center for
excellence in parallel
programming

CUDA Introduction

Georges-Emmanuel Moulard
Paul Karlshöfer

Bull
atos technologies

Why GPU computing ?

- What is A GPU ?

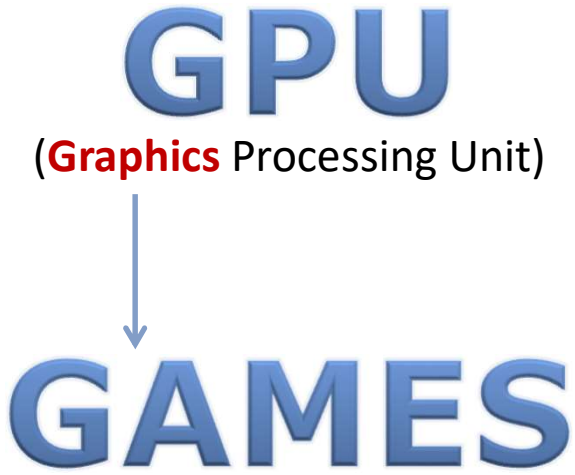
Nvidia

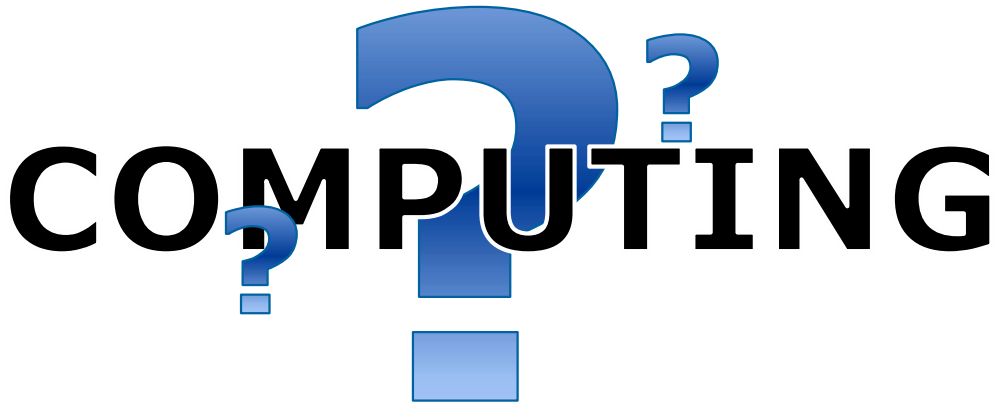
AMD

Why GPU computing ?

- What is A GPU ?

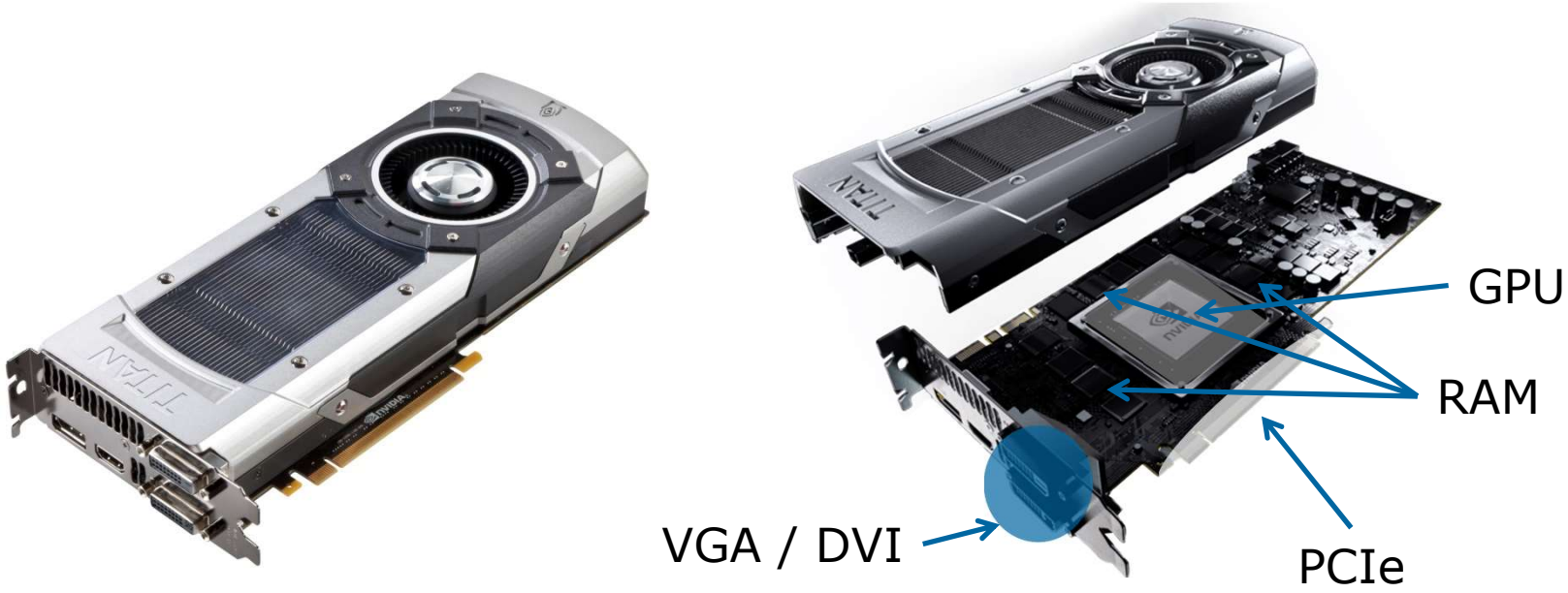
Why GPU computing ?





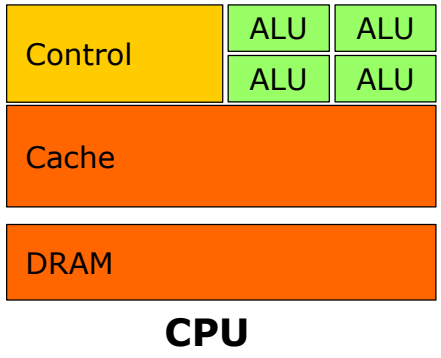
Why GPU computing ?

- GPU : COMPONENTS



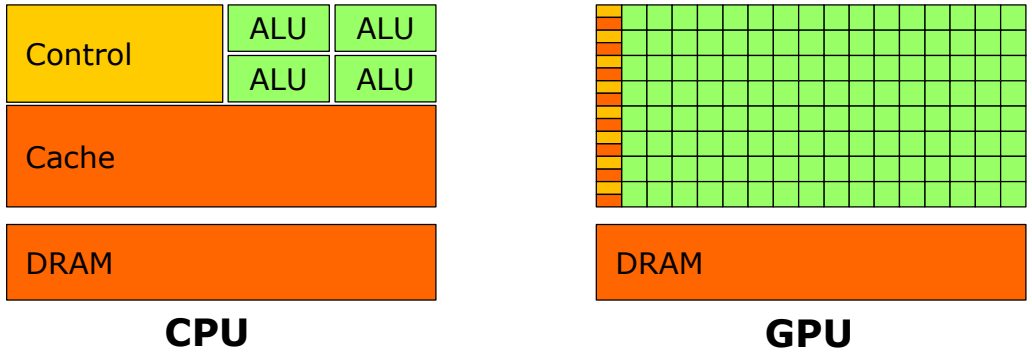
Why GPU Computing ?

