 2200 D2 Visual Computing System[f]	
Quadro Plex 2200 S4 Visual Computing System[f]	
C2050 GPU Computing Module[160]	Fermi
M2050 GPU Computing Module[161]	
C2070 GPU Computing Module[160]	
C2075 GPU Computing Module[162]	
M2070/M2070Q GPU Computing Module[163]	
M2090 GPU Computing Module[164]	
S2050 GPU Computing Server	
S2070 GPU Computing Server	
K10 GPU Accelerator[165]	Kepler
K20 GPU Accelerator[166][167]	
K20X GPU Accelerator[168]	
K40 GPU Accelerator[169]	
K80 GPU Accelerator[170]	
M4 GPU Accelerator[171][172]	Maxwell
M6 GPU Accelerator[173]	
M10 GPU Accelerator[174]	
M40 GPU Accelerator[172][175]	
M60 GPU Accelerator[176]	
P4 GPU Accelerator[177]	Pascal
P6 GPU Accelerator[178][179]	
P40 GPU Accelerator[177]	
P100 GPU Accelerator (Mezzanine)[180][181]	
P100 GPU Accelerator (16 GB Card)[182]	
P100 GPU Accelerator (12 GB Card)[182]	
V100 GPU Accelerator (Mezzanine)[183][184][185]	Volta
V100 GPU Accelerator (PCIe card)[183][184][185]	
T4 GPU Accelerator (PCIe card)[186][187]	Turing



Model	Micro-architecture	Launch	Chips	Core clock (MHz)	Shaders			Memory					Processing power (GFLOPS)[a]			CUDA compute ability[b]	TDP (watts)	Notes, form_factor
					Cuda cores (total)	Base clock (MHz)	Max boost clock (MHz)[c]	Bus type	Bus width (bit)	Size (GB)	Clock (MT/s)	Bandwidth (GB/s)	Single precision (MAD+MUL)	Single precision (MAD or FMA)	Double precision (FMA)			
Units						MHz	MHz										W	
K80 GPU Accelerator[170]		November 17, 2014	2× GK210	N/A	4992	560	875	GDDR5	2× 384	2× 12	5000	2× 240	No	5591–8736	1864–2912	3.7	300	Internal PCIe GPU (full-height, dual-slot)
T4 GPU Accelerator (PCIe card)[186][187]	Turing	September 12, 2018	1× TU104	N/A	2560	585	1590	GDDR6	256	16	Unknown	320	No	8100	Unknown	7.5	70	PCIe card

What is a GPU?

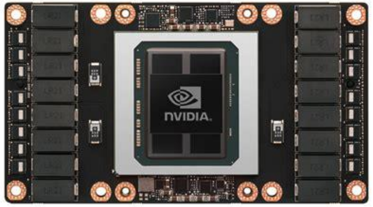
A little bit of background

3dfx Voodoo and NVIDIA GeForce 256

VGA interfaces

NVIDIA K80 (look familiar?)

You cannot even attach a monitor.
…still PCI though



NVIDIA Pascal

This one does not even have PCI.
Clips right onto the main board.

What is a GPU?

A little bit of background

1999

GEFORCE 256

NVIDIA introduced the first GPU, the GeForce 256, in 1999 and remains one of the major players in the market.

2002

RADEON 9700

ATI Technologies released Radeon 9700 in 2002 and called in VPU (visual).

2003

APIS

Before 2003 GPUs could only be addressed by OpenGL and Direct3D.

2004

BROOKGPU

First attempt of establishing GPGPU API. Last update: 2007

2007

CUDA

NVIDIA releases CUDA.

What is CUDA?

A little bit of background

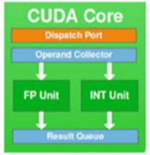
CUDA stands for **C**ompute **U**nified **D**evice **A**rchitecture.
It is a parallel computing platform (using a GPU) and a programming model (using code). CUDA is an extension of C and fully supports C++.

