NVIDIA CUDA Toolkit

Release Notes for CUDA 11.5.1
Table of Contents

Chapter 1. CUDA 11.5 Release Notes.................................................................1
  1.1. CUDA Toolkit Major Component Versions........................................1
  1.2. General CUDA..................................................................................4
  1.3. CUDA Compilers.............................................................................5
  1.4. Resolved Issues...............................................................................5
    1.4.1. CUDA Compilers....................................................................5
  1.5. Deprecated Features.......................................................................6
  1.6. Known Issues..................................................................................6
    1.6.1. General CUDA........................................................................6

Chapter 2. CUDA Libraries...........................................................................7
  2.1. cuBLAS Library...............................................................................7
    2.1.1. cuBLAS: Release 11.4 Update 3.............................................7
    2.1.2. cuBLAS: Release 11.4 Update 2.............................................7
    2.1.3. cuBLAS: Release 11.4.........................................................8
    2.1.4. cuBLAS: Release 11.3 Update 1.............................................8
    2.1.5. cuBLAS: Release 11.3.........................................................9
    2.1.6. cuBLAS: Release 11.2.........................................................9
    2.1.7. cuBLAS: Release 11.1 Update 1..........................................9
    2.1.8. cuBLAS: Release 11.1.........................................................10
    2.1.9. cuBLAS: Release 11.0 Update 1.........................................10
    2.1.10. cuBLAS: Release 11.0......................................................12
    2.1.11. cuBLAS: Release 11.0 RC................................................12
  2.2. cuFFT Library.................................................................................13
    2.2.1. cuFFT: Release 11.5..........................................................13
    2.2.2. cuFFT: Release 11.4 Update 2..........................................13
    2.2.3. cuFFT: Release 11.4 Update 1..........................................14
    2.2.4. cuFFT: Release 11.4..........................................................14
    2.2.5. cuFFT: Release 11.3..........................................................14
    2.2.6. cuFFT: Release 11.2 Update 2..........................................15
    2.2.7. cuFFT: Release 11.2 Update 1..........................................15
    2.2.8. cuFFT: Release 11.2..........................................................15
    2.2.9. cuFFT: Release 11.1..........................................................16
    2.2.10. cuFFT: Release 11.0 RC................................................16
  2.3. cuRAND Library.............................................................................17
    2.3.1. cuRAND: Release 11.5 Update 1.....................................17
2.3.2. cuRAND: Release 11.3................................................................. 17
2.3.3. cuRAND: Release 11.0 Update 1............................................... 18
2.3.4. cuRAND: Release 11.0................................................................. 18
2.3.5. cuRAND: Release 11.0 RC......................................................... 18
2.4. cuSOLVER Library........................................................................... 18
2.4.1. cuSOLVER: Release 11.4............................................................... 18
2.4.2. cuSOLVER: Release 11.3............................................................... 19
2.4.3. cuSOLVER: Release 11.2 Update 2............................................ 19
2.4.4. cuSOLVER: Release 11.2............................................................... 19
2.4.5. cuSOLVER: Release 11.1 Update 1............................................ 19
2.4.6. cuSOLVER: Release 11.1............................................................... 20
2.4.7. cuSOLVER: Release 11.0............................................................... 21
2.5. cuSPARSE Library............................................................................ 21
2.5.1. cuSPARSE: Release 11.5 Update 1............................................. 21
2.5.2. cuSPARSE: Release 11.4 Update 1............................................. 22
2.5.3. cuSPARSE: Release 11.4............................................................... 22
2.5.4. cuSPARSE: Release 11.3 Update 1............................................. 22
2.5.5. cuSPARSE: Release 11.3............................................................... 23
2.5.6. cuSPARSE: Release 11.2 Update 2............................................. 23
2.5.7. cuSPARSE: Release 11.2 Update 1............................................. 24
2.5.8. cuSPARSE: Release 11.2............................................................... 25
2.5.9. cuSPARSE: Release 11.1 Update 1............................................. 25
2.5.10. cuSPARSE: Release 11.0............................................................. 25
2.5.11. cuSPARSE: Release 11.0 RC...................................................... 26
2.6. Math Library.................................................................................... 27
2.6.1. CUDA Math: Release 11.5........................................................... 27
2.6.2. CUDA Math: Release 11.4........................................................... 28
2.6.3. CUDA Math: Release 11.3........................................................... 28
2.6.4. CUDA Math: Release 11.1........................................................... 29
2.6.5. CUDA Math: Release 11.0 Update 1......................................... 29
2.6.6. CUDA Math: Release 11.0 RC................................................... 29
2.7. NVIDIA Performance Primitives (NPP)......................................... 29
2.7.1. NPP: Release 11.5........................................................ DISPATCHING 29
2.7.2. NPP: Release 11.4........................................................ DISPATCHING 30
2.7.3. NPP: Release 11.3........................................................ DISPATCHING 30
2.7.4. NPP: Release 11.2 Update 2....................................................... 30
2.7.5. NPP: Release 11.2 Update 1....................................................... 30
2.7.6. NPP: Release 11.0........................................................ DISPATCHING 31
2.7.7. NPP: Release 11.0 RC

2.8. nvJPEG Library

2.8.1. nvJPEG: Release 11.5 Update 1

2.8.2. nvJPEG: Release 11.4

2.8.3. nvJPEG: Release 11.2 Update 1

2.8.4. nvJPEG: Release 11.1 Update 1

2.8.5. nvJPEG: Release 11.0 Update 1

2.8.6. nvJPEG: Release 11.0

2.8.7. nvJPEG: Release 11.0 RC
List of Tables

Table 1. CUDA 11.5 Component Versions ...........................................................................................................1

Table 2. CUDA Toolkit and Minimum Required Driver Version for CUDA Minor Version Compatibility........................................................................................................................................3

Table 3. CUDA Toolkit and Corresponding Driver Versions ................................................................................3
Chapter 1. CUDA 11.5 Release Notes

The release notes for the NVIDIA® CUDA® Toolkit can be found online at http://docs.nvidia.com/cuda/cuda-toolkit-release-notes/index.html.

Note: The release notes have been reorganized into two major sections: the general CUDA release notes, and the CUDA libraries release notes including historical information for 11.x releases.

