



# NVIDIA CUDA TOOLKIT V7.0

RN-06722-001 \_v7.0 | March 2015

**Release Notes for Windows, Linux, and Mac OS**



# TABLE OF CONTENTS

<b>Chapter 1. CUDA Toolkit Major Components.....</b>	<b>1</b>
<b>Chapter 2. New Features.....</b>	<b>3</b>
2.1. General CUDA.....	3
2.2. CUDA Tools.....	4
2.2.1. CUDA Compiler.....	4
2.2.2. CUDA-GDB.....	4
2.2.3. CUDA-MEMCHECK.....	4
2.2.4. CUDA Profiler.....	4
2.2.5. Nsight Eclipse Edition.....	5
2.3. CUDA Libraries.....	5
2.3.1. cuBLAS Library.....	5
2.3.2. cuFFT Library.....	5
2.3.3. cuSOLVER Library.....	5
2.3.4. cuSPARSE Library.....	5
2.3.5. CUDA Math Library.....	6
2.3.6. Thrust Library.....	6
2.4. CUDA Samples.....	6
<b>Chapter 3. Unsupported Features.....</b>	<b>7</b>
<b>Chapter 4. Deprecated Features.....</b>	<b>9</b>
<b>Chapter 5. Performance Improvements.....</b>	<b>10</b>
5.1. CUDA Libraries.....	10
5.1.1. cuFFT Library.....	10
5.1.2. CUDA Math Library.....	10
<b>Chapter 6. Resolved Issues.....</b>	<b>11</b>
6.1. General CUDA.....	11
6.2. CUDA Tools.....	11
6.2.1. CUDA Compiler.....	11
6.2.2. Nsight Eclipse Edition.....	11
6.2.3. NVIDIA Visual Profiler.....	12
6.3. CUDA Libraries.....	12
6.3.1. cuFFT Library.....	12
6.3.2. cuSPARSE Library.....	12
6.4. CUDA Samples.....	12
<b>Chapter 7. Known Issues.....</b>	<b>13</b>
7.1. General CUDA.....	13
7.2. CUDA Tools.....	14
7.2.1. CUDA Compiler.....	14
7.2.2. CUDA Profiler.....	14
7.3. CUDA Libraries.....	14
7.3.1. Thrust Library.....	14

# Chapter 1.

## CUDA TOOLKIT MAJOR COMPONENTS

This section provides an overview of the major components of the CUDA Toolkit and points to their locations after installation.

### Compiler

The CUDA-C and CUDA-C++ compiler, **nvcc**, is found in the **bin/** directory. It is built on top of the NVVM optimizer, which is itself built on top of the LLVM compiler infrastructure. Developers who want to target NVVM directly can do so using the Compiler SDK, which is available in the **nvvm/** directory.

### Tools

The following development tools are available in the **bin/** directory (except for Nsight Visual Studio Edition (VSE) which is installed as a plug-in to Microsoft Visual Studio).

- ▶ IDEs: **nsight** (Linux, Mac), Nsight VSE (Windows)
- ▶ Debuggers: **cuda-memcheck**, **cuda-gdb** (Linux, Mac), Nsight VSE (Windows)
- ▶ Profilers: **nvprof**, **nvvp**, Nsight VSE (Windows)
- ▶ Utilities: **cuobjdump**, **nvdiasm**

### Libraries

The scientific and utility libraries listed below are available in the **lib/** directory (DLLs on Windows are in **bin/**), and their interfaces are available in the **include/** directory.