- CPU vs GPU : Transistors repartition



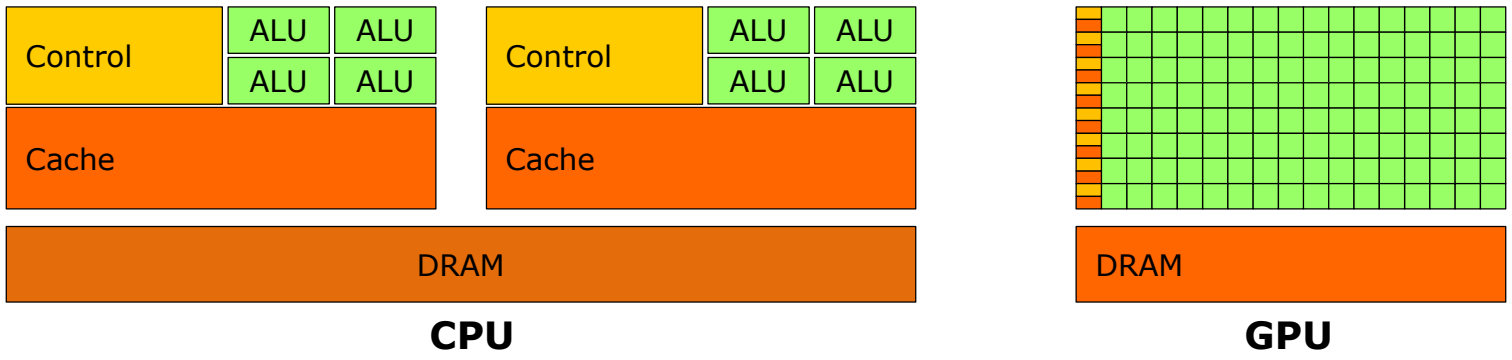
CPU vs GPU : Transistors Repartition

- GPU comes from graphics rendering :
  - Single Instruction Multiple Data (compute-intensive & few control)
  - Throughput oriented (thousands of pixels simultaneously)



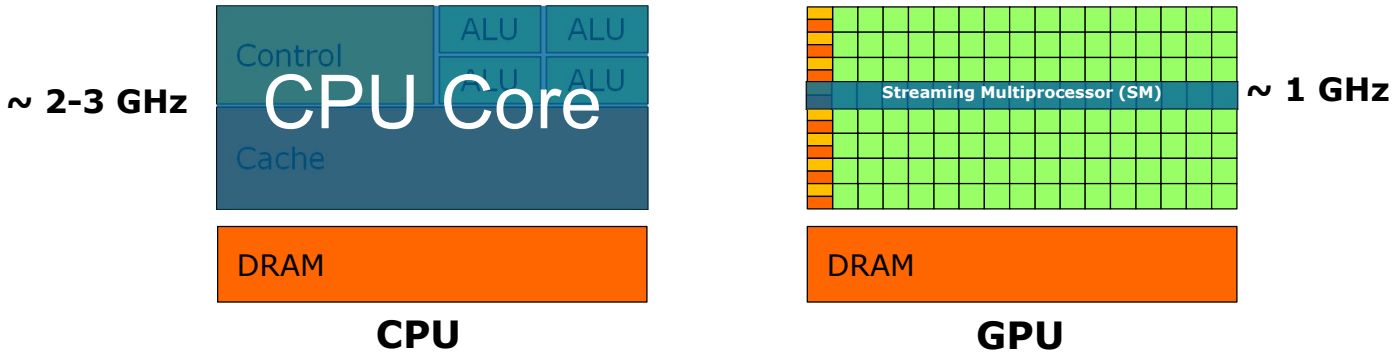
CPU vs GPU : Transistors Repartition

- GPU comes from graphics rendering :
  - Single Instruction Multiple Data (compute-intensive & few control)
  - Throughput oriented (thousands of pixels simultaneously)



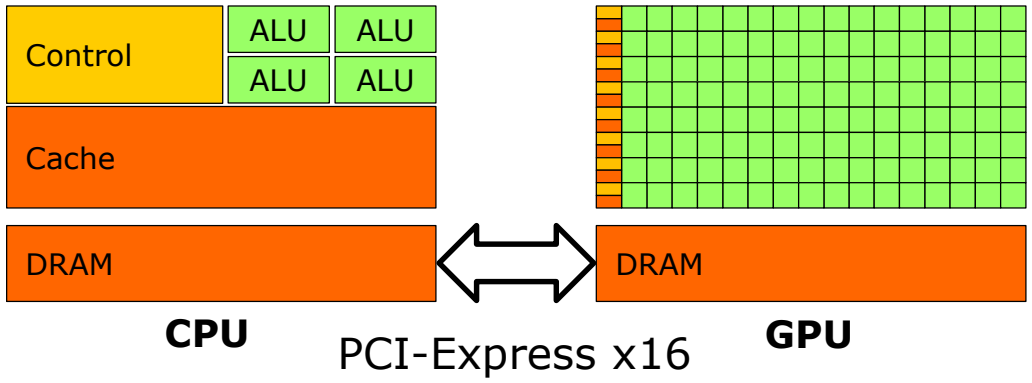
CPU vs GPU : Transistors Repartition

- GPU comes from graphics rendering :
  - Single Instruction Multiple Data (compute-intensive & few control)
  - Throughput oriented (thousands of pixels simultaneously)
  - 1 SM comparable to 1 CPU core : fetch, decode, load operand, execute



CPU vs GPU : Transistors Repartition

- GPU comes from graphics rendering :
  - Single Instruction Multiple Data (compute-intensive & few control)
  - Throughput oriented (thousands of pixels simultanesouly)
  - 1 SM comparable to 1 CPU core : fetch, decode, load operand, execute
  - Connected through PCI-Express



Why GPU Computing ?

