CUDA DEVELOPER TOOLS: NEW FEATURES AND CAPABILITIES

Rafael Campana, Nov. 19, 2019
DEVELOPER TOOLS PORTFOLIO

Application Development IDE integration

Debug CUDA
- CUDA GPU crash dump
- CUDA Profiling

Debug Gfx API
- Gfx GPU crash dump
- Graphics Profiling
- System Profiling

Nsight Eclipse Edition
Nsight Visual Studio Edition

Nsight Eclipse Edition
cuda-gdb
Nsight Visual Studio Edition
Nsight Compute
cuda-memcheck & Compute Sanitizer API
Nsight Graphics

Nsight-gdb/Nsight Eclipse Edition
Nsight Visual Studio Edition
Nsight Aftermath

Nsight Systems
• Plug-in to Eclipse
  Eclipse 4.7, 4.8 and 4.9 support

• Edit, build, debug CUDA-C applications

• CUDA aware source code editor - syntax highlighting, code completion and inline help

• Debugger - Seamless and simultaneous debugging of CPU and GPU code

• NVCC build integration to cross compile for various target platforms

• Docker support

• Documentation: https://docs.nvidia.com/cuda/nsight-eclipse-plugins-guide
NSIGHT VISUAL STUDIO EDITION

Plug-in

Visual Studio 2015, 2017 and 2019 support
Native CUDA C/C++ GPU Debugging
Source-correlated assembly debugging
(SASS / PTX / SASS+PTX)
Data breakpoints for CUDA C/C++ code
Expressions in Locals, Watch and Conditionals
CUDA info view
Warp Watch

CUDA-GDB
Overview

Provides the debugging features of Nsight Eclipse Edition

Command line source and assembly (SASS) level debugger

Simultaneous CPU and GPU debugging

Inspect and modify memory, register, variable state

Control program execution

Runtime GPU error detection

Support for multiple GPUs, multiple contexts, multiple kernels, Thread focus

Core dump support

CUDA-GDB

New Features

Support for ARM (Server Base System Architecture) preview

Performance improvements

- Module load time (~30% faster)

Quality improvements

- Improved handling of --lineinfo debug information (OptiX)
- Improved display of uniform registers (Turing)
CUDA-MEMCHECK

Functional correctness checking tool suite

Multiple tools

memcheck: reports out of bounds/misaligned memory access errors

racecheck: identifies races on __shared__ memory

initcheck: usage of uninitialized global memory

synccheck: identify invalid usage of __syncthreads() and __syncwarp()
SANITIZER

API

Provides finer control than cuda-memcheck through APIs to analyze memory patterns.

APIs are grouped in two:

- **Callback API** - CUDA events such as memory allocations/kernel
- **Patching API** - inserts patches for specific memory instructions


Samples: [https://github.com/NVIDIA/compute-sanitizer-samples](https://github.com/NVIDIA/compute-sanitizer-samples)
NSIGHT SUITE OF PROFILERS
NSIGHT TOOLS WORKFLOW

Start here

Nsight Systems
Comprehensive system-level performance

Nsight Compute
Detailed CUDA kernel performance

Dive into top CUDA kernels by using metrics/counter collection

Nsight Graphics
Detailed frame/render performance

Dive into graphics frames

Re-check overall performance

Re-check overall performance
NSIGHT SYSTEMS
Overview

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

OS: Linux (x86, Tegra), Windows, MacOSX (host only)

Processes and threads
CUDA and OpenGL API trace
cuDNN and cuBLAS trace
Multi-GPU
Kernel and memory transfer activities
Thread/core migration
Thread state
NSIGHT SYSTEMS

New Features

- ARM (SBSA) support
- ftrace collection
- NVTX correlated to GPU
- Event table
NSIGHT SYSTEMS

MPI API Trace

Select the MPI implementation used by the target application to trace a default set of synchronous MPI calls. If the application uses a different MPI implementation, see the documentation for additional setup required to trace MPI. Note that NVTX tracing will also be enabled on selecting MPI tracing.

- OpenMPI
- MPICH and its derivatives

Collect NVTX trace
NSIGHT SYSTEMS
New Trace Collections

CUDA Kernel Backtraces

FTrace events
NSIGHT SYSTEMS

FTrace
NSIGHT COMPUTE
Next-Gen Kernel Profiling Tool

Key Features:
• Interactive CUDA API debugging and kernel profiling
• Fast Data Collection
• Improved Workflow (Diff’ing Results)
• Fully Customizable (Programmable UI/Rules)
• Command Line, Standalone, IDE Integration

OS: Linux (x86, Power, Tegra), Windows, MacOSX (host only)
GPUs: Pascal, Volta, Turing

NSIGHT COMPUTE
Profile Report - Details Page

Focused Sections

All Data on Single Page

Ordered from Top-Level to Low-Level
NSIGHT COMPUTE

Section Example

Section Header
provides overview & context for other sections

Section Body
provides additional details (tables & charts)

Section Config
completely data driven
add/modify/change sections
NSIGHT COMPUTE
Unguided Analysis / Rules System

Analysis Rules
recommendations from nvvp and more

Rules Config
completely data driven
add/modify/change rules
NSIGHT COMPUTE

**Diff’ing kernel runs**

- **Metric delta**: current values and changes from baseline
- **Baseline**: from any previous profile report (different kernel, gpu, ...)

**Chart difference**: current values and baseline values

---

### Recommendations

- **Bottleneck**: Simple GPU bottleneck detection.

### GPU Utilization

- **% SM Busy**: 0.61 (+118.69%)
- **% Memory Busy**: 86.97 (+51.56%)
NSIGHT COMPUTE

New Features

Support for PowerPC target architecture

Support for ARM (SBSA) preview target architecture

Profile activity command line is shown
Detailed breakdown of high-level SOL metrics
NSIGHT COMPUTE

New Features

Improved help for Source page and Details page charts
NSIGHT COMPUTE

New Features

The Memory Workload Analysis chart now supports baselines.
NSIGHT COMPUTE

New Features

Command line interface has new profiling and output controls

```
C:\Program Files\NVIDIA Corporation\Nsight Compute 2019.5.0>nv-nsight-cu-cli.bat --list-chips gp102, gp104, gp106, gp107, gp108, gv100, gv11b, tu102, tu104, tu106, tu116, tu117

--cache-control arg (=all) Control the behavior of the GPU caches during profiling. Allowed values:
  all
  none

C:\Program Files\NVIDIA Corporation\Nsight Compute 2019.5.0\target\windows-desktop-win7-x64>nv-nsight-cu-cli.exe --kernel-base demangled [314] python3.5@127.0.0.1
  void mxnet::op::mxnet_op::softmax_compute_kernel<7, mxnet::op::mxnet_op::softmax_fwd, false, float, 2, mshadow::half::half_t, mshadow::Shape<2>, mshadow::Shape<2>, double), 2019-Aug-07 07:25:25, Context 1, Stream 80
```
SUMMARY

Full stack of Developer Tools available across platforms and APIs
Overall user consistent experience across many platforms
Compute & Graphics API support
Lower overhead (performance and footprint)

Visit:  https://developer.nvidia.com/tools-overview