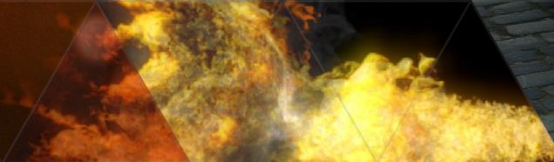
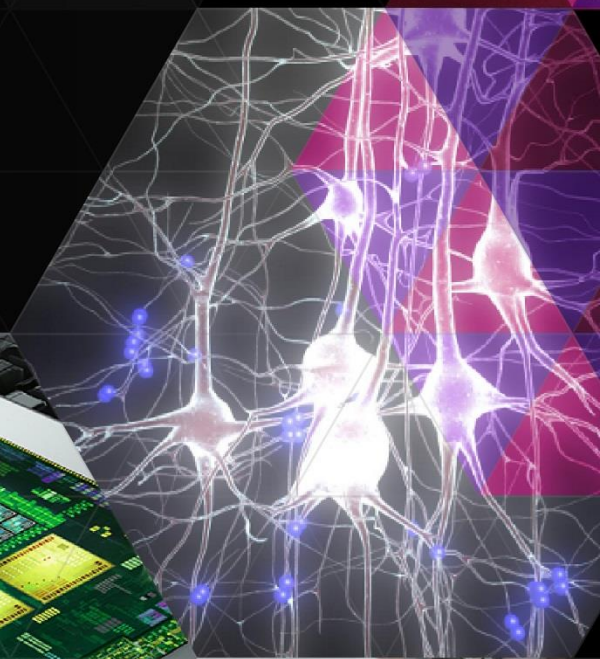
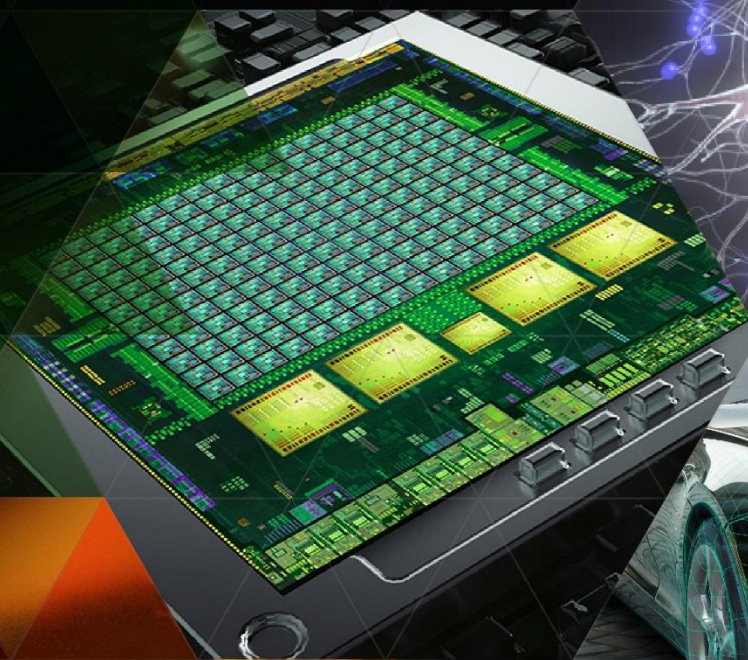




DRIVER/TOOLKIT/SAMPLES



ACCESS TO DEEPTHOUGHT2

- ▶ `ssh <username>@login.deepthought2.umd.edu`
- ▶ `cp -r ~gpu-k1c/maryland .`
- ▶ Use the modules environment system
 - ▶ `module load <ModuleName>` to load a module
 - ▶ `module list` to show which modules are loaded
- ▶ Use SLURM batch system. Every job submitted to the queue.
 - ▶ `sbatch`—submit a job to the queue
 - ▶ `squeue`—show all jobs in the queue
 - ▶ `qdel`—delete a job from the queue

SOFTWARE

- ▶ GPU Driver
- ▶ CUDA toolkit
 - ▶ Includes all the software necessary for developers to write applications
 - ▶ Compiler (nvcc), Libraries, Profiler, Documentation
- ▶ CUDA Samples
 - ▶ 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/Clockspeed
- ▶ `sbatch runit.nvidia-smi`
 - ▶ Open the resulting `slurm-*.out` file

nvidia-smi

```
login.deepthought2.umd.edu - PuTTY
Thu Oct 23 14:40:08 2014
-----+
| NVIDIA-SMI 340.32      Driver Version: 340.32      |
|-----+-----+-----+-----+-----+-----+-----+
| 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 K20m                Off   | 0000:03:00.0    Off  |           0         |
| N/A   25C    P0      48W / 225W | 11MiB / 4799MiB |      0%      Default  |
|-----+-----+-----+-----+-----+-----+
|   1   Tesla K20m                Off   | 0000:83:00.0    Off  |           0         |
| N/A   24C    P0      51W / 225W | 11MiB / 4799MiB |     75%      Default  |
|-----+-----+-----+-----+-----+-----+
|
| Compute processes:                                     GPU Memory |
| GPU      PID  Process name                                 Usage      |
|====+=====+=====+=====+=====+=====+=====+
|
| No running compute processes found
|
~
~
~
"slurm-3323285.out" 20L, 1552C
```