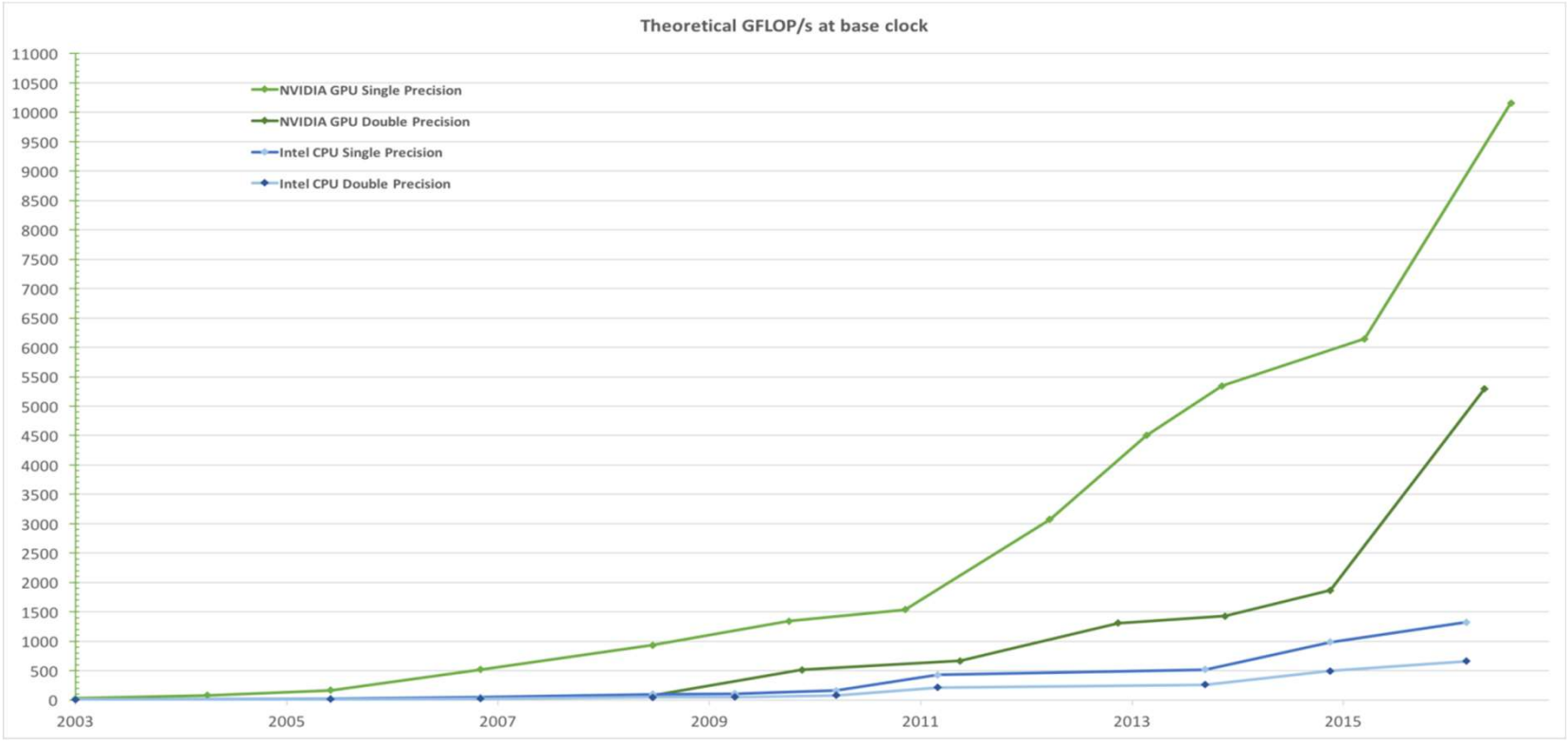
- Since years ~2000, frequency scaling is over… We are now scaling cores !

- GPU is a massively multi-threaded many-cores architecture
  - Thousands of threads executed in parallel
    - Kepler (K80) able to run 2048*13=26 624 threads in parallel
    - Pascal (P100) 2048*56=114 688 threads
    - Volta (V100) 2048*80=163 840 threads

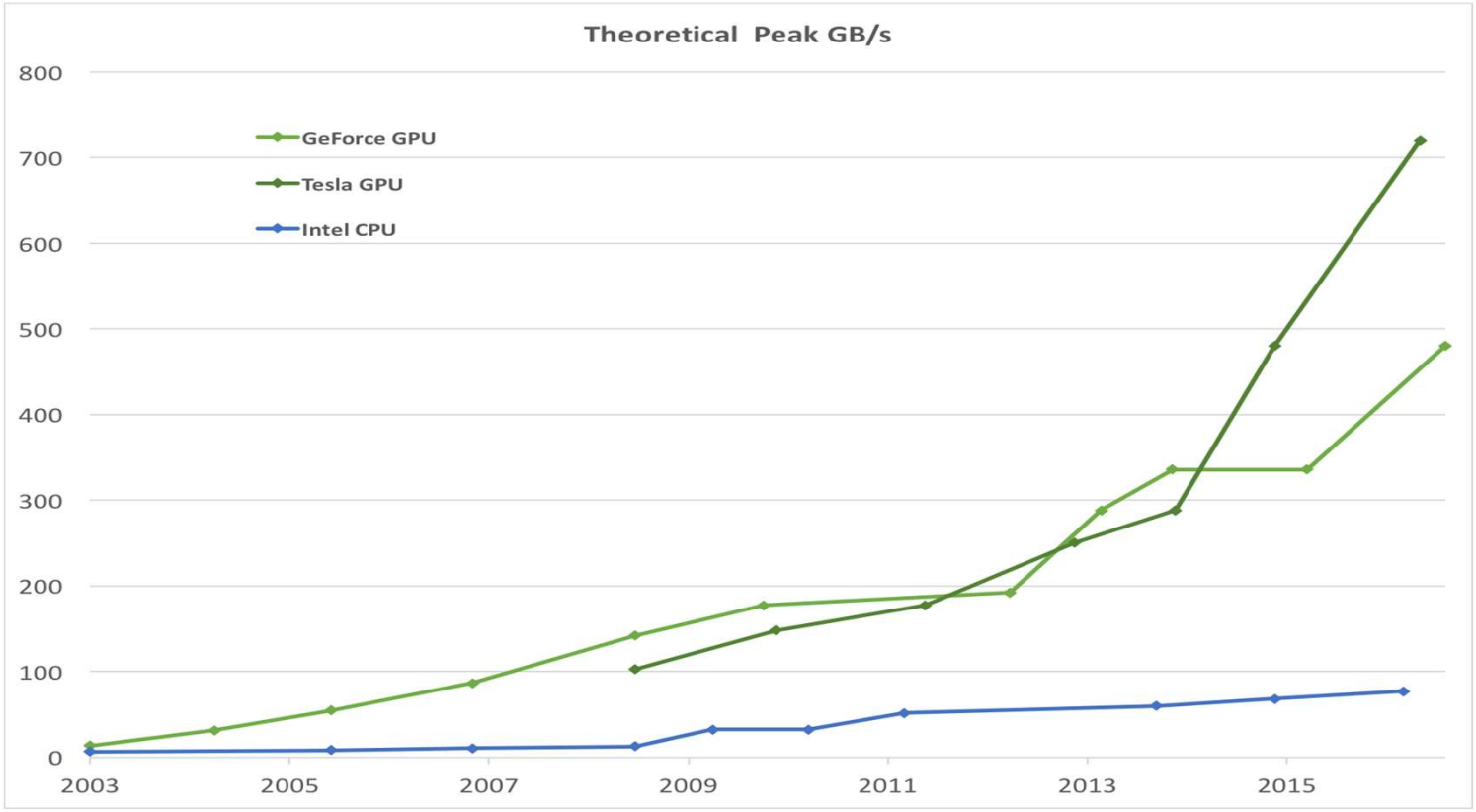
- GPU is a relatively cheap commodity component
  - Big Market, low production prices… comparable to CPU