1.1. CUDA Toolkit Major Component Versions

CUDA Components
Starting with CUDA 11, the various components in the toolkit are versioned independently. For CUDA 11.5, the table below indicates the versions:

Table 1. CUDA 11.5 Component Versions

<table>
<thead>
<tr>
<th>Component Name</th>
<th>Version Information</th>
<th>Supported Architectures</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUDA Runtime (cudart)</td>
<td>11.5.117</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>cuobjdump</td>
<td>11.5.119</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUPTI</td>
<td>11.5.114</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA cuxfilt (demangler)</td>
<td>11.5.119</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA Demo Suite</td>
<td>11.5.50</td>
<td>x86_64</td>
</tr>
<tr>
<td>CUDA GDB</td>
<td>11.5.114</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA Memcheck</td>
<td>11.5.114</td>
<td>x86_64, POWER</td>
</tr>
<tr>
<td>CUDA Nsight</td>
<td>11.5.114</td>
<td>x86_64, POWER</td>
</tr>
<tr>
<td>CUDA NVCC</td>
<td>11.5.119</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA nvdisasm</td>
<td>11.5.119</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA NVML Headers</td>
<td>11.5.50</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA nvprof</td>
<td>11.5.114</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA nvprune</td>
<td>11.5.119</td>
<td>x86_64, POWER, Arm64</td>
</tr>
</tbody>
</table>
CUDA 11.5 Release Notes

<table>
<thead>
<tr>
<th>Component Name</th>
<th>Version Information</th>
<th>Supported Architectures</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUDA NVRTC</td>
<td>11.5.119</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA NVTX</td>
<td>11.5.114</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA NVVP</td>
<td>11.5.114</td>
<td>x86_64, POWER</td>
</tr>
<tr>
<td>CUDA Samples</td>
<td>11.5.56</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA Compute Sanitizer API</td>
<td>11.5.114</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA Thrust</td>
<td>11.5.62</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA cuBLAS</td>
<td>11.7.4.6</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA cuFFT</td>
<td>10.6.0.107</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA cuFile</td>
<td>1.1.1.25</td>
<td>x86_64</td>
</tr>
<tr>
<td>CUDA cuRAND</td>
<td>10.2.7.107</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA cuSOLVER</td>
<td>11.3.2.107</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA cuSPARSE</td>
<td>11.7.0.107</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA NPP</td>
<td>11.5.1.107</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>CUDA nvJPEG</td>
<td>11.5.4.107</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>Nsight Compute</td>
<td>2021.3.1.4</td>
<td>x86_64, POWER, Arm64 [CLI only]</td>
</tr>
<tr>
<td>NVTX</td>
<td>1.21018621</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>Nsight Systems</td>
<td>2021.3.3.2</td>
<td>x86_64, POWER, Arm64 [CLI only]</td>
</tr>
<tr>
<td>nvidia_fs&lt;sup&gt;1&lt;/sup&gt;</td>
<td>2.9.5</td>
<td>x86_64</td>
</tr>
<tr>
<td>Visual Studio Integration</td>
<td>11.5.114</td>
<td>x86_64 [Windows]</td>
</tr>
<tr>
<td>NVIDIA Linux Driver</td>
<td>495.29.05</td>
<td>x86_64, POWER, Arm64</td>
</tr>
<tr>
<td>NVIDIA Windows Driver</td>
<td>496.13</td>
<td>x86_64 [Windows]</td>
</tr>
</tbody>
</table>

**CUDA Driver**

Running a CUDA application requires the system with at least one CUDA capable GPU and a driver that is compatible with the CUDA Toolkit. See Table 3. For more information various GPU products that are CUDA capable, visit [https://developer.nvidia.com/cuda-gpus](https://developer.nvidia.com/cuda-gpus).

Each release of the CUDA Toolkit requires a minimum version of the CUDA driver. The CUDA driver is backward compatible, meaning that applications compiled against a particular version of the CUDA will continue to work on subsequent (later) driver releases.


**Note:** Starting with CUDA 11.0, the toolkit components are individually versioned, and the toolkit itself is versioned as shown in the table below.

---

<sup>1</sup> Only available on select Linux distros
The minimum required driver version for CUDA minor version compatibility is shown below. CUDA minor version compatibility is described in detail in https://docs.nvidia.com/deploy/cuda-compatibility/index.html

**Table 2. CUDA Toolkit and Minimum Required Driver Version for CUDA Minor Version Compatibility**

<table>
<thead>
<tr>
<th>CUDA Toolkit</th>
<th>Minimum Required Driver Version for CUDA Minor Version Compatibility*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linux x86_64 Driver Version</td>
</tr>
<tr>
<td>CUDA 11.5.x</td>
<td>&gt;=450.80.02</td>
</tr>
<tr>
<td>CUDA 11.4.x</td>
<td>&gt;=450.80.02</td>
</tr>
<tr>
<td>CUDA 11.3.x</td>
<td>&gt;=450.80.02</td>
</tr>
<tr>
<td>CUDA 11.2.x</td>
<td>&gt;=450.80.02</td>
</tr>
<tr>
<td>CUDA 11.1 [11.1.0]</td>
<td>&gt;=450.80.02</td>
</tr>
<tr>
<td>CUDA 11.0 [11.0.3]</td>
<td>&gt;=450.36.06**</td>
</tr>
</tbody>
</table>

* Using a Minimum Required Version that is different from Toolkit Driver Version could be allowed in compatibility mode -- please read the CUDA Compatibility Guide for details.

** CUDA 11.0 was released with an earlier driver version, but by upgrading to drivers 450.80.02 (Linux) / 452.39 (Windows), minor version compatibility is possible across the CUDA 11.x family of toolkits.

The version of the development NVIDIA GPU Driver packaged in each CUDA Toolkit release is shown below.