- ▶ **cublas** (BLAS)
- ▶ **cublas\_device** (BLAS Kernel Interface)
- ▶ **cuda\_occupancy** (Kernel Occupancy Calculation [header file implementation])
- ▶ **cuda devrt** (CUDA Device Runtime)
- ▶ **cuda rt** (CUDA Runtime)
- ▶ **cufft** (Fast Fourier Transform [FFT])
- ▶ **cupti** (Profiling Tools Interface)
- ▶ **curand** (Random Number Generation)
- ▶ **cusolver** (Dense and Sparse Direct Linear Solvers and Eigen Solvers)
- ▶ **cuspars** (Sparse Matrix)
- ▶ **npp** (NVIDIA Performance Primitives [image and signal processing])
- ▶ **nvblas** ("Drop-in" BLAS)

- ▶ **nvcuvid** (CUDA Video Decoder [Windows, Linux])
- ▶ **nVRTC** (CUDA Runtime Compilation)
- ▶ **thrust** (Parallel Algorithm Library [header file implementation])

### CUDA Samples

Code samples that illustrate how to use various CUDA and library APIs are available in the **samples/** directory on Linux and Mac, and are installed to **C:\ProgramData\NVIDIA Corporation\CUDA Samples** on Windows. On Linux and Mac, the **samples/** directory is read-only and the samples must be copied to another location if they are to be modified. Further instructions can be found in the *Getting Started Guides* for Linux and Mac.

### Documentation

The most current version of these release notes can be found online at <http://docs.nvidia.com/cuda/cuda-toolkit-release-notes/index.html>.

Documentation can be found in PDF form in the **doc/pdf/** directory, or in HTML form at **doc/html/index.html** and online at <http://docs.nvidia.com/cuda/index.html>.

### CUDA-GDB Sources

CUDA-GDB sources are available as follows:

- ▶ For CUDA Toolkit 7.0 and newer, in the installation directory **extras/**. The directory is created by default during the toolkit installation unless the **.rpm/** **.deb** package installers are used. In this case, the **cuda-gdb-src** package must be manually installed.
- ▶ For CUDA Toolkit 6.5, 6.0, and 5.5, at <https://github.com/NVIDIA/cuda-gdb>.
- ▶ For CUDA Toolkit 5.0 and earlier, at <ftp://download.nvidia.com/CUDAOpen64/>.
- ▶ Upon request by sending an e-mail to <mailto:oss-requests@nvidia.com>.

# Chapter 2.

## NEW FEATURES

### 2.1. General CUDA

- ▶ The default stream, used either when `0` is passed as a `cudaStream_t` or by APIs that operate on a stream implicitly, can now be configured to be a separate stream per thread that will not synchronize with other streams. Currently, operations in the default stream will serialize with other streams (see *Stream Synchronization Behavior* in the API documentation). This new behavior is not enabled by default and can be controlled per compilation unit with the `--default-stream nvcc` option.
- ▶ Added a method to the CUDA Driver API, `cuDevicePrimaryCtxRetain()`, that allows a program to create (or to access if it already exists) the same CUDA context for a GPU device as the one used by the CUDART (CUDA Runtime API) library. This context is referred to as the primary context, and this new method allows for sharing the primary context between CUDART and other threads, which can reduce the performance overhead of creating and maintaining multiple contexts per device.
- ▶ Unified the device enumeration for CUDA, NVML, and related tools. Variable `CUDA_DEVICE_ORDER` can have a value of `FASTEST_FIRST` (default) or `PCI_BUS_ID`.
- ▶ Instrumented NVML (NVIDIA Management Library) and the CUDA driver to ignore GPUs that have been made inaccessible via cgroups (control groups). This enables schedulers that rely on cgroups to enforce device access restrictions for their jobs. Job schedulers wanting to use cgroups for device restriction need CUDA and NVML to handle those restrictions in a graceful way.
- ▶ Multi-Process Service (MPS) is now supported on GeForce products in addition to Tesla and Quadro products.
- ▶ Multi-Process Service now supports concurrent execution of GPU tasks on multiple GPUs at once. At startup, the MPS server attempts to spin up a context on each visible GPU and make it available to the client application.
- ▶ The Windows and Mac OS X installers are now also available as network installers. A network installer is much smaller than the traditional local installer and downloads only the components selected for installation.

