

GPU driver/CUDA



Software



- **GPU Driver**
- **CUDA toolkit**
 - Includes all the software necessary for developers to write applications
 - Compiler (nvcc), Libraries, Profiler, Documentation
- **SDK**
 - Not strictly required but a good idea for ensuring your system is running properly.
 - Many examples with code samples illustrating lots of the important programming constructs and techniques.
- **www.nvidia.com/getcuda Above software from NVIDIA is free**

Examine GPU h/w and driver



- `nvidia-smi`
- **-h for help**
- **-q for long query of all GPUs**
 - PCIe Bus ID
 - Driver version
 - ECC state
 - Power state
 - Fans/Temps/Clocks speeds

nvidia-smi



```
jonathan.bentz@mcmillan-001:~  
[jonathan.bentz@mcmillan-001 ~]$ nvidia-smi  
Thu Dec 6 11:38:42 2012  
+-----+  
| NVIDIA-SMI 3.295.59    Driver Version: 295.59      |  
+-----+  
| Nb.   Name                | Bus Id          Disp. | Volatile ECC SB / DB |  
| Fan   Temp   Power Usage /Cap | Memory Usage    | GPU Util. Compute M. |  
+-----+  
| 0.   Tesla M2090          | 0000:08:00.0   Off  |      0                0 |  
| N/A   N/A    P0      77W / 225W | 0%    9MB / 5375MB | 0%    Default        |  
+-----+  
| 1.   Tesla M2090          | 0000:09:00.0   Off  |      0                0 |  
| N/A   N/A    P0      81W / 225W | 0%    9MB / 5375MB | 0%    Default        |  
+-----+  
| 2.   Tesla M2090          | 0000:0A:00.0   Off  |      0                0 |  
| N/A   N/A    P0      81W / 225W | 0%    9MB / 5375MB | 0%    Default        |  
+-----+  
| 3.   Tesla M2090          | 0000:19:00.0   Off  |      0                0 |  
| N/A   N/A    P0      79W / 225W | 0%    9MB / 5375MB | 0%    Default        |  
+-----+  
| 4.   Tesla M2090          | 0000:1A:00.0   Off  |      0                0 |  
| N/A   N/A    P0      79W / 225W | 0%    9MB / 5375MB | 0%    Default        |  
+-----+  
| 5.   Tesla M2090          | 0000:1B:00.0   Off  |      0                0 |
```

CUDA toolkit



- `module load cudatoolkit/4.2.9`
 - `/usr/local/cudatoolkit/4.2.9`
- **Compiler (nvcc)**
- **Libraries**
 - BLAS, FFT, sparse, RNG, NPP, OpenCL
- **Profiler**
 - Visual or command-line profiling available.

SDK (free download from [nvidia.com](https://www.nvidia.com))



- **Sample programs to illustrate CUDA and OpenGL programming constructs and algorithms.**
- **Useful diagnostic tests as well to query the GPU and its performance**

SDK bandwidthTest



```
jonathan.bentz@mcmillan-001:~/NVIDIA_GPU_Computing_SDK/C/bin/linux/release
[jonathan.bentz@mcmillan-001 release]$ ./bandwidthTest
[bandwidthTest] starting...

./bandwidthTest Starting...

Running on...

Device 0: Tesla M2090
Quick Mode

Host to Device Bandwidth, 1 Device(s), Paged memory
Transfer Size (Bytes)      Bandwidth(MB/s)
33554432                   3390.2

Device to Host Bandwidth, 1 Device(s), Paged memory
Transfer Size (Bytes)      Bandwidth(MB/s)
33554432                   2983.5

Device to Device Bandwidth, 1 Device(s)
Transfer Size (Bytes)      Bandwidth(MB/s)
33554432                   120531.3

[bandwidthTest] test results...
PASSED

> exiting in 3 seconds: 3...2...1...done!

[jonathan.bentz@mcmillan-001 release]$
```