Flynn's Taxonomy introduced in 1966:
Single Instruction Single Datum (PC)
Single Instruction Multiple Data (GPU)
Multiple Instruction Single Datum (Fault Tolerance)
Multiple Instruction Multiple Data (distributed systems, autonomous processors)

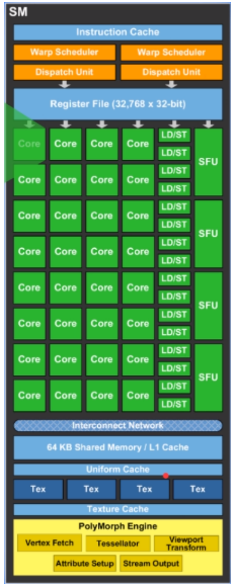
What is a GPU?

Architecture

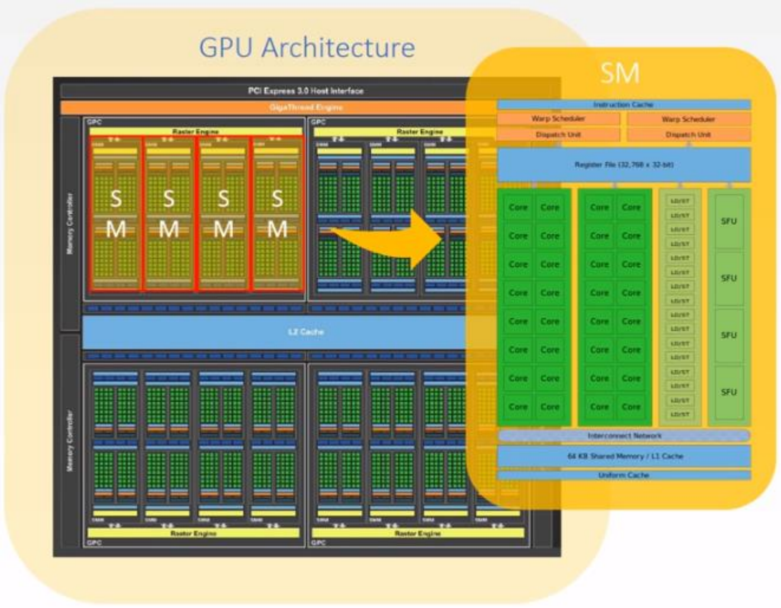
CUDA Core:
-Smallest building block of a GPU.
-Executes computations (“threads”)