## 2.2. CUDA Tools

### 2.2.1. CUDA Compiler

- ▶ On supported x86\_64 Linux operating systems, the PGI C/C++ compiler (**pgc++**) is supported as a host compiler by **nvcc**.
- ▶ POWER8 XL compiler 13.1.1 is supported.
- ▶ Added support for GCC 4.9.
- ▶ On Mac OS X, **libc++** is supported with XCode 5.x. Command-line option **-Xcompiler -stdlib=libc++** is no longer needed when invoking NVCC. Instead, NVCC uses the default library that Clang chooses on Mac OS X. Users are still able to choose between **libc++** and **libstdc++** by passing **-Xcompiler -stdlib=libc++** or **-Xcompiler -stdlib=libstdc++** to NVCC.
- ▶ The Runtime Compilation library (**nVRTC**) provides an API to compile CUDA-C++ device source code at runtime. The resulting compiled PTX can be launched on a GPU using the CUDA Driver API. More details can be found in the *libNVRTC User Guide*.
- ▶ Added C++11 support. The new **nvcc** flag **-std=c++11** turns on C++11 features in the CUDA compiler as well as the host compiler and linker. The flag is supported by host compilers Clang and GCC versions 4.7 and newer. In addition, any C++11 features that are enabled by default by a supported host compiler are also allowed to be used in device code. Please see the *CUDA Programming Guide* for further details.

### 2.2.2. CUDA-GDB

- ▶ Starting with CUDA 7.0, GPU core dumps can read by CUDA-GDB with the **target cudacore \${gpubcoredump}** and **target core \${cpucoredump} \${gpubcoredump}** commands.
- ▶ Enabled CUDA applications to generate a GPU core dump when an exception is hit on the GPU. The feature is supported on Windows, Mac OS X, and Linux desktops. (Android, L4T, and Vibrante support may come in the future.) On Windows, this feature is only supported in TCC mode. On Unix-like operating systems (Linux, OS X, etc.), a CPU core dump is generated along with a GPU core dump.

### 2.2.3. CUDA-MEMCHECK

- ▶ Enabled the reporting of divergent block synchronization.
- ▶ Enabled the tracking and reporting of uninitialized global memory.

### 2.2.4. CUDA Profiler

- ▶ On supported chips (sm\_30 and beyond), all hardware counters exposed by CUDA profiling tools (**nvprof**, **nvvp**, and Nsight Eclipse Edition) can now be profiled from multiple applications at the same time.

## 2.2.5. Nsight Eclipse Edition

- ▶ Cross compiling to the POWER8 target architecture using the GNU tool-chain is now supported within the Nsight IDE.

## 2.3. CUDA Libraries

### 2.3.1. cuBLAS Library

- ▶ A license is no longer required in order to use cuBLAS-XT with more than two GPUs.
- ▶ The batched LU solver `cublas{T}getrsBatched` routine has been added to cuBLAS. It takes the output of the batched factorization routines `cublas{T}getrfBatched` to compute the solution given the provided batch of right-hand-side matrices.

### 2.3.2. cuFFT Library

- ▶ For CUDA 7.0, support for callback routines, invoked when cuFFT loads and/or stores data, no longer requires an evaluation license file.
- ▶ For CUDA 7.0, cuFFT multiple-GPU execution is supported on up to four GPUs, except for single 1D complex-to-complex transforms, which are supported on two or four GPUs.
- ▶ In CUDA 7.0, transform execution may be distributed to four GPUs with the same CUDA architecture. In addition, multiple GPU support for two or four GPUs is no longer constrained to GPUs on a single board.
- ▶ For CUDA 7.0, single complex-to-complex 2D and 3D transforms with dimensions that can be factored into primes less than or equal to 127 are supported on multiple GPUs. Single complex-to-complex 1D transforms on multiple GPUs continue to be limited to sizes that are powers of 2.

### 2.3.3. cuSOLVER Library

- ▶ CUDA 7.0 introduces cuSOLVER, a new library that is a collection of routines to solve linear systems and Eigen problems. It includes dense and sparse linear solvers and sparse refactorization.
- ▶ Enabled offloading dense linear algebra calls to the GPUs in a sparse direct solver.

### 2.3.4. cuSPARSE Library

- ▶ Added a new `cusparse<t>csrgermm2()` routine optimized for small matrices and operations  $C = a \cdot A \cdot B + b \cdot D$ , where **A**, **B**, and **D** are CSR matrices.
- ▶ Added graph coloring.