**Table 3. CUDA Toolkit and Corresponding Driver Versions**

<table>
<thead>
<tr>
<th>CUDA Toolkit</th>
<th>Toolkit Driver Version</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linux x86_64 Driver Version</td>
</tr>
<tr>
<td>CUDA 11.5 Update 1</td>
<td>&gt;=495.29.05</td>
</tr>
<tr>
<td>CUDA 11.5 GA</td>
<td>&gt;=495.29.05</td>
</tr>
<tr>
<td>CUDA 11.4 Update 3</td>
<td>&gt;=470.82.01</td>
</tr>
<tr>
<td>CUDA 11.4 Update 2</td>
<td>&gt;=470.57.02</td>
</tr>
<tr>
<td>CUDA 11.4 Update 1</td>
<td>&gt;=470.57.02</td>
</tr>
<tr>
<td>CUDA 11.4.0 GA</td>
<td>&gt;=470.42.01</td>
</tr>
<tr>
<td>CUDA 11.3.1 Update 1</td>
<td>&gt;=465.19.01</td>
</tr>
<tr>
<td>CUDA 11.3.0 GA</td>
<td>&gt;=465.19.01</td>
</tr>
<tr>
<td>CUDA 11.2.2 Update 2</td>
<td>&gt;=460.32.03</td>
</tr>
<tr>
<td>CUDA 11.2.1 Update 1</td>
<td>&gt;=460.32.03</td>
</tr>
<tr>
<td>CUDA Toolkit</td>
<td>Linux x86_64 Driver Version</td>
</tr>
<tr>
<td>--------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>CUDA 11.2.0 GA</td>
<td>&gt;=460.27.03</td>
</tr>
<tr>
<td>CUDA 11.1.1 Update 1</td>
<td>&gt;=455.32</td>
</tr>
<tr>
<td>CUDA 11.1 GA</td>
<td>&gt;=455.23</td>
</tr>
<tr>
<td>CUDA 11.0.3 Update 1</td>
<td>&gt;= 450.51.06</td>
</tr>
<tr>
<td>CUDA 11.0.2 GA</td>
<td>&gt;= 450.51.05</td>
</tr>
<tr>
<td>CUDA 11.0.1 RC</td>
<td>&gt;= 450.36.06</td>
</tr>
<tr>
<td>CUDA 10.2.89</td>
<td>&gt;= 440.33</td>
</tr>
<tr>
<td>CUDA 10.1 [10.1.105 general release, and updates]</td>
<td>&gt;= 418.39</td>
</tr>
<tr>
<td>CUDA 10.0.130</td>
<td>&gt;= 410.48</td>
</tr>
<tr>
<td>CUDA 9.2 [9.2.148 Update 1]</td>
<td>&gt;= 396.37</td>
</tr>
<tr>
<td>CUDA 9.2 [9.2.88]</td>
<td>&gt;= 396.26</td>
</tr>
<tr>
<td>CUDA 9.1 [9.1.85]</td>
<td>&gt;= 390.46</td>
</tr>
<tr>
<td>CUDA 9.0 [9.0.76]</td>
<td>&gt;= 384.81</td>
</tr>
<tr>
<td>CUDA 8.0 [8.0.61 GA2]</td>
<td>&gt;= 375.26</td>
</tr>
<tr>
<td>CUDA 8.0 [8.0.44]</td>
<td>&gt;= 367.48</td>
</tr>
<tr>
<td>CUDA 7.5 [7.5.16]</td>
<td>&gt;= 352.31</td>
</tr>
<tr>
<td>CUDA 7.0 [7.0.28]</td>
<td>&gt;= 346.46</td>
</tr>
</tbody>
</table>

For convenience, the NVIDIA driver is installed as part of the CUDA Toolkit installation. Note that this driver is for development purposes and is not recommended for use in production with Tesla GPUs.

For running CUDA applications in production with Tesla GPUs, it is recommended to download the latest driver for Tesla GPUs from the NVIDIA driver downloads site at http://www.nvidia.com/drivers.

During the installation of the CUDA Toolkit, the installation of the NVIDIA driver may be skipped on Windows (when using the interactive or silent installation) or on Linux (by using meta packages).

For more information on customizing the install process on Windows, see http://docs.nvidia.com/cuda/cuda-installation-guide-microsoft-windows/index.html#install-cuda-software.

For meta packages on Linux, see https://docs.nvidia.com/cuda/cuda-installation-guide-linux/index.html#package-manager-metas

1.2. General CUDA
11.5
- Device-side caching behavior is now configurable with annotated pointers.
- Prefix sums (scans) for cooperative groups: Added four new functions for inclusive and exclusive scans.
- Support for `NvSciBufGeneralAttrKey_EnableGpuCompression` in CUDA will be available in r470TRD2.
- Preview release of a new data type, `__int128`, usable with compatible host compilers. As it is a preview, there is no broad support for math operations, library support, dev tools, and so on.
- Added native support for signed and unsigned normalized 8- and 16-bit types.
- Improved interoperability with graphics frameworks: Added support for normalized integer and block-compressed data types.
- For multi-process sharing of GPUs, CUDA now supports per-process memory access policies.
- Linking is supported with cubins larger than 2 GB.
- GSP-RM is enabled as opt-in for Turing+ Tesla GPUs.
- GA release of CUDA Python.

### 1.3. CUDA Compilers

11.5

- CUDA now provides a static version of the NVRTC library.
- `builtin_assume` can now be used to specify address space to allow for efficient loads and stores.
- Added support for numerous pragmas that offer more control over diagnostic messages.
- To generate code for multiple architectures at the same time, use the `--arch=all` or `--arch=all-major` options.
- Deterministic code generation: compiled bits will not change between `nvcc` invocations without semantic source code changes.

### 1.4. Resolved Issues

#### 1.4.1. CUDA Compilers

11.5. Update 1

- When using the `--fmad=false` compiler option, even the explicitly requested fused multiply-add instructions were decomposed into separate multiply and add, leading to loss
One of the consequences was that CUDA Math APIs could not be trusted to deliver correct results; worst case errors became unbounded. This issue was introduced in 11.5, and is now resolved.

- Fixed a compiler optimization bug that may move memory access instructions across memory barriers that may lead to incorrect runtime results with certain synchronization dependencies.
- An issue in the PTX optimizer sometimes produced incorrect results. This issue is resolved.

11.5

- Linking with cubins larger than 2 GB is now supported.
- Certain C++17 features that were backported to C++14 in MSVC are now supported.
- An issue with the use of lambda function when an object is passed-by-value is resolved. [https://github.com/Ahdhn/nvcc_bug_maybe](https://github.com/Ahdhn/nvcc_bug_maybe)

### 1.5. 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.

**General CUDA**

- NVIDIA Driver support for Kepler is removed beginning with R495. CUDA Toolkit development support for Kepler continues through CUDA 11.x.

### 1.6. Known Issues

#### 1.6.1. General CUDA

- Intermittent crashes were seen when CUDA binaries were running on a system with a GLIBC version older than 2.17-106.el7_2.1. This is due to a known bug in older versions of GLIBC (Bug reference: [https://bugzilla.redhat.com/show_bug.cgi?id=1293976](https://bugzilla.redhat.com/show_bug.cgi?id=1293976)) and has been fixed in later versions (>= glibc-2.17-107.el7).
Chapter 2. CUDA Libraries

This section covers CUDA Libraries release notes for 11.x releases.

