

# CUDA KERNEL PROFILING USING NVIDIA NSIGHT COMPUTE

Sanjiv Satoor, Magnus Strengert

## AGENDA

Overview of NVIDIA's Devtools What is New for CUDA Tools in CUDA 10.1 Working with NVIDIA Nsight Compute (Demo)

## **TOOLS OFFERINGS**

### IDE

• Nsight Eclipse Edition (EE) (plugin)



Nsight Visual Studio Edition (VSE) (plugin)

### Debug

- Nsight
- CUDA-GDB
- CUDA Debug API

### Memcheck

- CUDA-memcheck
- Nsight Visual Studio Edition built-in



### Profile

- Nsight Systems
- Nsight Compute
- CUDA Visual Profiler
- nvprof
- Nsight Visual Studio Edition
- CUPTI

NVTX (NVidia Tools eXtension)



## **CUDA PROFILING TOOLS**

Updates for CUDA 10.1

General:

• Support for latest Turing GPUs

Visual Profiler/nvprof/CUPTI:

Support for NVTX string registration API nvtxDomainRegisterStringA().

Nsight Visual Studio Edition:

• Nsight Compute improvements

# NSIGHT COMPUTE

Updates for CUDA 10.1

Updates:

- Lower performance overhead and reduced memory overhead
- Source Page reports metrics by functions or files
- Updates sections and rules; section descriptions

Improved Parity to NVPROF:

- Profiling of child processes (MPI, ...)
- CLI options: --summary, --quite, ...

Extended NVTX Support:

- Trigger profiling by NVTX ranges
- Print NVTX state in CLI

| ≣ # Ad         | dress Source  | Sampling Data (Not Issued)                | Instructions Executed        | Predicated-On Thread Instructions Executed        |
|----------------|---|---|------------------------------|---|
| > 1            | _Z11getLocation3Rayi  | 155                                       | 70,738                       | 2,259,774   |
| 2              | _Zmi6float2S_   | 54  | 31,926                       | 1,019,898   |
| 3              | _Z3dot6float2S_   | 65  | 28,170                       | 899 <mark>, 910</mark>                            |
| 4              | atanf   | 51  | 35,089                       | 1,119,919   |
| 5              | sqrtf   | 37  | 13,145                       | 419,958   |
| 6              | _Z20computeAngles_kernel3RayPf                                    | 612                                       | 244,213                      | 7,804,773   |
| 7              | _Z8getAngle3Ray6float2f   | 137                                       | 84,510                       | 2,699,730   |
| 8              | _Zml6float2f  | 57  | 23,475                       | 749,925   |
| 9              | _Z6length6float2  | 54  | 26,918                       | 859,914   |
| 10             | _ZN80_INTERNAL_58_tmpxft_0000                                     | 192                                       | 92,335                       | 2,949,705   |
| 11             | _Zpl6float2S_   | 70  | 31,926                       | 1,019,898   |
| 12             | cuda_sm3x_div_rn_noftz_f32  | 43  | 13,772                       | 439,956   |
| Scheduler      | Statistics A  |   |                              | Q   |
| Summary of the | activity of the schedulers issuing instructions. Each scheduler m | aintains a nool of warns that it can issu | e instructions for The upper | bound of warps in the pool (Theoretical Warps) is |

Jummary of the activity of the schedulers issuing instructions. Each scheduler maintains a pool of warps that it can issue instructions for. The upper bound of warps in the pool (Active Warps), Active warps that are not stalled (Eligible Warps) are ready to issue heir next instruction. From the set of eligible warps the scheduler selects a single warp from which to issue one or more instructions (Issued Warp). On cycles with no eligible warps, the issue slot is kipped and no instruction is issued. Having many skipped issue slots indicates poor latency hiding.

| Issued Warp Per Scheduler 0.2           | One or More Eligible [%] 25.99                      |
|---|---|
|   |   |
| Eligible Warps Per Scheduler [warp] 2.8 | No Eligible [%] 74.01                               |
| Active Warps Per Scheduler [warp] 12.2  | 7 Instructions Per Active Issue Slot [inst/cycle] 1 |

## **NSIGHT PRODUCT FAMILY**

### **Standalone Performance Tools**

Nsight Systems - System-wide application algorithm tuning

**Nsight Compute** - Debug/optimize specific CUDA kernel

**Nsight Graphics -** Debug/optimize specific graphics shader

### **IDE Plugins**

Nsight Eclipse Edition/Visual Studio - editor, debugger, some perf analysis





System-wide application algorithm tuning Multi-process tree support

Locate optimization opportunities Visualize millions of events on a very fast GUI timeline Or gaps of unused CPU and GPU time

Balance your workload across multiple CPUs and GPUs CPU algorithms, utilization, and thread state GPU streams, kernels, memory transfers, etc

#### WIDIA System Profiler 4.0

<u>File View H</u>elp

Select device for profiling... 🔹 🏠 More info...

Project 2 🔝 DGKV8-Im-GFU.gdrep 🔝 trace\_DGK11\_TF\_synthetic\_ResNet50-with-trace-backtraces.gdrep 🔝 trace\_DGK11\_TF\_synthetic\_ResNet50-with-trace-backtraces.gdrep 🔝