- GPU is fast

CPU vs GPU : Theoretical Computational Peak



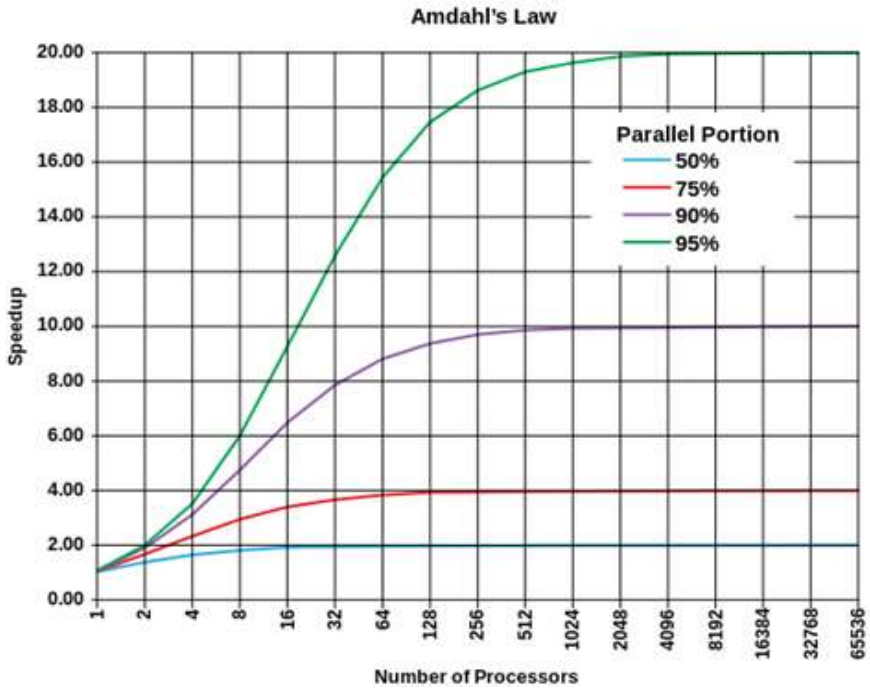
CPU vs GPU : Theoretical Bandwidth



Parallelism and Performance

- GPU Speed-up
  - x100-1000 performance gain?

- Amdahl's law:
  - P : fraction made parallel
  - 1-P : fraction kept serial
  - N : amount of processors
  - $$S(n) = \frac{1}{(1-P)+\frac{P}{N}}$$



« Debunking the 100X GPU vs. CPU myth »
« Closing the Ninja Performance Gap through Traditional Programming and Compiler Technology »

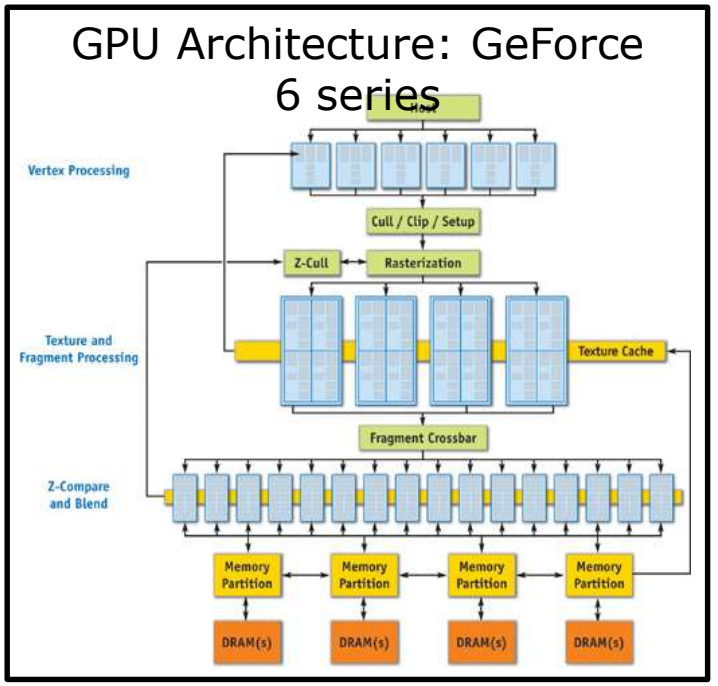
GPU : From Graphics to Computing

16/09/2019

Bull
atos technologies

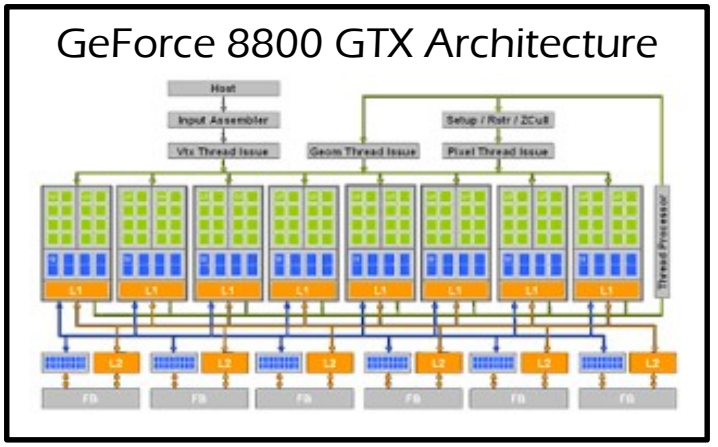
Hardware Architecture Evolution

- Before 2007:
  - Hardware pipeline
  - Lack of double precision support
  - Graphics APIs : directX, OpenGL, Cg



Hardware Architecture Evolution

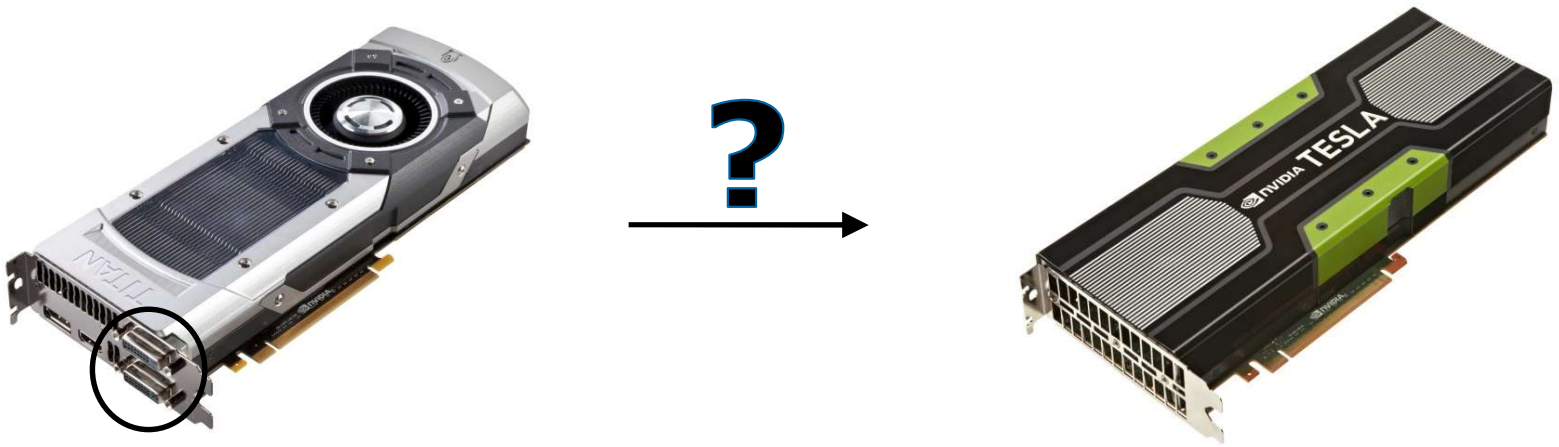
- 2007 : CUDA (Compute Unified Device Architecture)
  - Unified graphics & compute architecture : Tesla, Fermi, Kepler, …, Volta
  - Programming model: C extensions