- CUDA Math Libraries toolchain uses C++11 features, and a C++11-compatible standard library (libstdc++ >= 20150422) is required on the host.
- CUDA Math libraries are no longer shipped for SM30 and SM32.
- Support for the following compute capabilities are deprecated for all libraries:
  - sm_35 (Kepler)
  - sm_37 (Kepler)
  - sm_50 (Maxwell)

2.1. cuBLAS Library

2.1.1. cuBLAS: Release 11.4 Update 3

- Resolved Issues
  - Some cublas and cublasLt functions sometimes returned CUBLAS_STATUS_EXECUTION_FAILED if the dynamic library was loaded and unloaded several times during application lifetime within the same CUDA context. This issue has been resolved.

2.1.2. cuBLAS: Release 11.4 Update 2

- New Features
  - Vector (and batched) alpha support for per-row scaling in TN int32 math Matmul with int8 output. See CUBLASLT_POINTER_MODE_ALPHA_DEVICE_VECTOR_BETA_HOST and CUBLASLT_MATMUL_DESC_ALPHA_VECTOR_BATCHStride.
  - New epilogue options have been added to support fusion in DLtraining: CUBLASLT_EPILOGUE_BGRADA and CUBLASLT_EPILOGUE_BGRADB which compute bias gradients based on matrices A and B respectively.
CUDA Libraries

New auxiliary functions `cublasGetStatusName()`, `cublasGetStatusString()` have been added to cuBLAS that return the string representation and the description of the cuBLAS status (`cublasStatus_t`) respectively. Similarly, `cublasLtGetStatusName()`, `cublasLtGetStatusString()` have been added to cuBlasLt.

Known Issues

- `cublasGemmBatchedEx()` and `cublas<t>gemmBatched()` check the alignment of the input/output arrays of the pointers like they were the pointers to the actual matrices. These checks are irrelevant and will be disabled in future releases. This mostly affects half-precision inputGEMMs which might require 16-byte alignment, and array of pointers could only be aligned by 8-byte boundary.

Resolved Issues

- `cublasLtMatrixTransform` can now operate on matrices with dimensions greater than 65535.
- Fixed out-of-bound access in GEMM and Matmul functions, when split K or non-default epilogue is used and leading dimension of the output matrix exceeds `int32_t` limit.
- NVBLAS now uses lazy loading of the CPU BLAS library on Linux to avoid issues caused by preloading `libnvblas.so` in complex applications that use `fork` and similar APIs.
- Resolved symbols name conflict when using cuBlasLt static library with static TensorRT or cuDNN libraries.

2.1.3. cuBLAS: Release 11.4

Resolved Issues

- Some gemv cases were producing incorrect results if the matrix dimension (n or m) was large, for example $2^{20}$.

2.1.4. cuBLAS: Release 11.3 Update 1

New Features

- Some new kernels have been added for improved performance but have the limitation that only host pointers are supported for scalars (for example, alpha and beta parameters). This limitation is expected to be resolved in a future release.
- New epilogues have been added to support fusion in ML training. These include:
  - ReLuBias and GeluBias epilogues that produce an auxiliary output which is used on backward propagation to compute the corresponding gradients.
  - DReLuBGrad and DGeluBGrad epilogues that compute the backpropagation of the corresponding activation function on matrix C, and produce bias gradient as a separate output. These epilogues require auxiliary input mentioned in the bullet above.

Resolved Issues
CUDA Libraries

Some tensor core accelerated strided batched GEMM routines would result in misaligned memory access exceptions when batch stride wasn’t a multiple of 8.

Tensor core accelerated cublasGemmBatchedEx [pointer-array] routines would use slower variants of kernels assuming bad alignment of the pointers in the pointer array. Now it assumes that pointers are well aligned, as noted in the documentation.

**Known Issues**

To be able to access the fastest possible kernels through `cublasLtMatmulAlgoGetHeuristic()` you need to set `CUBLASLT_MATMUL_PREF_POINTER_MODE_MASK` in search preferences to `CUBLASLT_POINTER_MODE_MASK_HOST` or `CUBLASLT_POINTER_MODE_MASK_NO_FILTERING`. By default, heuristics query assumes the pointer mode may change later and only returns algo configurations that support both _HOST and _DEVICE modes. Without this, newly added kernels will be excluded and it will likely lead to a performance penalty on some problem sizes.

**Deprecated Features**

Linking with static cublas and cublasLt libraries on Linux now requires using gcc-5.2 and compatible or higher due to C++11 requirements in these libraries.

2.1.5. **cuBLAS: Release 11.3**

**Known Issues**

The planar complex matrix descriptor for batched matmul has inconsistent interpretation of batch offset.

Mixed precision operations with reduction scheme `CUBLASLT_REDUCTION_SCHEME_OUTPUT_TYPE` (might be automatically selected based on problem size by `cublasSgemmEx()` or `cublasGemmEx()` too, unless `CUBLAS_MATH_DISALLOW_REDUCED_PRECISION_REDUCTION` math mode bit is set) not only stores intermediate results in output type but also accumulates them internally in the same precision, which may result in lower than expected accuracy. Please use `CUBLASLT_MATMUL_PREF_REDUCTION_SCHEME_MASK` or `CUBLAS_MATH_DISALLOW_REDUCED_PRECISION_REDUCTION` if this results in numerical precision issues in your application.

2.1.6. **cuBLAS: Release 11.2**

**Known Issues**

cublas<s/d/c/z>Gemm() with very large n and m=k=1 may fail on Pascal devices.

2.1.7. **cuBLAS: Release 11.1 Update 1**

**New Features**

cuBLASLt Logging is officially stable and no longer experimental. cuBLASLt Logging APIs are still experimental and may change in future releases.
Resolved Issues

- cublasLt Matmul fails on Volta architecture GPUs with CUBLAS_STATUS_EXECUTION_FAILED when n dimension > 262,137 and epilogue bias feature is being used. This issue exists in 11.0 and 11.1 releases but has been corrected in 11.1 Update 1

2.1.8. cuBLAS: Release 11.1

- Resolved Issues
  - A performance regression in the cublasCgetrfBatched and cublasCgetriBatched routines has been fixed.
  - The IMMA kernels do not support padding in matrix C and may corrupt the data when matrix C with padding is supplied to cublasLtMatmul. A suggested work around is to supply matrix C with leading dimension equal to 32 times the number of rows when targeting the IMMA kernels: computeType = CUDA_R_32I and CUBLASLT_ORDER_COL32 for matrices A,C,D, and CUBLASLT_ORDER_COL4_4R2_8C (on NVIDIA Ampere GPU architecture or Turing architecture) or CUBLASLT_ORDER_COL32_2R_4R4 (on NVIDIA Ampere GPU architecture) for matrix B. Matmul descriptor must specify CUBLAS_OP_T on matrix B and CUBLAS_OP_N (default) on matrix A and C. The data corruption behavior was fixed so that CUBLAS_STATUS_NOT_SUPPORTED is returned instead.
  - Fixed an issue that caused an Address out of bounds error when calling cublasSgemm().
  - A performance regression in the cublasCgetrfBatched and cublasCgetriBatched routines has been fixed.