| Timeline View *   |   |   |  |  |   |                                     |                  |
|---|---|---|--|--|---|-------------------------------------|------------------|
| 🗻 🖂 Γτολί βλαιου 🔸  | 5s +850ms +900ms  | +950ms 6s   | +50ms +100ms   | +150ms +200ms  | +250ms  | +300ms +350ms                       | +400ms           |
| System<br>CUDA API  |   |   | Thread/core  |  | <u>1) (111-1111)</u> , (111) , (11) , (1) , (1)   |                                     |                  |
| cuDNN   |   | a second according to   | Thead/Core   | and the state of t | a a carlacta  | a cha di                            |                  |
| Profiler overhead   |   |   | migration  |  |   |                                     |                  |
| ✓ 🗹 [178] python →  |   |   |  |  | والمستعدية الكلا المراكل المراجع المتعادية والمحاد  |                                     |                  |
| System D  | rocesses and  | nthr  |  |  |   |                                     | ()               |
| cuDNN   |   |   | And the second second  | and the second second  |   |                                     |                  |
| cuBLAS<br>Profiler overhead   | threads   |   |  |  | Thread st   | ate                                 |                  |
| ∽ 🗹 [165] python →  | Amar, A   |   | . الالبالي الشكلية بينانية بيناني الشكلية  |  |   |                                     |                  |
| System  |   |   |  |  |   |                                     |                  |
| cuDNN   |   | and a second second and a second s |  |  |   |                                     |                  |
| cuBLAS<br>Profiler overhead   | CUDA and OpenGL   |   |  |  |   |                                     |                  |
| ∽ 🗹 [166] python →  | A DI traco  |   | ينور الأركاني اراد الجرية والمتقار   |  | کر کر پر کار ، ایک کر ا   |                                     |                  |
| System  | APILIACE  | - )   |  | 100  | .0115 ; 16 (011 ). 11 ). 5  |                                     | ), _ pth, p p.   |
| cuDNN   | х і і лі  |   |  | n ha mulanti ha kultu ni ana a a dan kuma.<br>A sa ƙasar a ƙasar a   |   |                                     |                  |
| cuBLAS  |   |   |  |  |   |                                     |                  |
| ✓ [159] python +  | cuDNN and   |   |  |  |   |                                     |                  |
| CUDA API  | CUDININ ALIU  |   | s s <b>,01</b> , s s <b>,01</b> , s s <b>,01</b> , s s <b>,01</b> , s s  | MILLER M   | III - J. 010, J <b>J.0</b> , 063. [   |                                     |                  |
| cuBLAS  | cuBLAS trace  |   | 1 a 1 1 a 1  | ala i calcura  | at the second second  |                                     | 11               |
| 33 threads hidden   |   |   |  |  |   |                                     |                  |
| ✓ Stream 174  | and the second  |   | and an an exception of the state of the stat |  |   |                                     | - A - A          |
| ✓ Kernels   | mio della in della in   | an a  | ւթյուն արտարին անվանգանին անվանություն   | e in min in min inmin the min initia   | io bran in ina. In inan in inim   | nijo ola in joli in jojoh           |                  |
| > maxwell_fp16_scudnn_fp16_128x128_stridedB_sp<br>> maxwell_fp16_scudnn_fp16_128x128_relu_interio                         |   | el and memory   |  |  |   |                                     |                  |
| > maxwell_fp16_scudnn_fp16_128x128_stridedB_int > dgrad_engine  | terior_nn   | ict and memory  |  |  |   |                                     |                  |
| > cudnn_maxwell_gcgemm_64x64_tn_batched   | tran  | nsfer activities 🛛 😐  | 11 1   |  | l i   | I                                   |                  |
| 28 kernel group(s) hidden<br><b>V Stream 12</b>   | Law Manager ( Manager )   | · · · · · · · · · · · · · · · · · · ·   | and the second   | ennistrative large in tall, terms daried et e  | الدفيتيات بالتدعي فالتقادية فافتعا  | Constant and Constant a             | Concentrations.  |
| ✓ Kernels > AllReduceKernelSmall  |   | · · · · · · · · · · · · · · · · · · ·   | a suma and a state of a state of the state o | talma literatura dallar dalar dalar dalar  | and a state of the second |                                     |                  |
| > AllReduceKernel   |   |   |  |  |   |                                     |                  |
| 1 kernel group(s) hidden<br>67 stream(s) hidden   |   |   |  |  |   |                                     |                  |
| <ul> <li>CUDA (Tesla P100-SXM2-16GB)</li> <li>Stream 173</li> </ul>   |   |   | nan kana kana kana kana kana kana kana   |  |   |                                     |                  |
| > Stream 20   | Multi-GPU   | •   | and a second   | na har bar atau an de la cara an   | an dia minimi ny mini na mini ha  | and the design of the second second | STATE OF COMPLEX |
| > CUDA (Tesla P100-SXM2-16GB)   |   |   |  |  |   |                                     |                  |
| <ul> <li>CUDA (Tesia P100-SXM2-1008)</li> <li>CUDA (Tesia P100-SXM2-16GB)</li> <li>CUDA (Tesia P100-SXM2-16GB)</li> </ul> |   |   |  |  |   |                                     |                  |
| <ul> <li>CUDA (Tesla P100-SXM2-16GB)</li> <li>CUDA (Tesla P100-SXM2-16GB)</li> </ul>                                      |   |   |  |  |   |                                     |                  |
| > CUDA (Tesla P100-SXM2-16GB)   | Contraction of the second s |   |  |  |   |                                     |                  |

## **TRANSITIONING TO PROFILE A KERNEL**

Dive into kernel analysis



📀 NVIDIA.

# NSIGHT COMPUTE INTRODUCTION



### **NVIDIA NSIGHT COMPUTE** Next-Gen Kernel Profiling Tool

#### Key Features:

- Interactive CUDA API debugging and kernel profiling •
- Graphical profile report •
- Comparison of multiple kernel reports •
- Fully Customizable (Reports and Analysis Rules) .
- Command Line, Standalone, IDE Integration •

OS: Linux, Windows, ARM, MacOSX (host only) GPUs: Pascal (GP10x), Volta, Turing

Docs/product: https://developer.nvidia.com/nsight-compute

| ▼ GPU Speed 0 | Light |  |      |           |
|---------------|-------|--|------|-----------|
| SOL SM [%]    |       |  | 8.74 | (-80.00%) |
| SOL TEX [%]   |       |  | 8.74 | (-2.82%)  |
| SOL L2 [%]    |       |  | 0.05 | (-99.35%) |
| SOL FB [%]    |       |  | 0.09 | (-99.76%) |
|               |       |  |      | GPU Ut    |
| SM [%]        |       |  |      |           |
| Memory [%]    |       |  |      |           |