### 2.3.5. CUDA Math Library

- ▶ Support for 3D and 4D Euclidean norm and 3D Euclidean reciprocal norm has been added to the math library.

### 2.3.6. Thrust Library

- ▶ Thrust version 1.8.0 introduces support for algorithm invocation from CUDA `__device__` code, support for CUDA streams, and algorithm performance improvements. Users may now invoke Thrust algorithms from CUDA `__device__` code, providing a parallel algorithms library to CUDA programmers authoring custom kernels as well as allowing Thrust programmers to nest their algorithm calls within functors. The `thrust::seq` execution policy allows users to require sequential algorithm execution in the calling thread and makes a sequential algorithms library available to individual CUDA threads. The `.on(stream)` syntax allows users to request a CUDA stream for kernels launched during algorithm execution. Finally, new CUDA algorithm implementations provide substantial performance improvements.

## 2.4. CUDA Samples

- ▶ The CUDA Samples makefile `x86_64=1` and `ARMv7=1` options have been deprecated. Please use `TARGET_ARCH` to set the targeted build architecture instead. The CUDA Samples makefile `GCC` option has been deprecated. Please use `HOST_COMPILER` to set the host compiler instead.



# Chapter 3.

## UNSUPPORTED FEATURES

The following features are officially unsupported in the current release. Developers must employ alternative solutions to these features in their software.

### **Support for 32-bit x86 Linux Systems**

The CUDA Toolkit and CUDA Driver no longer support developing and running CUDA and OpenCL Applications on 32-bit x86 Linux operating systems.

**Note 1:** Developing and running 32-bit applications on 64-bit (x86\_64) Linux operating systems is still supported, but that functionality is marked as deprecated and may be dropped in a future release. 64-bit applications are not impacted.

**Note 2:** This notice applies to x86 architectures only; 32-bit application support on the ARM architecture remains officially supported.

### **Red Hat Enterprise Linux 5 and CentOS 5**

CUDA no longer supports the RHEL 5 and CentOS 5 Linux distributions. Please note that RHEL 6, RHEL 7, CentOS 6, and CentOS 7 are all supported.

### **CUDA Toolkit and CUDA Driver Support for Tesla Architecture**

The CUDA Toolkit and CUDA Driver no longer supports the sm\_10, sm\_11, sm\_12, and sm\_13 architectures. As a consequence, `CU_TARGET_COMPUTE_1x` enum values have been removed from the CUDA headers.

### **Certain CUDA Features on 32-bit and 32-on-64-bit Windows Systems**

The CUDA Toolkit no longer supports 32-bit Windows operating systems. Furthermore, the Windows CUDA Driver no longer supports Tesla and Quadro products on 32-bit Windows operating systems. Additionally, on 64-bit Windows operating systems, the following features are no longer supported by the CUDA driver or CUDA toolkit:

- ▶ Running 32-bit applications on Tesla and Quadro products
- ▶ Using the Thrust library from 32-bit applications
- ▶ 32-bit versions of the CUDA Toolkit scientific libraries, including cuBLAS, cuSPARSE, cuFFT, cuRAND, and NPP
- ▶ 32-bit versions of the CUDA samples

Note the above list doesn't impact any 64-bit components on Windows.

### **Using gcc as a Host Compiler on Mac OS X**

On Mac OS X platforms, `nvcc` no longer supports the `gcc` toolchain for compiling host code, including the GNU stdlibc++ standard C++ library. Developers should use

the **clang/llvm** toolchain for compiling host code instead; **nvcc** does this by default in CUDA 7.0 on supported Mac OS X platforms.

# Chapter 4.

## DEPRECATED FEATURES

The following features are deprecated in the current release of the CUDA software. The features still work in the current release, but their documentation may have been removed, and they will become officially unsupported in a future release. We recommend that developers employ alternative solutions to these features in their software.

### Developing and Running 32-bit CUDA and OpenCL Applications on x86 Linux Platforms

Support for developing and running 32-bit CUDA and OpenCL applications on 64-bit x86 Linux platforms is deprecated.