2.1.9. cuBLAS: Release 11.0 Update 1

- New Features
  - The cuBLAS API was extended with a new function, cublasSetWorkspace(), which allows the user to set the cuBLAS library workspace to a user-owned device buffer, which will be used by cuBLAS to execute all subsequent calls to the library on the currently set stream.
  - cuBLASLt experimental logging mechanism can be enabled in two ways:
    - By setting the following environment variables before launching the target application:
      - CUBLASLT_LOG_LEVEL=<level> -- where level is one of the following levels:
        - “0” - Off - logging is disabled (default)
        - “1” - Error - only errors will be logged
        - “2” - Trace - API calls that launch CUDA kernels will log their parameters and important information
CUDA Libraries

- "3" - Hints - hints that can potentially improve the application's performance
- "4" - Heuristics - heuristics log that may help users to tune their parameters
- "5" - API Trace - API calls will log their parameter and important information

CUBLASLT_LOG_MASK=<mask> -- where mask is a combination of the following masks:
- "0" - Off
- "1" - Error
- "2" - Trace
- "4" - Hints
- "8" - Heuristics
- "16" - API Trace

CUBLASLT_LOG_FILE=<value> -- where value is a file name in the format of ", %i", %i will be replaced with process id. If CUBLASLT_LOG_FILE is not defined, the log messages are printed to stdout.

By using the runtime API functions defined in the cublasLt header:

typedef void(*cublasLtLoggerCallback_t)(int logLevel, const char* functionName, const char* message) -- A type of callback function pointer.

cublasStatus_t cublasLtLoggerSetCallback(cublasLtLoggerCallback_t callback) -- Allows to set a call back functions that will be called for every message that is logged by the library.

cublasStatus_t cublasLtLoggerSetFile(FILE* file) -- Allows to set the output file for the logger. The file must be open and have write permissions.

cublasStatus_t cublasLtLoggerOpenFile(const char* logFile) -- Allows to give a path in which the logger should create the log file.

cublasStatus_t cublasLtLoggerSetLevel(int level) -- Allows to set the log level to one of the above mentioned levels.

cublasStatus_t cublasLtLoggerSetMask(int mask) -- Allows to set the log mask to a combination of the above mentioned masks.

cublasStatus_t cublasLtLoggerForceDisable() -- Allows to disable to logger for the entire session. Once this API is being called, the logger cannot be reactivated in the current session.
2.1.10. cuBLAS: Release 11.0

- **New Features**
  - cuBLASLt Matrix Multiplication adds support for fused ReLU and bias operations for all floating point types except double precision (FP64).
  - Improved batched TRSM performance for matrices larger than 256.

2.1.11. cuBLAS: Release 11.0 RC

- **New Features**
  - Many performance improvements have been implemented for NVIDIA Ampere, Volta, and Turing Architecture based GPUs.
  - The cuBLASLt logging mechanism can be enabled by setting the following environment variables before launching the target application:
    - `CUBLASLT_LOG_LEVEL=<level>` - while level is one of the following levels:
      - "0" - Off - logging is disabled (default)
      - "1" - Error - only errors will be logged
      - "2" - Trace - API calls will be logged with their parameters and important information
    - `CUBLASLT_LOG_FILE=<value>` - while value is a file name in the format of "<file_name>.%i", %i will be replaced with process id. If `CUBLASLT_LOG_FILE` is not defined, the log messages are printed to stdout.
  - For matrix multiplication APIs:
    - `cublasGemmEx, cublasGemmBatchedEx, cublasGemmStridedBatchedEx` and `cublasLtMatmul` added new data type support for __nv_bfloat16 (CUDA_R_16BF).
    - A new compute type TensorFloat32 (TF32) has been added to provide tensor core acceleration for FP32 matrix multiplication routines with full dynamic range and increased precision compared to BFLOAT16.
    - New compute modes Default, Pedantic, and Fast have been introduced to offer more control over compute precision used.
    - Tensor cores are now enabled by default for half-, and mixed-precision- matrix multiplications.
    - Double precision tensor cores (DMMA) are used automatically.
    - Tensor cores can now be used for all sizes and data alignments and for all GPU architectures.
Selection of these kernels through cuBLAS heuristics is automatic and will depend on factors such as math mode setting as well as whether it will run faster than the non-tensor core kernels.

Users should note that while these new kernels that use tensor cores for all unaligned cases are expected to perform faster than non-tensor core based kernels but slower than kernels that can be run when all buffers are well aligned.

**Deprecated Features**

- Algorithm selection in `cublasGemmEx` APIs (including batched variants) is non-functional for NVIDIA Ampere Architecture GPUs. Regardless of selection it will default to a heuristics selection. Users are encouraged to use the `cublasLt` APIs for algorithm selection functionality.

- The matrix multiply math mode `CUBLAS_TENSOR_OP_MATH` is being deprecated and will be removed in a future release. Users are encouraged to use the new `cublasComputeType_t` enumeration to define compute precision.

### 2.2. cuFFT Library

#### 2.2.1. cuFFT: Release 11.5

**Known Issues**

- FFTs of certain sizes in single and double precision (multiples of size 6) could fail on future devices. This issue will be fixed in an upcoming release.

#### 2.2.2. cuFFT: Release 11.4 Update 2

**Resolved Issues**

- Since cuFFT 10.3.0 (CUDA Toolkit 11.1), cuFFT may require user to make sure that all operations on input and output buffers are complete before calling `cufft[Xd]Exec*` if:
  - Sm70 or later, 3D FFT, batch > 1, total size of transform is greater than 4.5MB
  - FFT size for all dimensions is in the set of the following sizes: \{2, 4, 8, 16, 32, 64, 128, 3, 9, 81, 243, 729, 2187, 6561, 5, 25, 125, 625, 3125, 6, 36, 216, 1296, 7776, 7, 49, 343, 2401, 11, 121\}
  - Some V100 FFTs were slower than expected. This issue is resolved.