0.0

nst, executed lins

|                   | 10.0 | 20.0 | 30.0       |               | 40.0       |             | 20   |
|-------------------|------|------|------------|---------------|------------|-------------|------|
|                   |      |      |            |               |            | Speed 0     | )f I |
|                   |      |      |            |               |            | Recon       | nn   |
|                   |      |      |            |               |            |             | _    |
|                   |      |      | 16,528.00; | 16,528.00; _  | 13,476.00; | 13,476.00;  | -    |
|                   |      |      |            | 14.3          | 3          | r           | ı/a  |
|                   |      |      |            | 128.0         | 0          | 128.        | 66   |
|                   |      |      | 47,        | 611,587,968.0 | 0          | 12,273,728. | .00  |
|                   |      |      |            | 4,132.0       | 0          | 3,369.      | .00  |
| cks [block]       |      |      |            | 32.0          | 0          | 32.         | 00   |
| isters [register] | 1    |      |            | 21.0          | 0          | 21.         | 00   |
| ared mem [byte    | es]  |      |            | 384.0         | 0          | 384.        | ee   |
|                   | •    |      | -          |               |            |             |      |

| tex_sol_pct [%]                                | 14.33                  | n/a                    |
|--|------------------------|------------------------|
| unchblock_size                                 | 128.00                 | 128.00                 |
| unchfunction_pcs                               | 47,611,587,968.00      | 12,273,728.00          |
| unchgrid_size                                  | 4,132.00               | 3,369.00               |
| unchoccupancy_limit_blocks [block]             | 32.00                  | 32.00                  |
| unchoccupancy_limit_registers [register]       | 21.00                  | 21.00                  |
| unchoccupancy_limit_shared_mem [bytes]         | 384.00                 | 384.00                 |
| unchoccupancy_limit_warps [warps]              | 16.00                  | 16.00                  |
| unchoccupancy_per_block_size                   | 3,638.00               | 3,638.00               |
| unchoccupancy_per_register_count               | 5,792.00               | 5,792.00               |
| unchoccupancy_per_shared_mem_size              | 2,260.00               | 2,260.00               |
| unchregisters_per_thread [register/thread]     | 17.00                  | 17.00                  |
| unchshared_mem_config_size [bytes]             | 49,152.00              | 49,152.00              |
| unchshared_mem_per_block_dynamic [bytes/block] | 0.00                   | 0.00                   |
| unchshared_mem_per_block_static [bytes/block]  | 20.00                  | 20.00                  |
| unchthread_count [thread]                      | 528,896.00             | 431,232.00             |
| unchwaves_per_multiprocessor                   | 3.23                   | 42.11                  |
| :sol_pct [%]                                   | 6.93                   | 7.18                   |
| emory_access_size_type [bytes]                 | 2.00; 32.00; 32.00; 32 | 2.00; 32.00; 32.00; 32 |
|  | 2 00. 4 00. 2 00. 2 00 | 2 00. 4 00. 2 00. 2 00 |