Stream Multiprocessor:
-Collection of CUDA Cores including a Scheduler

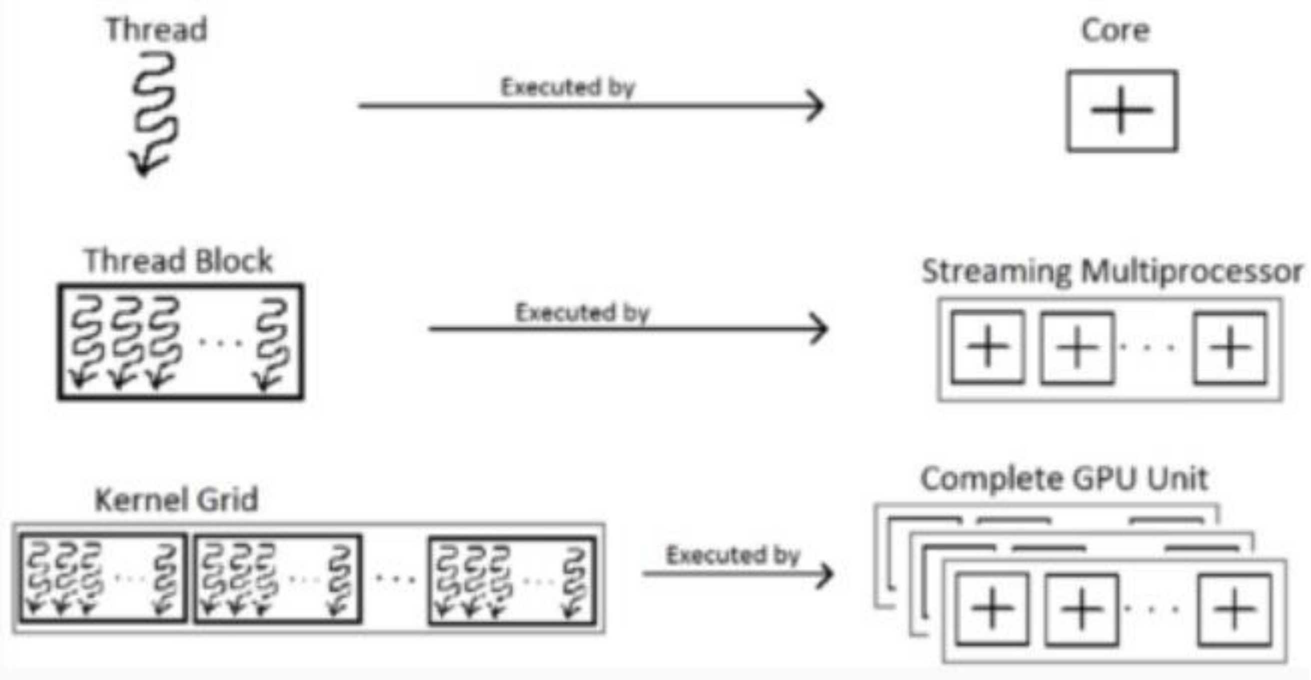


GPU:
-Collection of SMs



What is a GPU?

Execution



Heterogeneous Computing

A little bit of background

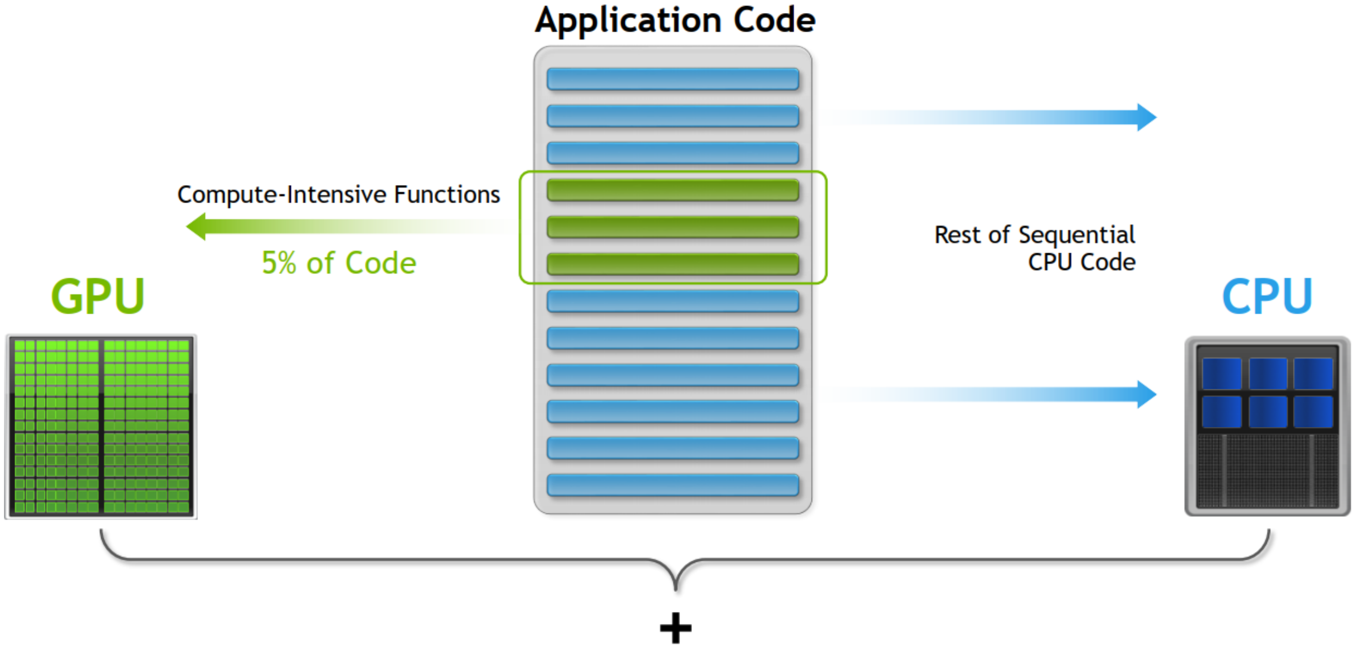
Host:
The computer that has it own CPU and memory (“host memory”)