CUDA TOOLKIT

- ▶ `module load cuda/6.5.14`
- ▶ Compiler (nvcc)
- ▶ Libraries
 - ▶ BLAS, FFT, sparse, RNG, NPP, OpenCL
- ▶ Profiler
 - ▶ Visual or command-line profiling available.

SAMPLES (COMES WITH CUDA TOOLKIT)

- ▶ `~gpu-k1c/CUDA65Samples/`
- ▶ Sample programs to illustrate CUDA and OpenGL programming constructs and algorithms.
- ▶ Useful diagnostic tests to query the GPU and its performance

sbatch runit.bandwidth

```
login.deepthought2.umd.edu - PuTTY
[ CUDA Bandwidth Test ] - Starting...
Running on...

Device 0: Tesla K20m
Quick Mode

Host to Device Bandwidth, 1 Device(s)
PINNED Memory Transfers
  Transfer Size (Bytes)      Bandwidth(MB/s)
  33554432                   6165.3

Device to Host Bandwidth, 1 Device(s)
PINNED Memory Transfers
  Transfer Size (Bytes)      Bandwidth(MB/s)
  33554432                   6547.8

Device to Device Bandwidth, 1 Device(s)
PINNED Memory Transfers
  Transfer Size (Bytes)      Bandwidth(MB/s)
  33554432                   147309.5

Result = PASS
~
"slurm-3324212.out" 22L, 485C
```


sbatch runit.query ~gpu-k1c

```
login.deepthought2.umd.edu - PuTTY
Device 0: "Tesla K20m"
  CUDA Driver Version / Runtime Version      6.5 / 6.5
  CUDA Capability Major/Minor version number: 3.5
  Total amount of global memory:             4800 MBytes (5032706048 bytes)
  (13) Multiprocessors, (192) CUDA Cores/MP: 2496 CUDA Cores
  GPU Clock rate:                            706 MHz (0.71 GHz)
  Memory Clock rate:                         2600 Mhz
  Memory Bus Width:                          320-bit
  L2 Cache Size:                             1310720 bytes
  Maximum Texture Dimension Size (x,y,z)     1D=(65536), 2D=(65536, 65536),
  3D=(4096, 4096, 4096)
  Maximum Layered 1D Texture Size, (num) layers 1D=(16384), 2048 layers
  Maximum Layered 2D Texture Size, (num) layers 2D=(16384, 16384), 2048 layers
  Total amount of constant memory:           65536 bytes
  Total amount of shared memory per block:    49152 bytes
  Total number of registers available per block: 65536
  Warp size:                                 32
  Maximum number of threads per multiprocessor: 2048
  Maximum number of threads per block:       1024
  Max dimension size of a thread block (x,y,z): (1024, 1024, 64)
  Max dimension size of a grid size (x,y,z): (2147483647, 65535, 65535)
  Maximum memory pitch:                      2147483647 bytes
  Texture alignment:                          512 bytes
```

`sbatch runit.matmul ~gpu-k1c`

```
login.deepthought2.umd.edu - PuTTY
[Matrix Multiply Using CUDA] - Starting...
GPU Device 0: "Tesla K20m" with compute capability 3.5

MatrixA(320,320), MatrixB(640,320)
Computing result using CUDA Kernel...
done
Performance= 247.93 GFlop/s, Time= 0.529 msec, Size= 131072000 Ops, WorkgroupSize= 1024 threads/block
Checking computed result for correctness: Result = PASS

Note: For peak performance, please refer to the matrixMulCUBLAS example.
[Matrix Multiply CUBLAS] - Starting...
GPU Device 0: "Tesla K20m" with compute capability 3.5

MatrixA(320,640), MatrixB(320,640), MatrixC(320,640)
Computing result using CUBLAS...done.
Performance= 1159.98 GFlop/s, Time= 0.113 msec, Size= 131072000 Ops
Computing result using host CPU...done.
Comparing CUBLAS Matrix Multiply with CPU results: PASS
~
~
~
~
"slurm-3324218.out" 18L, 759C
```

RECAP

- ▶ Driver
 - ▶ `nvidia-smi` to query the GPU hardware and state
- ▶ CUDA Toolkit
 - ▶ Development tools for GPU programming
- ▶ CUDA Samples
 - ▶ Sample code as well as diagnostic tests