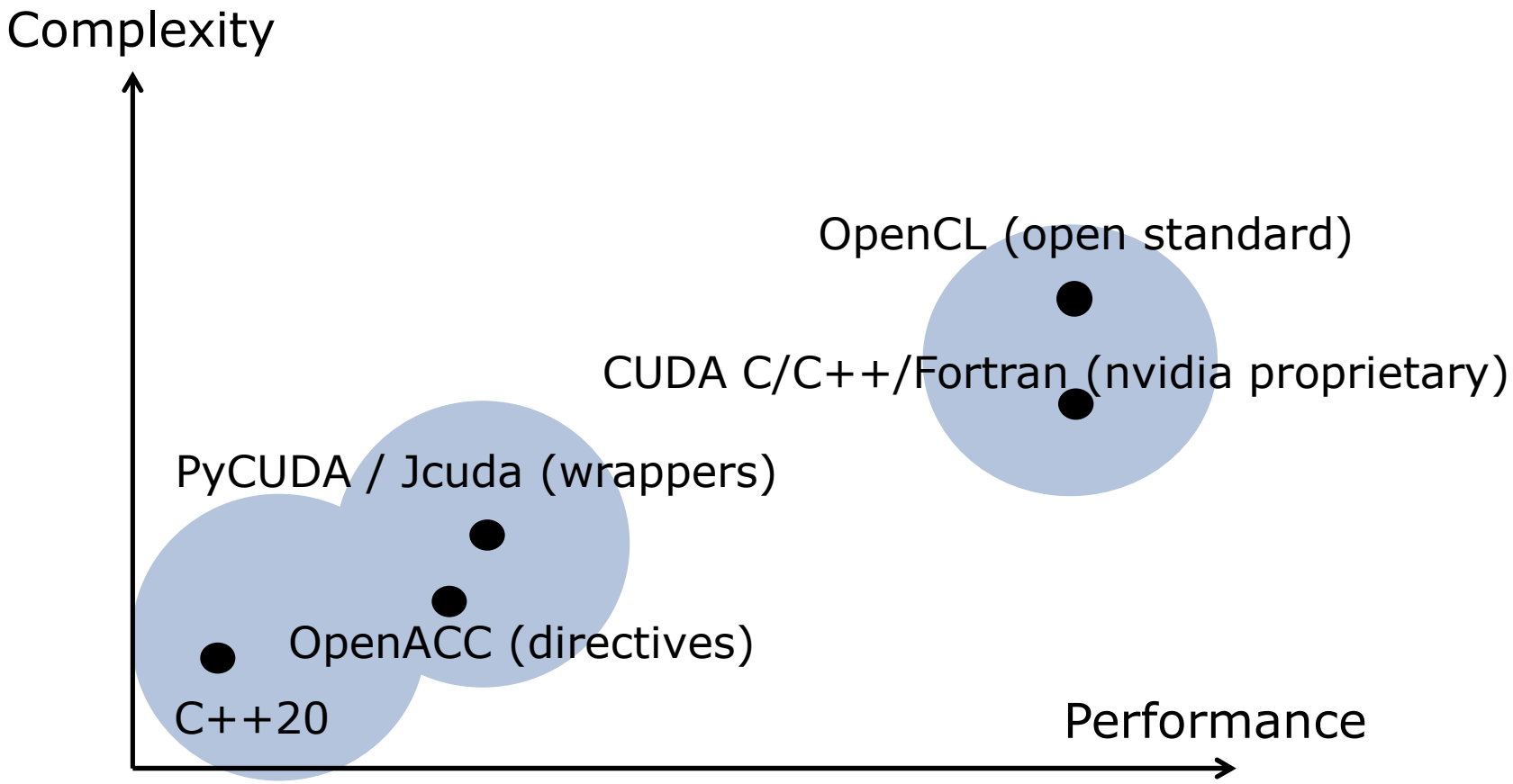
Hardware Architecture Evolution

- 2007 : CUDA (Compute Unified Device Architecture)
  - Unified graphics & compute architecture : Tesla, Fermi, Kepler, …, Volta
  - Programming model: C extensions

- Nvidia Tesla GPU has since then become the product range for HPC



Programming Languages



Nvidia GPU Product Range

Nvidia Product	Market	Graphics	Double Precision	ECC
GeForce	gamers	x	~	
Quadro	CAD	x	~	~
Tesla	HPC / AI		x	x
Tegra	Mobile	x	~	

Software Overview

16/09/2019

Bull
atos technologies

Software Overview

- Get CUDA from https://developer.Nvidia.com/cuda-downloads

- Package contains:
  - CUDA Driver
  - CUDA Programming model
  - CUDA Toolkit (sdk):

    - GPU accelerated libraries : cuFFT, cuBLAS, cuSPARSE, cuRAND, NPP, Thrust, CUDA Math Library
    - Debugging
      - **cuda-gdb**
      - **cuda-memcheck** for memory errors, race conditions, bad memory accesses, etcProfiling
      - **nvprof** command-line profiling
      - **nvvp** visual profiler
      - **Nsight**

Extended C

- **Declspecs**
  - global, device, shared, local, constant
- **Keywords**
  - threadIdx, blockIdx
- **Intrinsics**
  - __syncthreads
- **Runtime API**
  - Memory, symbol, execution management
- **Function launch**

```
__device__ float filter[N];
__global__ void convolve (float *image)  {
  __shared__ float region[M];
  ...

  region[threadIdx.x] = image[i];
  __syncthreads()
  ...
  image[j] = result;
}
```

```
// Allocate GPU memory
void *myimage = cudaMalloc(bytes)
```

```
// 100 blocks, 10 threads per block
convolve<<<100, 10>>> (myimage);
```

Software Overview

- ONLINE CUDA Documentation:
  - GPU Management & Deployment Documentation:
    http://docs.nvidia.com/deploy/index.html
  - CUDA Toolkit Documentation
    http://docs.nvidia.com/cuda/index.html
- OFFLINE CUDA Documentation:
  - All documentations can be found in directory:
    /opt/cuda/(*CUDA_VERSION*)/doc/pdf/ (/opt/cuda default installation path)
- Start with "CUDA C Programming Guide"
  https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html