| Source         Live Registers         Sampling Data (AII)         Sampling Data (No Issue)           @!PT SHFL.IDX PT, RZ, RZ, RZ, RZ;         0         223         0           MOV R1, c[0x0][0x28];         1         13         44           S2R R0, SR_CTAID.X;         2         143         75           S2R R2, SR_TID.X;         3         0         38           IMAD R0, R0, c[0x0][0x0], R2;         3         599         94           ISETP.GE.AND P0, PT, R0, c[0x0][0x170]         2         125         26           @P0         EXIT;         2         259         86           MOV R2, R0;         3         386         259           @P1 SHFL.IDX PT, RZ, RZ, RZ, RZ;         2         0         6           MOV R4, 0x4;         3         0         0         0           IMAD.WIDE R4, R2, R4, c[0x0][0x160];         4         0         0         0           IMAD.WIDE R4, R2, R4, c[0x0][0x160];         4         0         0         0           BSSY B0, 0xb00976780;         3         0         0         0         0           BSSY B0, 0xb00976780;         3         0         0         0         0         0           SHF.R.S32.HI R0, RZ, 0x1f, R |      |  |                |                     |               |            |
|---|------|--|----------------|---------------------|---------------|------------|
| @!PT SHFL.IDX PT, RZ, RZ, RZ;       0       223       0         MOV R1, c[0x0][0x28];       1       13       44         S2R R0, SR_CTAID.X;       2       143       75         S2R R2, SR_TID.X;       3       0       38         IMAD R0, R0, c[0x0][0x0], R2;       3       599       94         ISETP.GE.AND P0, PT, R0, c[0x0][0x170]       2       125       26         @P0 EXIT;       2       259       86         MOV R2, R0;       3       386       25         @P1 SHFL.IDX PT, RZ, RZ, RZ, RZ;       2       0       6         MOV R4, 0x4;       3       0       6         IMAD.WIDE R4, R2, R4, c[0x0][0x160];       4       0       6         LDG.E.SYS R3, [R4];       3       0       6         BSSY B0, 0xb00976780;       3       0       6         SHF.R.S32.HI R0, RZ, 0x1f, R2;       4       0       6  |      | Source                                 | Live Registers | Sampling Data (All) | Sampling Data | (No Issue) |
| MOV R1, c[@x0][@x28];       1       13       44         S2R R0, SR_CTAID.X;       2       143       75         S2R R2, SR_TID.X;       3       0       38         IMAD R0, R0, c[@x0][@x0], R2;       3       599       94         ISETP.GE.AND P0, PT, R0, c[@x0][@x170]       2       125       26         @P0 EXIT;       2       259       86         MOV R2, R0;       3       386       25         @P1 SHFL.IDX PT, RZ, RZ, RZ;       2       0       6         MOV R4, 0x4;       3       0       6         IMAD.WIDE R4, R2, R4, c[@x0][@x160];       4       0       6         LDG.E.SYS R3, [R4];       3       0       6         BSSY B0, @xb00976780;       3       0       6         SHF.R.S32.HI R0, RZ, @x1f, R2;       4       0       6  | @!PT | SHFL.IDX PT, RZ, RZ, RZ, RZ;           | 0              | 223                 |               | e          |
| S2R R0, SR_CTAID.X;       2       143       75         S2R R2, SR_TID.X;       3       0       38         IMAD R0, R0, c[0x0][0x0], R2;       3       599       94         ISETP.GE.AND P0, PT, R0, c[0x0][0x170]       2       125       26         @P0 EXIT;       2       259       86       25         MOV R2, R0;       3       386       25         @P1 SHFL.IDX PT, RZ, RZ, RZ;       2       0       6         MOV R4, 0x4;       3       0       6         IMAD.WIDE R4, R2, R4, c[0x0][0x160];       4       0       6         LDG.E.SYS R3, [R4];       3       0       6         BSSY B0, 0xb00976780;       3       0       6         SHF.R.S32.HI R0, RZ, 0x1f, R2;       4       0       6   |      | MOV R1, c[0x0][0x28];                  | 1              | 13                  |               | 44         |
| S2R R2, SR_TID.X;       3       0       38         IMAD R0, R0, c[@x0][@x0], R2;       3       599       94         ISETP.GE.AND P0, PT, R0, c[@x0][@x170]       2       125       26         @P0 EXIT;       2       259       86       25         MOV R2, R0;       3       386       25         @!PT SHFL.IDX PT, RZ, RZ, RZ;       2       0       6         MOV R4, 0x4;       3       0       6         IMAD.WIDE R4, R2, R4, c[@x0][@x160];       4       0       6         LDG.E.SYS R3, [R4];       3       0       6         BSSY B0, @xb00976780;       3       0       6         SHF.R.S32.HI R0, RZ, @x1f, R2;       4       0       6   |      | S2R RØ, SR_CTAID.X;                    | 2              | 143                 |               | 75         |
| IMAD R0, R0, c[0x0][0x0], R2;       3       599       94         ISETP.GE.AND P0, PT, R0, c[0x0][0x170]       2       125       26         @P0 EXIT;       2       259       86         MOV R2, R0;       3       386       25         @P1 SHFL.IDX PT, RZ, RZ, RZ;       2       0       6         MOV R4, 0x4;       3       0       6         IMAD.WIDE R4, R2, R4, c[0x0][0x160];       4       0       6         LDG.E.SYS R3, [R4];       3       0       6         BSSY B0, 0xb00976780;       3       0       6         SHF.R.S32.HI R0, RZ, 0x1f, R2;       4       0       6  |      | S2R R2, SR_TID.X;                      | 3              | 0                   |               | 38         |
| ISETP.GE.AND P0, PT, R0, c[0x0][0x170]       2       125       26         @P0 EXIT;       2       259       86         MOV R2, R0;       3       386       25         @P1 SHFL.IDX PT, RZ, RZ, RZ, RZ;       2       0       6         MOV R4, 0x4;       3       0       6         IMAD.WIDE R4, R2, R4, c[0x0][0x160];       4       0       6         LDG.E.SYS R3, [R4];       3       0       6         BSSY B0, 0xb00976780;       3       0       6         SHF.R.S32.HI R0, RZ, 0x1f, R2;       4       0       6   |      | IMAD R0, R0, c[0x0][0x0], R2;          | 3              | 599                 |               | 94         |
| @P0 EXIT;       2       259       86         MOV R2, R0;       3       386       25         @!PT SHFL.IDX PT, RZ, RZ, RZ;       2       0       6         MOV R4, 0x4;       3       0       6         IMAD.WIDE R4, R2, R4, c[0x0][0x160];       4       0       6         LDG.E.SYS R3, [R4];       3       0       6         BSSY B0, 0xb00976780;       3       0       6         SHF.R.S32.HI R0, RZ, 0x1f, R2;       4       0       6  |      | ISETP.GE.AND P0, PT, R0, c[0x0][0x170] | 2              | 125                 |               | 26         |
| MOV R2, R0;     3     386     25       @!PT SHFL.IDX PT, RZ, RZ, RZ, RZ;     2     0     6       MOV R4, 0x4;     3     0     6       IMAD.WIDE R4, R2, R4, c[0x0][0x160];     4     0     6       LDG.E.SYS R3, [R4];     3     0     6       BSSY B0, 0xb00976780;     3     0     6       SHF.R.S32.HI R0, RZ, 0x1f, R2;     4     0     6   | @PØ  | EXIT;                                  | 2              | 259                 |               | 86         |
| @!PT SHFL.IDX PT, RZ, RZ, RZ, RZ;       2       0       0         MOV R4, 0x4;       3       0       0         IMAD.WIDE R4, R2, R4, c[0x0][0x160];       4       0       0         LDG.E.SYS R3, [R4];       3       0       0         BSSY B0, 0xb00976780;       3       0       0         SHF.R.S32.HI R0, RZ, 0x1f, R2;       4       0       0  |      | MOV R2, R0;                            | 3              | 386                 |               | 29         |
| MOV R4, 0x4;     3     0     0       IMAD.WIDE R4, R2, R4, c[0x0][0x160];     4     0     0       LDG.E.SYS R3, [R4];     3     0     0       BSSY B0, 0xb00976780;     3     0     0       SHF.R.S32.HI R0, RZ, 0x1f, R2;     4     0     0  | @!PT | SHFL.IDX PT, RZ, RZ, RZ, RZ;           | 2              | 0                   |               | e          |
| IMAD.WIDE R4, R2, R4, c[0x0][0x160];       4       0       6         LDG.E.SYS R3, [R4];       3       0       6         BSSY B0, 0xb00976780;       3       0       6         SHF.R.S32.HI R0, RZ, 0x1f, R2;       4       0       6   |      | MOV R4, 0x4;                           | 3              | 0                   |               | e          |
| LDG.E.SYS R3, [R4]; 3 0 6<br>BSSY B0, 0xb00976780; 3 0 6<br>SHF.R.S32.HI R0, RZ, 0x1f, R2; 4 0 6  |      | IMAD.WIDE R4, R2, R4, c[0x0][0x160];   | 4              | 0                   |               | e          |
| BSSY B0, 0xb00976780; 3 0 0<br>SHF.R.S32.HI R0, RZ, 0x1f, R2; 4 0 0   |      | LDG.E.SYS R3, [R4];                    | 3              | 0                   |               | e          |
| SHF.R.S32.HI R0, RZ, 0x1f, R2; 0 0  |      | BSSY B0, 0xb00976780;                  | 3              | 0                   |               | e          |
|   |      | SHF.R.S32.HI R0, RZ, 0x1f, R2;         | 4              | 0                   |               | e          |