**Known Issues**

- Some T4 FFTs are slower than expected.

- Plans for FFTs of certain sizes in single precision (including some multiples of 1024 sizes, and some large prime numbers) could fail on future devices with less than 64 kB of shared memory. This issue will be fixed in an upcoming release.
2.2.3. cuFFT: Release 11.4 Update 1

Resolved Issues

- Some cuFFT multi-GPU plans exhibited very long creation times.
- cuFFT sometimes produced incorrect results for real-to-complex and complex-to-real transforms when the total number of elements across all batches in a single execution exceeded 2147483647.

Known Issues

- Some V100 FFTs are slower than expected.
- Some T4 FFTs are slower than expected.

2.2.4. cuFFT: Release 11.4

New Features

- Performance improvements.

Known Issues

- Some T4 FFTs are slower than expected.
- cuFFT may produce incorrect results for real-to-complex and complex-to-real transforms when the total number of elements across all batches in a single execution exceeds 2147483647.
- Some cuFFT multi-GPU plans may exhibit very long creation time. Issue will be fixed in the next update.
- cuFFT may produce incorrect results for transforms with strides when the index of the last element across all batches exceeds 2147483647 [see Advanced Data Layout].

Deprecated Features

- Support for callback functionality using separately compiled device code is deprecated on all GPU architectures. Callback functionality will continue to be supported for all GPU architectures.

2.2.5. cuFFT: Release 11.3

New Features

- cuFFT shared libraries are now linked statically against libstdc++ on Linux platforms.
- Improved performance of certain sizes (multiples of large powers of 3, powers of 11) in SM86.

Known Issues
CUDA Libraries

- cuFFT planning and plan estimation functions may not restore correct context affecting CUDA driver API applications.
- Plans with strides, primes larger than 127 in FFT size decomposition and total size of transform including strides bigger than 32GB produce incorrect results.

2.2.6. cuFFT: Release 11.2 Update 2

- Known Issues
  - cuFFT planning and plan estimation functions may not restore correct context affecting CUDA driver API applications.
  - Plans with strides, primes larger than 127 in FFT size decomposition and total size of transform including strides bigger than 32GB produce incorrect results.

2.2.7. cuFFT: Release 11.2 Update 1

- Resolved Issues
  - Previously, reduced performance of power-of-2 single precision FFTs was observed on GPUs with sm_86 architecture. This issue has been resolved.
  - Large prime factors in size decomposition and real to complex or complex to real FFT type no longer cause cuFFT plan functions to fail.

- Known Issues
  - cuFFT planning and plan estimation functions may not restore correct context affecting CUDA driver API applications.
  - Plans with strides, primes larger than 127 in FFT size decomposition and total size of transform including strides bigger than 32GB produce incorrect results.

2.2.8. cuFFT: Release 11.2

- New Features
  - Multi-GPU plans can be associated with a stream using the `cufftSetStream` API function call.
  - Performance improvements for R2C/C2C/C2R transforms.
  - Performance improvements for multi-GPU systems.

- Resolved Issues
  - cuFFT is no longer stuck in a bad state if previous plan creation fails with `CUFFT_ALLOC_FAILED`.
  - Previously, single dimensional multi-GPU FFT plans ignored user input on `cufftXtSetGPUSwhichGPUS` argument and assumed that GPUs IDs are always numbered from 0 to N-1. This issue has been resolved.
- Plans with primes larger than 127 in FFT size decomposition or FFT size being a prime number bigger than 4093 do not perform calculations on second and subsequent `cufftExecute*` calls. The regression was introduced in cuFFT 11.1.

### Known Issues
- cuFFT planning and plan estimation functions may not restore correct context affecting CUDA driver API applications.

#### 2.2.9. cuFFT: Release 11.1

- **New Features**
  - cuFFT is now L2-cache aware and uses L2 cache for GPUs with more than 4.5MB of L2 cache. Performance may improve in certain single-GPU 3D C2C FFT cases.
  - After successfully creating a plan, cuFFT now enforces a lock on the cufftHandle. Subsequent calls to any planning function with the same cufftHandle will fail.
  - Added support for very large sizes [3k cube] to multi-GPU cuFFT on DGX-2.
  - Improved performance on multi-gpu cuFFT for certain sizes [1k cube].

- **Resolved Issues**
  - Resolved an issue that caused cuFFT to crash when reusing a handle after clearing a callback.
  - Fixed an error which produced incorrect results / NaN values when running a real-to-complex FFT in half precision.

- **Known Issues**
  - cuFFT will always overwrite the input for out-of-place C2R transform.
  - Single dimensional multi-GPU FFT plans ignore user input on the `whichGPUs` parameter of `cufftXtSetGPUs()` and assume that GPUs IDs are always numbered from 0 to N-1.

#### 2.2.10. cuFFT: Release 11.0 RC

- **New Features**
  - cuFFT now accepts `__nv_bfloat16` input and output data type for power-of-two sizes with single precision computations within the kernels.
  - Reoptimized power of 2 FFT kernels on Volta and Turing architectures.

- **Resolved Issues**
  - Reduced R2C/C2R plan memory usage to previous levels.
  - Resolved bug introduced in 10.1 update 1 that caused incorrect results when using custom strides, batched 2D plans and certain sizes on Volta and later.
2.3. cuRAND Library

2.3.1. cuRAND: Release 11.5 Update 1

- Known Issues
  - cuFFT modifies C2R input buffer for some non-strided FFT plans.
  - There is a known issue with certain cuFFT plans that causes an assertion in the execution phase of certain plans. This applies 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.

- New Features
  - Improved performance of CURAND_RNG_PSEUDO_MRG32K3A pseudorandom number generator when using ordering CURAND_ORDERING_PSEUDO_DEFAULT or CURAND_ORDERING_PSEUDO_BEST.
  - Added a new type of order parameter: CURAND_ORDERING_PSEUDO_DYNAMIC.
  - Supported PRNGs:
    - CURAND_RNG_PSEUDO_XORWOW
    - CURAND_RNG_PSEUDO_MRG32K3A
    - CURAND_RNG_PSEUDO_MTGP32
    - CURAND_RNG_PSEUDO_PHILOX4_32_10
  - Improved performance compared to CURAND_ORDERING_PSEUDO_DEFAULT, especially on NVIDIA Ampere architecture GPUs.
  - The output ordering of generated random numbers for CURAND_ORDERING_PSEUDO_DYNAMIC depends on the number of SMs on a GPU, and thus can be different on different GPUs.
  - The CURAND_ORDERING_PSEUDO_DYNAMIC ordering can’t be used with a host generator created using curandCreateGeneratorHost().