Device:
GPU that has it own memory (“device memory”)

NVIDIA K80

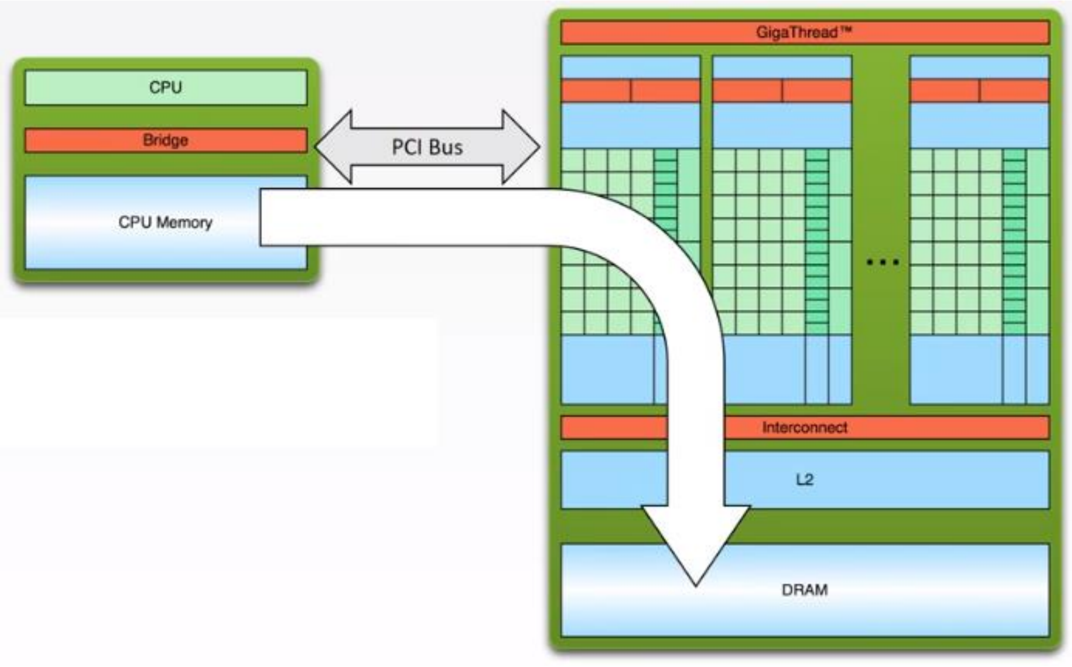
Putting these two together

How GPU acceleration works



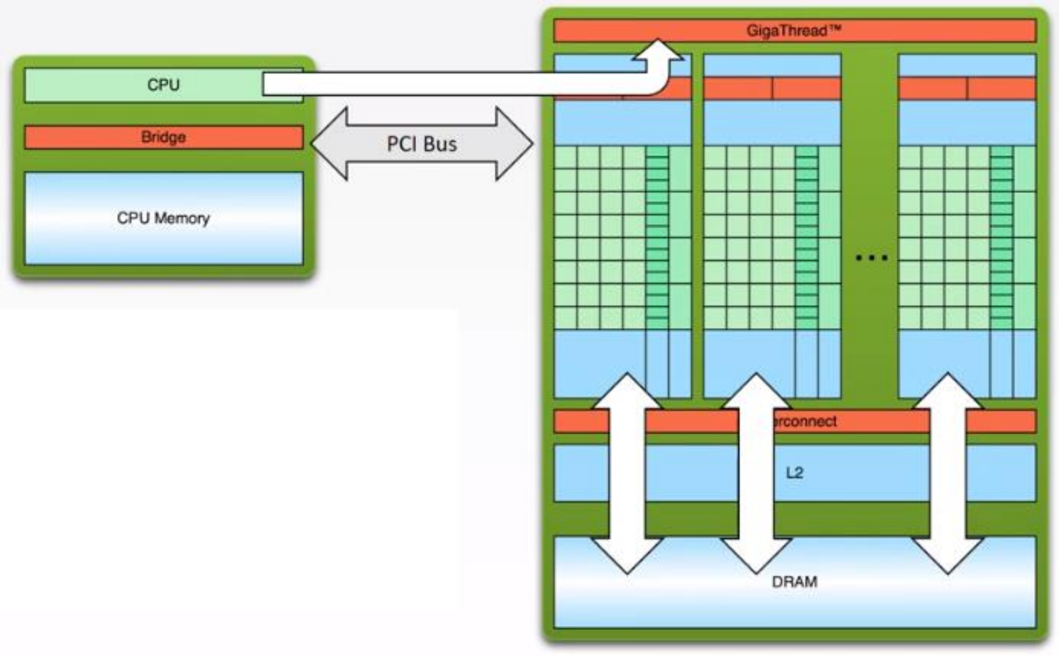
How GPU acceleration works

Copy input data from CPU memory to GPU memory



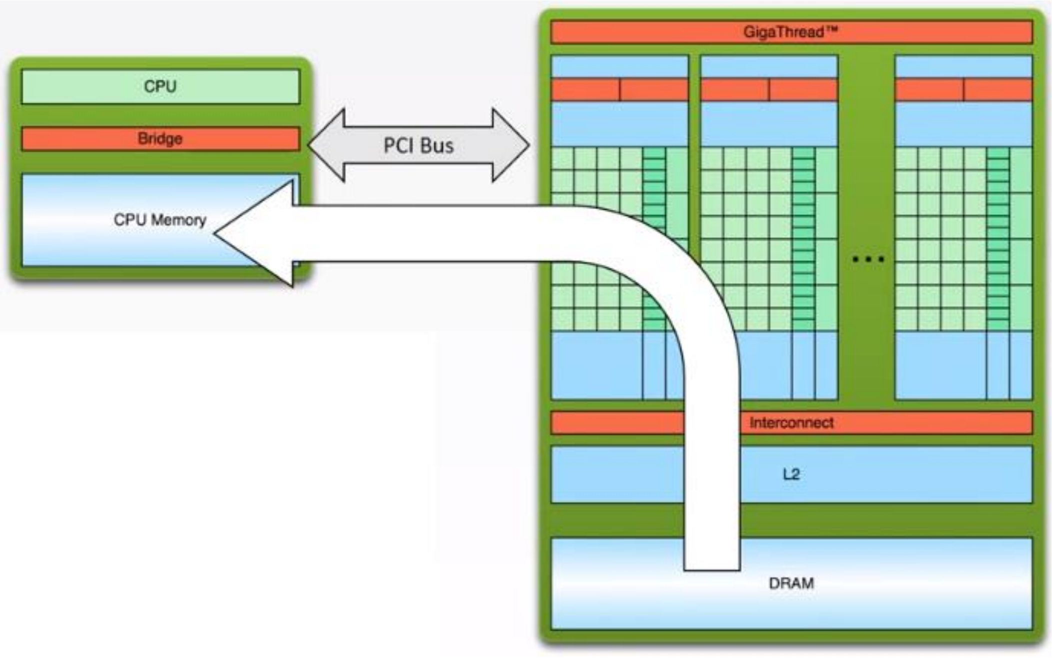
How GPU acceleration works

Load GPU program and execute, caching data on chip for performance



How GPU acceleration works

Copy results from GPU memory to CPU memory



CUDA Toolkits

…and over to Heiko

Programming Approaches

Libraries

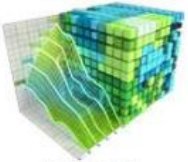
"Drop-in" Acceleration

Programming Languages

Maximum Flexibility

Development Environment

Nsight IDE



NVIDIA Visual Profiler

CUDA Profiling Tools Interface

CUDA-GDB Debugger

CUDA MEMCHECK

Language Support

C

C++

Fortran

python

LLVM COMPILER INFRASTRUCTURE

Compile new languages to CUDA

DELLEMC

DELL EMC