# NSIGHT COMPUTE DEMO

### Volta Architecture

|  |  |  |   |  |   |                                   | L1 Instruc     | tion Cache   |  |  |  |  |                                       |  |       |
|--|--|--|---|--|---|-----------------------------------|----------------|--|--|--|--|--|---------------------------------------|--|-------|
|  |  | L0 Ir  | nstruc  | tion C   | ache  |                                   |                |  |  | L0 Ir  | nstruc   | tion C   | ache                                  |  |       |
|  | War  | p Sch  | edule   | r (32 t  | hread/  | clk)                              |                |  | Wai  | rp Sch   | nedule   | r (32 t  | hread                                 | /clk)  |       |
|  | Dis  | patch  | h Unit  | (32 th   | read/c  | lk)                               |                |  | Di   | spatcl   |  | (32 th   | read/                                 | clk)   |       |
|  | Regi   | ister  | File (1   | 16,384   | 4 x 32·                                       | -bit)                             |                |  | Reg  |  | File ('  | 16,384   | 4 x 32                                | 2-bit)   |       |
| FP64   | INT  | INT  | FP32  | FP32   |   |                                   |                | FP64   |  |  | FP32   | FP32   |                                       |  |       |
| FP64   | INT  | INT  | FP32  | FP32   |   |                                   |                | FP64   |  |  | FP32   | FP32   |                                       |  |       |
| FP64   | INT  | INT  | FP32  | FP32   |   |                                   |                | FP64   |  |  | FP32   | FP32   |                                       |  |       |
| FP64   | INT  | INT  | FP32  | FP32   | TEN   | SOR                               | TENSOR         | FP64   |  |  | FP32   | FP32   | TEN                                   | ISOR   | TENSO |
| FP64   | INT  | INT  | FP32  | FP32   |   | KE                                | CORE           | FP64   |  |  | FP32   | FP32   |                                       | JRE  | CORE  |
| FP64   | INT  | INT  | FP32  | FP32   |   |                                   |                | FP64   |  | INT  | FP32   | FP32   |                                       |  |       |
| FP64   | INT  | INT  | FP32  | FP32   |   |                                   |                | FP64   |  |  | FP32   | FP32   |                                       |  |       |
| FP64   | INT  | INT  | FP32  | FP32   |   |                                   |                | FP64   |  |  | FP32   | FP32   |                                       |  |       |
| LD/ LD/<br>ST ST   | LD/<br>ST  | LD/<br>ST  | LD/<br>ST   | LD/<br>ST  | LD/<br>ST                                     | LD/<br>ST                         | SFU            | LD/ LD/<br>ST ST   |  |  |  |  |                                       |  | SFU   |
|  | 111111   | n Sch  | edule   |  | hread/  | clk)                              |                |  | Wai  | m Sch  | edule  | r (32 t  | hread                                 | (clk)  |       |
|  | Dis  | p Sch<br>patch<br>ister  | edule<br>h Unit<br>File (1  | (32 th<br>(32 th   | hread/<br>read/ci<br>4 x 32·                  | clk)<br>lk)<br>-bit)              |                |  | War<br>Di<br>Reg   | rp Sch<br>spatcl<br>jister   | nedule<br>h Unit<br>File ('  | r (32 t<br>(32 th<br>16,384  | hread<br>read/<br>4 x 32              | /clk)<br>clk)<br>2-bit)                              |       |
| FP64   | Regi   | p Sch<br>patch<br>ister  | FP32  | (32 th<br>(32 th<br>16,384   | hread/<br>read/c<br>4 x 32·                   | clk)<br>lk)<br>-bit)              |                | FP64   | War<br>Di<br>Reg   | rp Sch<br>spatcl<br>gister   | File (*  | r (32 th<br>(32 th<br>16,384   | hread<br>read/<br>4 x 32              | //clk)<br>clk)<br>2-bit)                             |       |
| FP64<br>FP64   | Dis<br>Regi<br>INT   | p Sch<br>patch<br>ister<br>INT<br>INT  | File (1<br>FP32<br>FP32   | (32 th<br>(32 th<br>6,384<br>FP32<br>FP32  | hread/c<br>read/c<br>4 x 32·                  | clk)<br>lk)<br>-bit)              |                | FP64<br>FP64   | Wai<br>Di<br>Reg<br>INT  | rp Sch<br>spatcl<br>jister<br>INT<br>INT                           | FP32   | r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32   | hread<br>read/<br>4 x 32              | //clk)<br>clk)<br>2-bit)                             |       |
| FP64<br>FP64<br>FP64   | INT  | p Sch<br>patch<br>ister<br>INT<br>INT<br>INT   | File (1<br>FP32<br>FP32<br>FP32<br>FP32                                 | (32 th<br>(32 th<br>6,384<br>FP32<br>FP32<br>FP32  | hread/c<br>read/c<br>4 x 32                   | clk)<br>lk)<br>-bit)              |                | FP64<br>FP64<br>FP64   | War<br>Di<br>Reg<br>INT<br>INT   | rp Sch<br>spatcl<br>lister<br>INT<br>INT<br>INT                    | File (*<br>FP32<br>FP32<br>FP32  | r (32 t<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32  | hread<br>read/<br>4 x 32              | VcIk)<br>cIk)<br>2-bit)                              |       |
| FP64<br>FP64<br>FP64<br>FP64                                 | INT<br>INT<br>INT<br>INT   | p Sch<br>patci<br>ister<br>INT<br>INT<br>INT   | FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                            | (32 th<br>(32 th<br>6,384<br>FP32<br>FP32<br>FP32<br>FP32  | hread/c<br>read/c<br>4 x 32<br>4 x 32<br>TEN  | cik)<br>ik)<br>-bit)<br>SOR       | TENSOR         | FP64<br>FP64<br>FP64<br>FP64                                 | Wai<br>Di<br>Reg<br>INT<br>INT<br>INT  | rp Sch<br>spatcl<br>ister<br>INT<br>INT<br>INT<br>INT              | File (*<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32  | r (32 t<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32                                  | hread<br>read/4<br>4 x 32             | VcIk)<br>cIk)<br>2-bit)<br>ISOR                      | TENSO |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64                         | INT<br>INT<br>INT<br>INT<br>INT<br>INT                             | p Sch<br>patch<br>ister<br>INT<br>INT<br>INT<br>INT                                    | File (1<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                 | (32 th<br>(32 th<br>6,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32  | hread/c<br>4 x 32<br>TENS                     | cik)<br>ik)<br>-bit)<br>SOR<br>RE | TENSOR         | FP64<br>FP64<br>FP64<br>FP64<br>FP64                         | Wa<br>Di<br>Reg<br>INT<br>INT<br>INT<br>INT                                    | rp Sch<br>spatcl<br>jister<br>INT<br>INT<br>INT<br>INT             | FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                                       | r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                         | hread<br>read/<br>4 x 32<br>TEN<br>CC | VcIk)<br>cIk)<br>2-bit)<br>ISOR<br>DRE               | TENSC |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64                 | INT<br>INT<br>INT<br>INT<br>INT<br>INT                             | p Sch<br>patch<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT                             | File (1<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | (32 th<br>(32 th<br>6,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32  | hread/c<br>4 x 32-<br>TENS                    | clk)<br>lk)<br>-bit)<br>SOR<br>RE | TENSOR         | FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64                 | Wa<br>Di<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT                             | INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                      | FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                               | r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                 | hread/<br>read/<br>4 x 32             | /clk)<br>clk)<br>2-bit)<br>ISOR<br>DRE               | TENSC |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64         | INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                      | p Sch<br>spatch<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                     | FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32            | (32 th<br>(32 th<br>6,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32  | hread/c<br>read/c<br>4 x 32<br>4 x 32<br>TENS | clk)<br>lk)<br>-bit)<br>SOR<br>RE | TENSOR         | FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64         | Water<br>Di<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                          | INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT               | FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32                               | r (32 t<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32          | hread/<br>4 x 32<br>TEN<br>CC         | /clk)<br>clk)<br>2-bit)<br>SSOR<br>DRE               | TENSC |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64 | INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT               | p Sch<br>patch<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                      | FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32            | (32 th<br>(32 th<br>(32 th<br>(33 th<br>(34 th<br>(34 th))<br>(34 th))<br>(34 th)<br>(34 th)<br>(34 th)<br>(34 th)<br>(34 th)<br>(34 th)<br>(34 th)<br>(34 th)<br>(34 th)<br>(32 th)<br>( | hread/c<br>read/c<br>4 x 32<br>TEN:<br>CO     | clk)<br>lk)<br>-bit)<br>SOR<br>RE | TENSOR         | FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64 | Want<br>Di<br>Regg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT                   | INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT        | edule<br>h Unit<br>File (*<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | r (32 t<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32  | hread/4<br>4 x 32                     | /clk)<br>2-bit)<br>SSOR                              | TENSC |
| FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64 | INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT | p Sch<br>patch<br>ister<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT | File (1<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32   | hread/<br>read/ci<br>4 x 32-<br>TEN:<br>CO    | clk)<br>lk)<br>-bit)<br>SOR<br>RE | TENSOR<br>CORE | FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64<br>FP64 | Wan<br>Di<br>Reg<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>LD/<br>ST | INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT<br>INT | edule<br>h Unit<br>File ('<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | r (32 th<br>(32 th<br>16,384<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32<br>FP32 | hread//<br>4 x 32<br>TEN<br>CC        | Icik)<br>Cik)<br>2-bit)<br>SSOR<br>BSOR<br>LD/<br>ST | TENS  |