### CUDA Samples Makefile `x86_64=1` and `ARMv7=1` Options

The CUDA Samples Makefile `x86_64=1` and `ARMv7=1` options have been deprecated. Please use `TARGET_ARCH` to set the targeted build architecture instead. The CUDA Samples Makefile `GCC` option has been deprecated. Please use `HOST_COMPILER` to set the host compiler instead.

### Header File `sobol_direction_vectors.h`.

The `sobol_direction_vectors.h` header file is deprecated. This file allowed developers to employ the cuRAND device API with sobol distributions. However, since it is large and takes significant amounts of time and RAM to compile, this file is deprecated in favor of the `curandGetDirectionVectors{32,64}()` and `curandGetScrambleConstants{32,64}()` functions. These functions return a memory pointer to the direction vectors that are precompiled into the cuRAND library, and developers should use this pointer to copy the vectors to the GPU device memory.

# Chapter 5.

## PERFORMANCE IMPROVEMENTS

### 5.1. CUDA Libraries

#### 5.1.1. cuFFT Library

- ▶ For CUDA 7.0, a new mode has been added for copying single 1D transform input from the host to the GPU. This mode redistributes the inputs as required by the first computation phase of the algorithm and eliminates the overhead of this data redistribution from the execution phase.
- ▶ In CUDA 7.0, new composite-sized FFT kernels have been added for many sizes which can be factored into small primes. Composite sizes up to 256 are nearly completely represented, and there is some coverage for sizes up to 1920. This was done to reduce the number of kernel invocations done on the host and will result in significant speed improvements for many composite sizes. Note that these composite-size kernels are not used in combination with the cuFFT callback feature.

#### 5.1.2. CUDA Math Library

- ▶ The performance of the double-precision reciprocal instruction **rcp(x)** in round-to-nearest mode was significantly improved.

# Chapter 6.

## RESOLVED ISSUES

### 6.1. General CUDA

- ▶ On openSUSE and SLES, X no longer fails to load if the CUDA Toolkit RPM packages are installed using relocation immediately following an installation of the `cuda-drivers` package.
- ▶ The Windows toolkit installation no longer fails if Visual Studio, `Nvda.Launcher.exe`, `Nsight.Monitor.exe`, or `Nvda.CrashReporter.exe` is running.
- ▶ The `cuda` and `gpu-deployment-kit` packages must be installed by separate executions of `yum`. See the *Linux Getting Started Guide* for more details.
- ▶ The CUDA reference manual now correctly describes the CUDA device pointer `CUdeviceptr` as an unsigned integer type whose size matches the size of a pointer on the target platform.

### 6.2. CUDA Tools

#### 6.2.1. CUDA Compiler

- ▶ A compiler bug that caused a lack of synchronization at the end of loop iterations was fixed. The bug would cause a barrier instruction to be reached by some threads and not others. For these cases to work correctly, threads should be synchronized at the end of each loop iteration and begin each new iteration in lockstep.

#### 6.2.2. Nsight Eclipse Edition

- ▶ Starting with CUDA 7.0, to cross-compile a CUDA project, Nsight uses the default cross-compiler available on the host (or remote) operating system.

### 6.2.3. NVIDIA Visual Profiler

- ▶ In the Visual Profiler timeline, the colors of intervals on the **Compute** and **Stream** timelines are no longer incorrect. Also the timeline modes **color by stream** and **color by process** do work.

## 6.3. CUDA Libraries

### 6.3.1. cuFFT Library

- ▶ An issue with certain cuFFT plans that caused an assertion in the execution phase has been fixed. The issue applied to plans with all of the following characteristics: real input to complex output (R2C), in-place, native compatibility mode, certain even transform sizes, and more than one batch.