Troubleshooting

16/09/2019

Bull
atos technologies

Troubleshooting

- Setting CUDA environment (example)

```
export CUDA_HOME=/opt/cuda/9.2
export PATH=$CUDA_HOME/bin:$PATH
export LD_LIBRARY_PATH=$CUDA_HOME/lib64:$LD_LIBRARY_PATH
export CUDA_INC=$CUDA_HOME/include
```

- CUDA environment test
  - $> nvcc –V

```
nvcc: NVIDIA (R) Cuda compiler driver
Copyright (c) 2005-2018 NVIDIA Corporation
Built on Wed_Apr_11_23:16:29_CDT_2018
Cuda compilation tools, release 9.2, V9.2.88
```

Installation & Troubleshooting

- Is Nvidia Tesla GPU visible inside Linux ?
  - $> /sbin/lspci | grep –i Nvidia

```
04:00.0 3D controller: Nvidia Corporation Device 102d (rev a1)
05:00.0 3D controller: Nvidia Corporation Device 102d (rev a1)
86:00.0 3D controller: Nvidia Corporation Device 102d (rev a1)
87:00.0 3D controller: Nvidia Corporation Device 102d (rev a1)
```

- Which nvidia driver is currently running ?
  - $> cat /proc/driver/nvidia/version

```
NVRM version: NVIDIA UNIX x86_64 Kernel Module  396.26  Mon Apr 30 18:01:39 PDT 2018
GCC version:  gcc version 4.8.5 20150623 (Red Hat 4.8.5-16) (GCC)
```

Installation & Troubleshooting

- Nvidia tool for quick check
  - $> nvidia-smi
    - NVIDIA System Management Interface
    - It provides monitoring information for Tesla and selected Quadro devices

- Download the exercise from Spartan:

/home_nfs_robin_ib/bkarlshoeferp/work/CUDA_intern/cuda_tps_intern.zip

```
srun -t 00:05:00 -p CSL-6248_GPU_hdr100_192gb_2933 --gres=gpu:1
nvidia-smi
```

LAB: nvidia-smi

- Using nvidia-smi find the:
  - Driver version
  - GPU model
  - Memory usage
  - GPU usage
  - GPU temperature
  - Power consumption
  - Compute processes

NVIDIA-SMI Basic Functionalities

- Nvidia tool for quick check
  - $> nvidia-smi

```
+-----------------------------------------------------------------------------+
| Nvidia-SMI 340.29     Driver Version: 340.29                                |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80           On   | 0000:04:00.0     Off |                    0 |
| N/A   66C    P0   108W / 149W |    240MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+
|   1  Tesla K80           On   | 0000:05:00.0     Off |                    0 |
| N/A   55C    P0   136W / 149W |    260MiB / 11519MiB |     68%   E. Process |
+-------------------------------+----------------------+----------------------+
|   2  Tesla K80           On   | 0000:86:00.0     Off |                    0 |
| N/A   47C    P0   123W / 149W |    252MiB / 11519MiB |     61%   E. Process |
+-------------------------------+----------------------+----------------------+
|   3  Tesla K80           On   | 0000:87:00.0     Off |                    0 |
| N/A   57C    P0   115W / 149W |    242MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Compute processes:                                               GPU Memory |
|  GPU       PID  Process name                                     Usage      |
|=============================================================================|
|    0     11205  Nvidia-cuda-mps-server                          182MiB |
|    1     11211  Nvidia-cuda-mps-server                          202MiB |
|    2     11213  Nvidia-cuda-mps-server                          194MiB |
|    3     11212  Nvidia-cuda-mps-server                          184MiB |
+-----------------------------------------------------------------------------+
```

NVIDIA-SMI Basic Functionalities

- Nvidia tool for quick check
  - $> nvidia-smi
    - **Driver**
    - GPU model
    - Memory usage
    - GPU usage
    - GPU temperature
    - Power consumption
    - Compute processes

```
+-----------------------------------------------------------------------------+
| Nvidia-SMI 340.29     Driver Version: 340.29                                |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80           On   | 0000:04:00.0     Off |                    0 |
| N/A   66C    P0   108W / 149W |    240MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+
|   1  Tesla K80           On   | 0000:05:00.0     Off |                    0 |
| N/A   55C    P0   136W / 149W |    260MiB / 11519MiB |     68%   E. Process |
+-------------------------------+----------------------+----------------------+
|   2  Tesla K80           On   | 0000:86:00.0     Off |                    0 |
| N/A   47C    P0   123W / 149W |    252MiB / 11519MiB |     61%   E. Process |
+-------------------------------+----------------------+----------------------+
|   3  Tesla K80           On   | 0000:87:00.0     Off |                    0 |
| N/A   57C    P0   115W / 149W |    242MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Compute processes:                                               GPU Memory |
|  GPU       PID  Process name                                     Usage      |
|=============================================================================|
|    0     11205  Nvidia-cuda-mps-server                          182MiB |
|    1     11211  Nvidia-cuda-mps-server                          202MiB |
|    2     11213  Nvidia-cuda-mps-server                          194MiB |
|    3     11212  Nvidia-cuda-mps-server                          184MiB |
+-----------------------------------------------------------------------------+
```

NVIDIA-SMI Basic Functionalities

- Nvidia tool for quick check
  - $> nvidia-smi
    - Driver
    - **GPU model**
    - Memory usage
    - GPU usage
    - GPU temperature
    - Power consumption
    - Compute processes

```
+-----------------------------------------------------------------------------+
| Nvidia-SMI 340.29     Driver Version: 340.29                                |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80          On    | 0000:04:00.0     Off |                    0 |
| N/A   66C    P0   108W / 149W |    240MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+
|   1  Tesla K80          On    | 0000:05:00.0     Off |                    0 |
| N/A   55C    P0   136W / 149W |    260MiB / 11519MiB |     68%   E. Process |
+-------------------------------+----------------------+----------------------+
|   2  Tesla K80          On    | 0000:86:00.0     Off |                    0 |
| N/A   47C    P0   123W / 149W |    252MiB / 11519MiB |     61%   E. Process |
+-------------------------------+----------------------+----------------------+
|   3  Tesla K80          On    | 0000:87:00.0     Off |                    0 |
| N/A   57C    P0   115W / 149W |    242MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Compute processes:                                               GPU Memory |
|  GPU       PID  Process name                                     Usage      |
|=============================================================================|
|    0     11205  Nvidia-cuda-mps-server                          182MiB |
|    1     11211  Nvidia-cuda-mps-server                          202MiB |
|    2     11213  Nvidia-cuda-mps-server                          194MiB |
|    3     11212  Nvidia-cuda-mps-server                          184MiB |
+-----------------------------------------------------------------------------+
```