- Resolved Issues
  - Added information about cuRAND thread safety.

- Known Issues
  - CURAND_RNG_PSEUDO_XORWOW with ordering CURAND_ORDERING_PSEUDO_DYNAMIC can produce incorrect results on architectures newer than SM86.

2.3.2. cuRAND: Release 11.3

- Resolved Issues
Fixed inconsistency between random numbers generated by GPU and host generators when CURAND_ORDERING_PSEUDO_LEGACY ordering is selected for certain generator types.

### 2.3.3. cuRAND: Release 11.0 Update 1

- **Resolved Issues**
  - Fixed an issue that caused linker errors about the multiple definitions of mtgp32dc_params_fast_11213 and mtgpdc_params_11213_num when including curand_mtgp32dc_p_11213.h in different compilation units.

### 2.3.4. cuRAND: Release 11.0

- **Resolved Issues**
  - Fixed an issue that caused linker errors about the multiple definitions of mtgp32dc_params_fast_11213 and mtgpdc_params_11213_num when including curand_mtgp32dc_p_11213.h in different compilation units.

### 2.3.5. cuRAND: Release 11.0 RC

- **Resolved Issues**
  - Introduced CURAND_ORDERING_PSEUDO_LEGACY ordering. Starting with CUDA 10.0, the ordering of random numbers returned by MTGP32 and MRG32k3a generators are no longer the same as previous releases despite being guaranteed by the documentation for the CURAND_ORDERING_PSEUDO_DEFAULT setting. The CURAND_ORDERING_PSEUDO_LEGACY provides pre-CUDA 10.0 ordering for MTGP32 and MRG32k3a generators.
  - Starting with CUDA 11.0 CURAND_ORDERING_PSEUDO_DEFAULT is the same as CURAND_ORDERING_PSEUDO_BEST for all generators except MT19937. Only CURAND_ORDERING_PSEUDO_LEGACY is guaranteed to provide the same for all future cuRAND releases.

### 2.4. cuSOLVER Library

#### 2.4.1. cuSOLVER: Release 11.4

- **New Features**
  - Introducing cusolverDnXtrtri, a new generic API for triangular matrix inversion (trtri).
  - Introducing cusolverDnXsytrs, a new generic API for solving systems of linear equations using a given factorized symmetric matrix from SYTRF.
2.4.2. cuSOLVER: Release 11.3

Known Issues

- For values \( N \leq 16 \), \texttt{cusolverDn[S|D|C|Z]syevjBatched} hits out-of-bound access and may deliver the wrong result. The workaround is to pad the matrix \( A \) with a diagonal matrix \( D \) such that the dimension of \([A \ 0 \ 0 \ D] \) is bigger than 16. The diagonal entry \( D_{jj} \) must be bigger than maximum eigenvalue of \( A \), for example, \( \text{norm}(A, \text{‘fro’}) \). After the \texttt{syevj}, \( W(0:n-1) \) contains the eigenvalues and \( A(0:n-1,0:n-1) \) contains the eigenvectors.

2.4.3. cuSOLVER: Release 11.2 Update 2

New Features

- New singular value decomposition (GESVDR) is added. GESVDR computes partial spectrum with random sampling, an order of magnitude faster than GESVD.
- \texttt{libcusolver.so} no longer links \texttt{libcublas_static.a}; instead, it depends on \texttt{libcublas.so}. This reduces the binary size of \texttt{libcusolver.so}. However, it breaks backward compatibility. The user has to link \texttt{libcusolver.so} with the correct version of \texttt{libcublas.so}.

2.4.4. cuSOLVER: Release 11.2

Resolved Issues

- \texttt{cusolverDnIRSXgels} sometimes returned \texttt{CUSOLVER_STATUS_INTERNAL_ERROR} when the precision is ‘z’. This issue has been fixed in CUDA 11.2; now \texttt{cusolverDnIRSXgels} works for all precisions.
- \texttt{ZSYTRF} sometimes returned \texttt{CUSOLVER_STATUS_INTERNAL_ERROR} due to insufficient resources to launch the kernel. This issue has been fixed in CUDA 11.2.
- \texttt{GETRF} returned early without finishing the whole factorization when the matrix was singular. This issue has been fixed in CUDA 11.2.

2.4.5. cuSOLVER: Release 11.1 Update 1

Resolved Issues

- \texttt{cusolverDnDDgels} reports \texttt{IRS_NOT_SUPPORTED} when \( m > n \). The issue has been fixed in release 11.1 U1, so \texttt{cusolverDnDDgels} will support \( m > n \).
- \texttt{cusolverMgDeviceSelect} can consume over 1GB device memory. The issue has been fixed in release 11.1 U1. The hidden memory allocation inside \texttt{cusolverMG} handle is about 30 MB per device.

Known Issues

- \texttt{cusolverDnIRSXgels} may return \texttt{CUSOLVER_STATUS_INTERNAL_ERROR} when the precision is ‘z’ due to insufficient workspace which causes illegal memory access.
The `cusolverDnIRSXgels_bufferSize()` does not report the correct size of workspace. To workaround the issue, the user has to add more workspace than what is reported by `cusolverDnIRSXgels_bufferSize()`. For example, if `x` is the size of workspace returned by `cusolverDnIRSXgels_bufferSize()`, then the user has to allocate `(x + \min(m,n) \times \text{sizeof}(\text{cuDoubleComplex}))` bytes.

### 2.4.6. cuSOLVER: Release 11.1

- **New Features**
  - Added new 64-bit APIs:
    - `cusolverDnXpotrf_bufferSize`
    - `cusolverDnXpotrf`
    - `cusolverDnXpotrs`
    - `cusolverDnXgeqrf_bufferSize`
    - `cusolverDnXgeqrf`
    - `cusolverDnXgetrf_bufferSize`
    - `cusolverDnXgetrf`
    - `cusolverDnXgetrs`
    - `cusolverDnXsyevd_bufferSize`
    - `cusolverDnXsyevd`
    - `cusolverDnXsyevdx_bufferSize`
    - `cusolverDnXsyevdx`
    - `cusolverDnXgesvd_bufferSize`
    - `cusolverDnXgesvd`
    - `cusolverDnXgesvdx`
  - Added a new SVD algorithm based on polar decomposition, called GESVDP which uses the new 64-bit API, including `cusolverDnXgesvdp_bufferSize` and `cusolverDnXgesvdp`.