bandwidthTest --memory=pinned



```
jonathan.bentz@mcmillan-001:~/NVIDIA_GPU_Computing_SDK/C/bin/linux/release
[jonathan.bentz@mcmillan-001 release]$ ./bandwidthTest --memory=pinned
[bandwidthTest] starting...

./bandwidthTest Starting...

Running on...

Device 0: Tesla M2090
Quick Mode

Host to Device Bandwidth, 1 Device(s), Pinned memory
Transfer Size (Bytes)      Bandwidth (MB/s)
33554432                   5700.6

Device to Host Bandwidth, 1 Device(s), Pinned memory
Transfer Size (Bytes)      Bandwidth (MB/s)
33554432                   6268.8

Device to Device Bandwidth, 1 Device(s)
Transfer Size (Bytes)      Bandwidth (MB/s)
33554432                   120701.5

[bandwidthTest] test results...
PASSED

> exiting in 3 seconds: 3...2...1...done!

[jonathan.bentz@mcmillan-001 release]$
```


SDK deviceQuery



```
jonathan.bentz@mcmillan-001:~/NVIDIA_GPU_Computing_SDK/C/bin/linux/release
Device 7: "Tesla M2090"
  CUDA Driver Version / Runtime Version          4.2 / 4.2
  CUDA Capability Major/Minor version number:    2.0
  Total amount of global memory:                 5375 MBytes (5636554752 bytes)
  (16) Multiprocessors x ( 32) CUDA Cores/MP:   512 CUDA Cores
  GPU Clock rate:                               1301 MHz (1.30 GHz)
  Memory Clock rate:                            1848 Mhz
  Memory Bus Width:                             384-bit
  L2 Cache Size:                                786432 bytes
  Max Texture Dimension Size (x,y,z)            1D=(65536), 2D=(65536,65535), 3
D=(2048,2048,2048)
  Max Layered Texture Size (dim) x layers        1D=(16384) x 2048, 2D=(16384,16
384) x 2048
  Total amount of constant memory:               65536 bytes
  Total amount of shared memory per block:       49152 bytes
  Total number of registers available per block: 32768
  Warp size:                                     32
  Maximum number of threads per multiprocessor: 1536
  Maximum number of threads per block:           1024
  Maximum sizes of each dimension of a block:    1024 x 1024 x 64
  Maximum sizes of each dimension of a grid:     65535 x 65535 x 65535
  Maximum memory pitch:                          2147483647 bytes
  Texture alignment:                             512 bytes
  Concurrent copy and execution:                 Yes with 2 copy engine(s)
  Run time limit on kernels:                     No
  Integrated GPU sharing Host Memory:            No
  Support host page-locked memory mapping:       Yes
  Concurrent kernel execution:                   Yes
  Alignment requirement for Surfaces:            Yes
  Device has ECC support enabled:                 Yes
  Device is using TCC driver mode:               No
  Device supports Unified Addressing (UVA):      Yes
  Device PCI Bus ID / PCI location ID:           22 / 0
```

matrixMul



```
jonathan.bentz@mcmillan-001:~/NVIDIA_GPU_Computing_SDK/C/bin/linux/release
[jonathan.bentz@mcmillan-001 release]$ ./matrixMul
[matrixMul] starting...

[ matrixMul ]
./matrixMul
    Starting (CUDA and CUBLAS tests)...

Device 0: "Tesla M2090" with Compute 2.0 capability

Using Matrix Sizes: A(640 x 960), B(640 x 640), C(640 x 960)

Runing Kernels...

> CUBLAS          758.4941 GFlop/s, Time = 0.00104 s, Size = 786432000 Ops

> CUDA matrixMul 229.0823 GFlop/s, Time = 0.00343 s, Size = 786432000 Ops, NumDe
vsUsed = 1, Workgroup = 1024

Comparing GPU results with Host computation...

Comparing CUBLAS & Host results
CUBLAS compares OK

Comparing CUDA matrixMul & Host results
CUDA matrixMul compares OK

[matrixMul] test results...
PASSED

> exiting in 3 seconds: 3...2...1...done!

[jonathan.bentz@mcmillan-001 release]$
```

Recap



- **Driver**
 - **nvidia-smi to query the GPU hardware and state**
- **Toolkit**
 - **Development tools for GPU programming**
- **SDK**
 - **Sample code as well as diagnostic tests**