NVIDIA-SMI Basic Functionalities

- Nvidia tool for quick check
  - $> nvidia-smi
    - Driver
    - GPU model
    - **Memory usage**
    - GPU usage
    - GPU temperature
    - Power consumption
    - Compute processes

```
+-----------------------------------------------------------------------------+
| Nvidia-SMI 340.29     Driver Version: 340.29                                |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80           On   | 0000:04:00.0     Off |                    0 |
| N/A   66C    P0   108W / 149W |   240MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+
|   1  Tesla K80           On   | 0000:05:00.0     Off |                    0 |
| N/A   55C    P0   136W / 149W |   260MiB / 11519MiB |     68%   E. Process |
+-------------------------------+----------------------+----------------------+
|   2  Tesla K80           On   | 0000:86:00.0     Off |                    0 |
| N/A   47C    P0   123W / 149W |   252MiB / 11519MiB |     61%   E. Process |
+-------------------------------+----------------------+----------------------+
|   3  Tesla K80           On   | 0000:87:00.0     Off |                    0 |
| N/A   57C    P0   115W / 149W |   242MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Compute processes:                                               GPU Memory |
|  GPU       PID  Process name                                     Usage      |
|=============================================================================|
|    0     11205  Nvidia-cuda-mps-server                          182MiB |
|    1     11211  Nvidia-cuda-mps-server                          202MiB |
|    2     11213  Nvidia-cuda-mps-server                          194MiB |
|    3     11212  Nvidia-cuda-mps-server                          184MiB |
+-----------------------------------------------------------------------------+
```

NVIDIA-SMI Basic Functionalities

- Nvidia tool for quick check
  - $> nvidia-smi
    - Driver
    - GPU model
    - Memory usage
    - **GPU usage**
    - GPU temperature
    - Power consumption
    - Compute processes

```
+-----------------------------------------------------------------------------+
| Nvidia-SMI 340.29     Driver Version: 340.29                                |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80           On   | 0000:04:00.0     Off |                    0 |
| N/A   66C    P0   108W / 149W |    240MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+
|   1  Tesla K80           On   | 0000:05:00.0     Off |                    0 |
| N/A   55C    P0   136W / 149W |    260MiB / 11519MiB |     68%   E. Process |
+-------------------------------+----------------------+----------------------+
|   2  Tesla K80           On   | 0000:86:00.0     Off |                    0 |
| N/A   47C    P0   123W / 149W |    252MiB / 11519MiB |     61%   E. Process |
+-------------------------------+----------------------+----------------------+
|   3  Tesla K80           On   | 0000:87:00.0     Off |                    0 |
| N/A   57C    P0   115W / 149W |    242MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Compute processes:                                               GPU Memory |
|  GPU       PID  Process name                                     Usage      |
|=============================================================================|
|    0     11205  Nvidia-cuda-mps-server                          182MiB |
|    1     11211  Nvidia-cuda-mps-server                          202MiB |
|    2     11213  Nvidia-cuda-mps-server                          194MiB |
|    3     11212  Nvidia-cuda-mps-server                          184MiB |
+-----------------------------------------------------------------------------+
```

NVIDIA-SMI Basic Functionalities

- Nvidia tool for quick check
  - $> nvidia-smi
    - Driver
    - GPU model
    - Memory usage
    - GPU usage
    - **GPU temperature**
    - Power consumption
    - Compute processes

```
+-----------------------------------------------------------------------------+
| Nvidia-SMI 340.29     Driver Version: 340.29                                |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80           On   | 0000:04:00.0     Off |                    0 |
| N/A   66C    P0   108W / 149W |    240MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+
|   1  Tesla K80           On   | 0000:05:00.0     Off |                    0 |
| N/A   55C    P0   136W / 149W |    260MiB / 11519MiB |     68%   E. Process |
+-------------------------------+----------------------+----------------------+
|   2  Tesla K80           On   | 0000:86:00.0     Off |                    0 |
| N/A   47C    P0   123W / 149W |    252MiB / 11519MiB |     61%   E. Process |
+-------------------------------+----------------------+----------------------+
|   3  Tesla K80           On   | 0000:87:00.0     Off |                    0 |
| N/A   57C    P0   115W / 149W |    242MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Compute processes:                                               GPU Memory |
|  GPU       PID  Process name                                     Usage      |
|=============================================================================|
|    0     11205  Nvidia-cuda-mps-server                          182MiB |
|    1     11211  Nvidia-cuda-mps-server                          202MiB |
|    2     11213  Nvidia-cuda-mps-server                          194MiB |
|    3     11212  Nvidia-cuda-mps-server                          184MiB |
+-----------------------------------------------------------------------------+
```

NVIDIA-SMI Basic Functionalities

- Nvidia tool for quick check
  - $> nvidia-smi
    - Driver
    - GPU model
    - Memory usage
    - GPU usage
    - GPU temperature
    - **Power consumption**
    - Compute processes

```
+-----------------------------------------------------------------------------+
| Nvidia-SMI 340.29     Driver Version: 340.29                                |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80           On   | 0000:04:00.0     Off |                    0 |
| N/A   66C    P0  108W / 149W  |    240MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+
|   1  Tesla K80           On   | 0000:05:00.0     Off |                    0 |
| N/A   55C    P0  136W / 149W  |    260MiB / 11519MiB |     68%   E. Process |
+-------------------------------+----------------------+----------------------+
|   2  Tesla K80           On   | 0000:86:00.0     Off |                    0 |
| N/A   47C    P0  123W / 149W  |    252MiB / 11519MiB |     61%   E. Process |
+-------------------------------+----------------------+----------------------+
|   3  Tesla K80           On   | 0000:87:00.0     Off |                    0 |
| N/A   57C    P0  115W / 149W  |    242MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Compute processes:                                               GPU Memory |
|  GPU       PID  Process name                                     Usage      |
|=============================================================================|
|    0     11205  Nvidia-cuda-mps-server                          182MiB |
|    1     11211  Nvidia-cuda-mps-server                          202MiB |
|    2     11213  Nvidia-cuda-mps-server                          194MiB |
|    3     11212  Nvidia-cuda-mps-server                          184MiB |
+-----------------------------------------------------------------------------+
```

NVIDIA-SMI Basic Functionalities

- Nvidia tool for quick check
  - $> nvidia-smi
    - Driver
    - GPU model
    - Memory usage
    - GPU usage
    - GPU temperature
    - Power consumption
    - **Compute processes**

```
+-----------------------------------------------------------------------------+
| Nvidia-SMI 340.29     Driver Version: 340.29                                |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80           On   | 0000:04:00.0     Off |                    0 |
| N/A   66C    P0   108W / 149W |    240MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+
|   1  Tesla K80           On   | 0000:05:00.0     Off |                    0 |
| N/A   55C    P0   136W / 149W |    260MiB / 11519MiB |     68%   E. Process |
+-------------------------------+----------------------+----------------------+
|   2  Tesla K80           On   | 0000:86:00.0     Off |                    0 |
| N/A   47C    P0   123W / 149W |    252MiB / 11519MiB |     61%   E. Process |
+-------------------------------+----------------------+----------------------+
|   3  Tesla K80           On   | 0000:87:00.0     Off |                    0 |
| N/A   57C    P0   115W / 149W |    242MiB / 11519MiB |     40%   E. Process |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Compute processes:                                               GPU Memory |
|  GPU       PID  Process name                                     Usage      |
|=============================================================================|
|    0     11205  Nvidia-cuda-mps-server                          182MiB |
|    1     11211  Nvidia-cuda-mps-server                          202MiB |
|    2     11213  Nvidia-cuda-mps-server                          194MiB |
|    3     11212  Nvidia-cuda-mps-server                          184MiB |
+-----------------------------------------------------------------------------+
```