|           | L0 Instruction Cache            |           |           |           |           |           |           |        |  |
|-----------|---------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|--------|--|
|           |                                 | War       | 'p Sch    | nedule    | r (32 t   | hread     | /clk)     |        |  |
|           |                                 | Dis       | spatc     | h Unit    | (32 th    | read/o    | :lk)      |        |  |
|           | Register File (16,384 x 32-bit) |           |           |           |           |           |           |        |  |
| FP        | 64                              | INT       | INT       | FP32      | FP32      | $\square$ |           |        |  |
| FP        | 64                              | INT       | INT       | FP32      | FP32      |           |           |        |  |
| FP        | 64                              | INT       | INT       | FP32      | FP32      |           |           |        |  |
| FP        | 64                              | INT       | INT       | FP32      | FP32      | TEN       | SOR       | TENSOR |  |
| FP        | 64                              | INT       | INT       | FP32      | FP32      | cc        | ORE       | CORE   |  |
| FP        | 64                              | INT       | INT       | FP32      | FP32      |           |           |        |  |
| FP        | 64                              | INT       | INT       | FP32      | FP32      |           |           |        |  |
| FP        | FP64 INT INT FP32 FP32          |           |           |           |           |           |           |        |  |
| LD/<br>ST | LD/<br>ST                       | LD/<br>ST | LD/<br>ST | LD/<br>ST | LD/<br>ST | LD/<br>ST | LD/<br>ST | SFU    |  |