### 6.3.2. cuSPARSE Library

- ▶ CUDA 7.0 fixed bugs in the `csr2csc()` and `bsr2bsc()` routines that were in the CUDA 6.0 and 6.5 releases. As a consequence, `csrsv()`, `csrsv2()`, `csrsm()`, `bsrsv2()`, `bsrsm2()`, and `csrgemm()` now produce correct results when working with transpose (`CUSPARSE_OPERATION_TRANSPOSE`) or conjugate-transpose (`CUSPARSE_OPERATION_CONJUGATE_TRANSPOSE`) operations.

## 6.4. CUDA Samples

- ▶ To properly build the `simpleCUFFT_callback` sample, the `-dc` compiler flag no longer must be added to the compilation commands.

# Chapter 7.

## KNOWN ISSUES

### 7.1. General CUDA

- ▶ Prior to the removal of the NVIDIA driver debian packages, the `nvidia-persistenced` daemon must be shut down manually or the removal will fail.
- ▶ The Canonical repackaging of the NVIDIA drivers on Ubuntu 14.04 may fail to upgrade cleanly to the latest version due to conflicts with the `/etc/OpenCL/vendors/nvidia.icd` file. To address this issue, remove the `nvidia-opencl-icd-*` package and then try the upgrade again. For more information, see <https://bugs.launchpad.net/ubuntu/+source/nvidia-graphics-drivers-331-updates/+bug/1328762>.
- ▶ If the Windows toolkit installation fails, it may be because Visual Studio, `Nvda.Launcher.exe`, `Nsight.Monitor.exe`, or `Nvda.CrashReporter.exe` is running. Make sure these programs are closed and try to install again.
- ▶ Peer access is disabled between two devices if either of them is in SLI mode.
- ▶ The following is a list of known CUDA 7.0 issues with the IBM POWER8 platform.
  - ▶ The POWER8 driver installation incorrectly overrides the mesa GL alternative. To access the mesa GL libraries, manually select the mesa GL alternative by running the following command: `sudo update-alternatives --config powerpc64le-linux-gnu_gl_conf`.
  - ▶ The `libcuvud` library isn't supported on POWER8.
  - ▶ NVIDIA video display driver and OpenGL implementations are not available on POWER8 at this time. Some samples may build and execute using alternative OpenGL implementations, but samples requiring NVIDIA's OpenGL will not work.
  - ▶ The IBM XL C/C++ compiler does not currently support OpenMP. When using that compiler as the host compiler with `nvcc`, samples that require OpenMP will fail to build.
  - ▶ The VDPAU API is not supported.
  - ▶ NVIDIA GPUDirect RDMA and GPUDirect P2P are not supported on POWER8, although they are not explicitly disabled.

## 7.2. CUDA Tools

### 7.2.1. CUDA Compiler

- ▶ When C++11 code (`-std=c++11`) is compiled on Linux with `gcc` as the host compiler, invoking `pow()` or `std::pow()` from device code with `(float, int)` or `(double, int)` arguments will not compile successfully. This will work from host functions or if the `-std=c++11` compilation option is not used. Developers can work around this issue by explicitly casting the second arguments to `float` or `double`, as appropriate.

### 7.2.2. CUDA Profiler

- ▶ The profiler may fail to collect events or metrics when **application replay** mode is turned on for an application that uses CUDA driver APIs to launch the kernel. In this case, use **kernel replay** mode.

## 7.3. CUDA Libraries

### 7.3.1. Thrust Library

- ▶ On the SLES 11 Linux distribution, there is a known issue that causes the `TestGetTemporaryBufferDispatchExplicit` and `TestGetTemporaryBufferDispatchImplicit` unit tests provided with the Thrust library to fail.
- ▶ On the SLES 11 Linux distribution, there is a known issue that causes the `segmentationTreeThrust` CUDA sample in the `6_Advanced` directory to fail.



## **Notice**

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

Information furnished is believed to be accurate and reliable. However, NVIDIA Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication of otherwise under any patent rights of NVIDIA Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all other information previously supplied. NVIDIA Corporation products are not authorized as critical components in life support devices or systems without express written approval of NVIDIA Corporation.

## **Trademarks**

NVIDIA and the NVIDIA logo are trademarks or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

## **Copyright**

© 2007-2015 NVIDIA Corporation. All rights reserved.