NVIDIA-SMI Basic Functionalities

- What is **Application Clock**?
  - Since the K20, it is possible to set the GPU frequency to a subset of supported clocks.
- Check:
  - $> nvidia-smi -q -d SUPPORTED_CLOCKS [-i X]
- Set:
  - $> nvidia-smi --reset-applications-clocks [-i X]
- Or
  - $> nvidia-smi --applications-clocks=2505,875 [-i X]

```
Supported Clocks
    Memory                          : 2505 MHz
        Graphics                    : 875 MHz
        Graphics                    : 862 MHz
        Graphics                    : 849 MHz
        Graphics                    : 836 MHz
        Graphics                    : 823 MHz
        Graphics                    : 810 MHz
        Graphics                    : 797 MHz
        Graphics                    : 784 MHz
        Graphics                    : 771 MHz
        Graphics                    : 758 MHz
        Graphics                    : 745 MHz
        Graphics                    : 732 MHz
        Graphics                    : 719 MHz
        Graphics                    : 705 MHz
        Graphics                    : 692 MHz
        Graphics                    : 679 MHz
        Graphics                    : 666 MHz
        Graphics                    : 653 MHz
        Graphics                    : 640 MHz
        Graphics                    : 627 MHz
        Graphics                    : 614 MHz
        Graphics                    : 601 MHz
        Graphics                    : 588 MHz
        Graphics                    : 575 MHz
        Graphics                    : 562 MHz
    Memory                          : 324 MHz
        Graphics                    : 324 MHz
```

NVIDIA-SMI Basic Functionalities

- What is **Auto-Boost** ?
  - When enabled, the GPU behaves as in a Turbo Mode:
    - frequency increases until power consumption hits the TDP, then frequency drops to default value, …

- Check:
  - $> nvidia-smi -q -d CLOCK [-i X]

- Set:
  - $> nvidia-smi --auto-boost-default=ENABLED [-i X]

```
Applications Clocks
    Graphics                          : 562 MHz
    Memory                            : 2505 MHz
Default Applications Clocks
    Graphics                          : 562 MHz
    Memory                            : 2505 MHz
Max Clocks
    Graphics                          : 875 MHz
    SM                                : 875 MHz
    Memory                            : 2505 MHz
```

```
Clock Policy
    Auto Boost                        : On
    Auto Boost Default                : On
```

NVIDIA-SMI Basic Functionalities

- What is **ECC Mode** ?
  - ECC errors are either single or double bit. Single bit errors are automatically corrected by the HW and do not result in data corruption. Double bit errors are detected but not corrected.
  - Volatile error counters track the number of errors detected since the last driver load.
  - Aggregate error counts persist indefinitely and thus act as a lifetime counter.

- Check:
  - $> nvidia-smi -q -d ECC [-i X]
- Set:
  - $> nvidia-smi --ecc-config=ENABLED/DISABLED [-i X]
- Reset is mandatory for ECC mode to take effect
  - $> nvidia-smi --gpu-reset -i X

NVIDIA-SMI Basic Functionalities

- What is **Persistence Mode** ?
  - When disabled, if the GPU is not used, the driver goes into idle mode.
  - This implies a “wake up” overhead each time the GPU is solicited and the driver loaded.
  - Also, some configuration commands reset each time the driver unloads…
- Check:
  - $> nvidia-smi --query-gpu=persistence_mode --format=csv [-i X]
- Set:
  - $> nvidia-smi --persistence-mode=ENABLED [-i X]

NVIDIA-SMI Basic Functionalities

- What is **Compute Mode** ?
  - It indicates whether individual or multiple compute applications may run on the same GPU

- Check:
  - $> nvidia-smi -q -d COMPUTE [–i X]
    - "Default" means multiple contexts are allowed per device.
    - "Exclusive Process" means only one context is allowed per device, usable from multiple threads at a time.
    - "Prohibited" means no contexts are allowed per device (no compute apps).
    - *"Exclusive Thread" means only one context is allowed per device, usable from one thread at a time. (deprecated)*

- Set:
  - $> nvidia-smi --compute-mode=DEFAULT [-i X]

Device Query and Bandwidth Test

- 2 useful tests provided in the CUDA sdk:
  - ${CUDA_HOME}/samples/1_Utilities
- Device Query provides information about the GPU
  - nb cores, CUDA driver version, CUDA capability, total amount of global memory, ...
- Bandwidth Test allows to test performance of data transfers
  - HtoD, DtoD, DtoH
- Performance issues?
  - first binaries to execute! (with nvidia-smi)

LAB: Device Query and Bandwidth Test

- Copy deviceQuery.cpp and bandwidthTest.cu from the nvidia samples
  - #cp ${CUDA_HOME}/samples/1_Utilities/deviceQuery/deviceQuery.cpp ./
  - #cp ${CUDA_HOME}/samples/1_Utilities/bandwidthTest/bandwidthTest.cu ./

- Compile both files:
  - #nvcc -I ${CUDA_HOME} ./deviceQuery.cpp -o ./deviceQuery.exe
  - #nvcc -I ${CUDA_HOME}/samples/common/inc ./bandwidthTest.cu -o ./bandwidthTest.exe

- Execute binaries:
  - #./deviceQuery.exe
  - #./bandwidthTest.exe

LAB: Device Query and Bandwidth Test

- What is **CUDA_VISIBLE_DEVICES** ?
  - On a multi-GPU node you can choose a subset of GPU visible devices

- How to check:
  - $> echo $CUDA_VISIBLE_DEVICES
- How to set:
  - $> export CUDA_VISIBLE_DEVICES=0,1,2,…,N

LAB: Device Query and Bandwidth Test

- What is **CUDA_VISIBLE_DEVICES** ?
  - On a multi-GPU node you can choose a subset of GPU visible devices

- How to check:
  - $> echo $CUDA_VISIBLE_DEVICES
- How to set:
  - $> export CUDA_VISIBLE_DEVICES=0,1,2,…,N

- Set CUDA_VISIBLE_DEVICES variable to 1. Execute binaries

- Set CUDA_VISIBLE_DEVICES variable to NAN. Execute binaries

- Unset CUDA_VISIBLE_DEVICES. Execute binaries

Copyright

Thanks

For more information please contact:

Georges-Emmanuel Moulard
georges-emmanuel.moulard@atos.net

Paul Karlshöfer
paul.Karlshoefer@atos.net

16-09-2019

Bull
atos technologies