4 Warp Scheduler per SM

Manages a pool of warps:

Volta: 16 warp slots Turing: 8 warp slots

Each scheduler can issue 1 warp/cycle

Offers simplified mental model for profiling and SM metrics

Mental Model for Profiling



### Mental Model for Profiling



Issue Slot:

### Mental Model for Profiling





### Mental Model for Profiling



Mental Model for Profiling



Mental Model for Profiling



Mental Model for Profiling



Mental Model for Profiling



4

Mental Model for Profiling



4 20

Mental Model for Profiling



Mental Model for Profiling



Mental Model for Profiling



Mental Model for Profiling



Mental Model for Profiling





### Application to kernel\_B



#### Metrics (theoretical; every 8 cycles):

| 8 | 1                |
|---|------------------|
| 7 | 7/8              |
| 1 | 1/8              |
| 1 | 1/8              |
|   | 8<br>7<br>1<br>1 |



### Application to kernel\_B



#### Metrics (theoretical; every 8 cycles):

| 8 | 1                |
|---|------------------|
| 7 | 7/8              |
| 1 | 1/8              |
| 1 | 1/8              |
|   | 8<br>7<br>1<br>1 |

#### Metrics (from report):

| 1.00 |
|------|
| 0.87 |
| 0.13 |
| 0.13 |
|      |



### Application to kernel\_A



#### FP64 Pipeline on Volta:

Dependent Issue Rate: Issue Rate: 8 cycles 4 cycles

| L0 Instruction Cache            |                                |                  |                  |        |  |  |  |
|---------------------------------|--------------------------------|------------------|------------------|--------|--|--|--|
|                                 | Warp Scheduler (32 thread/clk) |                  |                  |        |  |  |  |
|                                 | Dispato                        | h Unit (32 th    | read/clk)        |        |  |  |  |
| Register File (16,384 x 32-bit) |                                |                  |                  |        |  |  |  |
| FP64                            | INT INT                        | FP32 FP32        |                  |        |  |  |  |
| FP64                            | INT INT                        | FP32 FP32        |                  |        |  |  |  |
| FP64                            | INT INT                        | FP32 FP32        |                  |        |  |  |  |
| FP64                            | INT INT                        | FP32 FP32        | TENSOR           | TENSOR |  |  |  |
| FP64                            | INT INT                        | FP32 FP32        | CORE             | CORE   |  |  |  |
| FP64                            | INT INT                        | FP32 FP32        |                  |        |  |  |  |
| FP64                            | INT INT                        | FP32 FP32        |                  |        |  |  |  |
| FP64                            | INT INT FP32 FP32              |                  |                  |        |  |  |  |
| LD/ LD/<br>ST ST                | LD/ LD/<br>ST ST               | LD/ LD/<br>ST ST | LD/ LD/<br>ST ST | SFU    |  |  |  |

30 📀 💿 🕺 30



### Application to kernel\_A



#### FP64 Pipeline on Volta:

Dependent Issue Rate: Issue Rate: 8 cycles 4 cycles

| L0 Instruction Cache            |                   |                                    |        |        |           |      |        |  |
|---------------------------------|-------------------|------------------------------------|--------|--------|-----------|------|--------|--|
| Warp Scheduler (32 thread/clk)  |                   |                                    |        |        |           |      |        |  |
|                                 | Di                | spatcl                             | n Unit | (32 th | read/o    | clk) |        |  |
| Register File (16,384 x 32-bit) |                   |                                    |        |        |           |      |        |  |
| õ                               | INT               | INT                                | FP32   | FP32   | $\square$ |      |        |  |
| F                               | INT               | INT                                | FP32   | FP32   | H         |      |        |  |
| \a<br>l                         | INT               | INT                                | FP32   | FP32   |           |      |        |  |
| F                               | INT               | INT                                | FP32   | FP32   | TEN       | SOR  | TENSOR |  |
| FP64                            | INT               | INT                                | FP32   | FP32   | cc        | DRE  | CORE   |  |
| FP64                            | INT               | INT                                | FP32   | FP32   |           |      |        |  |
| FP64                            | INT               | INT                                | FP32   | FP32   |           |      |        |  |
| FP64                            | INT INT FP32 FP32 |                                    |        |        |           |      |        |  |
| LD/ LD/<br>ST ST                | LD/<br>ST         | LD/ LD/ LD/ LD/ LD/ ST ST ST ST ST |        |        |           |      |        |  |

31 📀 💿 🕺 31



### Application to kernel\_A



#### FP64 Pipeline on Volta:

Dependent Issue Rate: Issue Rate: 8 cycles 4 cycles

| L0 Instruction Cache            |                        |           |           |           |           |           |        |
|---------------------------------|------------------------|-----------|-----------|-----------|-----------|-----------|--------|
| Warp Scheduler (32 thread/clk)  |                        |           |           |           |           |           |        |
|                                 | Di                     | spatcl    | n Unit (  | 32 th     | read/c    | :lk)      |        |
| Register File (16,384 x 32-bit) |                        |           |           |           |           |           |        |
| Ö                               | INT                    | INT       | FP32      | FP32      |           |           |        |
| F                               | INT                    | INT       | FP32      | FP32      |           |           |        |
| \a_                             | INT                    | INT       | FP32      | FP32      |           |           |        |
| F                               | INT                    | INT       | FP32      | FP32      | TEN       | SOR       | TENSOR |
| FP64                            | INT                    | INT       | FP32      | FP32      | cc        | RE        | CORE   |
| FP64                            | INT                    | INT       | FP32      | FP32      |           |           |        |
| FP64                            | INT                    | INT       | FP32      | FP32      |           |           |        |
| FP64                            | FP64 INT INT FP32 FP32 |           |           |           |           |           |        |
| LD/ LD/<br>ST ST                | LD/<br>ST              | LD/<br>ST | LD/<br>ST | LD/<br>ST | LD/<br>ST | LD/<br>ST | SFU    |

32 📀 💿 NIDIA



### Application to kernel\_A



#### FP64 Pipeline on Volta:

Dependent Issue Rate: Issue Rate: 8 cycles 4 cycles

| L0 Instruction Cache            |                     |                                 |        |        |        |      |        |  |
|---------------------------------|---------------------|---------------------------------|--------|--------|--------|------|--------|--|
| Warp Scheduler (32 thread/clk)  |                     |                                 |        |        |        |      |        |  |
|                                 | Dis                 | patch                           | n Unit | (32 th | read/o | clk) |        |  |
| Register File (16,384 x 32-bit) |                     |                                 |        |        |        |      |        |  |
| Õ                               | INT                 | INT                             | FP32   | FP32   |        |      |        |  |
| F                               | INT                 | INT                             | FP32   | FP32   |        |      |        |  |
| /a⊓                             | INT                 | INT                             | FP32   | FP32   |        |      |        |  |
| F                               | INT                 | INT                             | FP32   | FP32   | TEN    | SOR  | TENSOR |  |
| FP64                            | INT                 | INT                             | FP32   | FP32   | cc     | DRE  | CORE   |  |
| FP64                            | INT                 | INT                             | FP32   | FP32   |        |      |        |  |
| FP64                            | INT                 | INT                             | FP32   | FP32   |        |      |        |  |
| FP64                            | 4 INT INT FP32 FP32 |                                 |        |        |        |      |        |  |
| LD/ LD/<br>ST ST                | LD/<br>ST           | LD/ LD/ LD/ LD/ LD/ ST ST ST ST |        |        |        |      |        |  |

33 💿 💿 🕺 🕺 🕺 33



### Application to kernel\_A



#### FP64 Pipeline on Volta:

Dependent Issue Rate: Issue Rate: 8 cycles 4 cycles



34 💿 💿 🕺 🕺 🕺 34



### Application to kernel\_A



Metrics (theoretical; every 8 cycles):

| warps_active   | 96 |
|----------------|----|
| warps_stalled  | 74 |
| warps_eligible | 22 |
| warps_selected | 2  |



### Application to kernel\_A



#### Metrics (theoretical; every 8 cycles):

| 96 | 96/8=12.00          |
|----|---------------------|
| 74 | 74/8= 9.25          |
| 22 | 22/8= 2.75          |
| 2  | 2/8= 0.25           |
|    | 96<br>74<br>22<br>2 |

#### Metrics (from report):

| warps_active   | 12.27 |
|----------------|-------|
| warps_stalled  | 9.08  |
| warps_eligible | 2.80  |
| warps_selected | 0.26  |

### TAKEAWAYS Profiling with Nsight Compute

#### General:

- Use tools whenever possible
- Understanding app/gpu/system and tools go hand in hand (on us to make this as easy as possible)

#### Nsight Compute:

- Use SOL metrics to understand overall limiter (reading real-world reports is a lot more difficult; but the very same workflow applies)
- Read rules' output for guidance throughout the report (we'll add more rules in future releases)
- Use top-down approach; no need to jump directly into SASS code
- Do not optimize stalls, if you already use all/sufficient issue slots
- Let us know what works for you and what doesn't: devtalk.nvidia.com > Development Tools > Nsight Compute

## **DEVELOPER TOOLS AT GTC19**

#### Talks:

S9751: Accelerate Your CUDA Development with Latest Debugging and Code Analysis Developer Tools, Tue @9am S9866 - Optimizing Facebook AI Workloads for NVIDIA GPUs, Tue @9am

S9345: CUDA Kernel Profiling using NVIDIA Nsight Compute, Tue @1pm

S9661: Nsight Graphics - DXR/Vulkan Profiling/Vulkan Raytracing, Wed @10am

S9503: Using Nsight Tools to Optimize the NAMD Molecular Dynamics Simulation Program, Wed @1pm

#### Hands-on labs:

L9102: Jetson Developer Tools Training Lab, Mon @9am, 11:30am

L9124: Debugging and optimizing CUDA applications with Nsight products on Linux training lab, Tue <a>0</a>8am, 10am</a>

Connect with the Experts (where Developer Tools will be available):

CE9123: CUDA & Graphics Developer Tools, Tue @2pm, Wed @3pm CE9137: Jetson Embedded Platform, Tue @12pm, 5pm, Wed @1pm, 4pm, Thu @12pm

#### Podium: Demos of DevTools products on Linux, DRIVE AGX & Jetson AGX at the showfloor

Tue @12pm - 7pm Wed @12pm - 7pm Thu @11am - 2pm

# THANK YOU

Any Questions?



### **BACKUP SLIDES**

### **CUDA PROFILING TOOLS**

NVIDIA® Visual Profiler - CUDA trace/kernel profiling

GUI - runs on Linux, Windows, Mac

nvprof - CUDA trace/kernel profiling

Command line tool - runs on Linux, Windows, Mac

NVIDIA® Nsight<sup>™</sup> Visual Studio Edition - CUDA, OpenGL and system trace and CUDA kernel profiling Integration into Visual Studio 2012 and newer - Windows only

## **NSIGHT SYSTEMS**

### Features

Multicore CPU and multi-GPU system activity trace

CPU utilization, thread state and core/thread migration

OS library trace with blocking call backtraces

pthread, semaphore, file I/O, network I/O, poll/select, sleep/yield, ioctl, syscall, fork

CUDA, OpenACC, OpenGL API and GPU Workload trace Includes UVM events cuDNN and cuBLAS

NVidia Tools eXtension (NVTX) user annotations

Command line