- **Deprecated Features**
  - The following 64-bit APIs are deprecated:
    - `cusolverDnPotrf_bufferSize`
    - `cusolverDnPotrf`
    - `cusolverDnPotrs`
2.4.7. cuSOLVER: Release 11.0

- **New Features**
  - Add 64-bit API of GESVD. The new routine `cusolverDnGesvd_bufferSize()` fills the missing parameters in 32-bit API `cusolverDn[S|D|C|Z]gesvd_bufferSize()` such that it can estimate the size of the workspace accurately.
  - Added the single process multi-GPU Cholesky factorization capabilities POTRF, POTRS and POTRI in cuSolverMG library.

- **Resolved Issues**
  - Fixed an issue where SYEVD/SYGVD would fail and return error code 7 if the matrix is zero and the dimension is bigger than 25.

2.5. cuSPARSE Library

2.5.1. cuSPARSE: Release 11.5 Update 1

- **New Features**

- **Resolved Issues**
- Added memory requirements, graph capture, and asynchronous notes for cusparseXcsrsm2_analysis.
- CSR, CSC, and COO format descriptions wrongly reported sorted column indices requirement. All routines support unsorted column indices, except where strictly indicated
- Clarified cusparseSpSV and cusparseSpSM memory management.
- cusparseSpSM produced wrong results in some cases when the matB operation is CUSPARSE_OPERATION_NON_TRANSPOSE or CUSPARSE_OPERATION_CONJUGATE_TRANSPOSE.
- cusparseSpSM produced wrong results in some cases when the matrix layout is row-major.

2.5.2. cuSPARSE: Release 11.4 Update 1
- Resolved Issues
  - cusparseSpSV and cusparseSpSM could produce wrong results
  - cusparseSpSV and cusparseSpSM did not work correctly when vecX == vecY or matB == matC.

2.5.3. cuSPARSE: Release 11.4
- Known Issues
  - cusparseSpSV and cusparseSpSM could produce wrong results
  - cusparseSpSV and cusparseSpSM do not work correctly when vecX == vecY or matB == matC.

2.5.4. cuSPARSE: Release 11.3 Update 1
- New Features
  - Introduced a new routine for sparse matrix - sparse matrix multiplication (cusparseSpGEMMreuse) where the output matrix structure is reused for multiple computation. The new routine supports CSR storage format and mixed-precision computation.
  - Sparse triangular solver adds support for COO format.
  - Introduced a new routine for sparse triangular solver with multiple right-hand sides cusparseSpSM().
  - cusparseDenseToSparse() routine adds the conversion from dense matrix [row-major/column-major] to Blocked-ELL format.
  - Blocke-ELL format now support empty blocks
CUDA Libraries

- Better performance for Blocked-ELL SpMM with block size > 64, double data type, and alignments smaller than 128-byte on NVIDIA Ampere sm80.
- All cuSPARSE APIs are now asynchronous on platforms that support stream ordered memory allocators [https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#stream-ordered-querying-memory-support](https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html#stream-ordered-querying-memory-support).
- Improved NTVX trace with distinction between light calls and kernel routines

### Resolved Issues

- cusparseCnnz_compress produced wrong results when the number of rows are greater than 128 * resident CTAs.
- cusparseSnnz produced wrong results for some particular sparsity pattern.

### Deprecated Features

- cusparseXcscrsrm2_zeroPivot, cusparseXcscrsrm2_solve, cusparseXcscrsrm2_analysis, and cusparseScsrsm2_bufferSizeExt have been deprecated in favor of cusparseSpSM Generic APIs

#### 2.5.5. cuSPARSE: Release 11.3

### New Features

- Added new routine cusparseSpSV for sparse triangular solver with better performance. The new Generic API supports:
  - CSR storage format
  - Non-transpose, transpose, and transpose-conjugate operations
  - Upper, lower fill mode
  - Unit, non-unit diagonal type
  - 32-bit and 64-bit indices
  - Uniform data type computation

### Deprecated Features

- cusparseScsrsrv2_analysis, cusparseScsrsrv2_solve, cusparseXcscrsrv2_zeroPivot, and cusparseScsrsrv2_bufferSize have been deprecated in favor of cusparseSpSV.

#### 2.5.6. cuSPARSE: Release 11.2 Update 2

### Resolved Issues

- cusparseDestroy(NULL) no longer crashes on Windows.

### Known Issues
CUDA Libraries

- cusparseDestroySpVec, cusparseDestroyDnVec, cusparseDestroySpMat, cusparseDestroyDnMat, cusparseDestroy with NULL argument could cause segmentation fault on Windows.

2.5.7. cuSPARSE: Release 11.2 Update 1

- New Features
  - New Tensor Core-accelerated Block Sparse Matrix - Matrix Multiplication (cusparseSpMM) and introduction of the Blocked-Ellpack storage format.
  - New algorithms for CSR/COO Sparse Matrix - Vector Multiplication (cusparseSpMV) with better performance.
  - Extended functionalities for cusparseSpMV:
    - Support for the CSC format.
    - Support for regular/complex bfloat16 data types for both uniform and mixed-precision computation.
    - Support for mixed regular-complex data type computation.
    - Support for deterministic and non-deterministic computation.
  - New routine for Sampled Dense Matrix - Dense Matrix Multiplication (cusparseSDDMM) which deprecated cusparseConstrainedGeMM and provides better performance.
  - Better accuracy of cusparseAxpby, cusparseRot, cusparseSpVV for bfloat16 and half regular/complex data types.
  - All routines support NVTX annotation for enhancing the profiler time line on complex applications.

- Resolved Issues
  - cusparseAxpby, cusparseGather, cusparseScatter, cusparseRot, cusparseSpVV, cusparseSpMV now support zero-size matrices.
  - cusparseCsr2cscEx2 now correctly handles empty matrices (nnz = 0).
  - cusparseXcsr2csr_compress now uses 2-norm for the comparison of complex values instead of only the real part.

- Known Issues
  - cusparseDestroySpVec, cusparseDestroyDnVec, cusparseDestroySpMat, cusparseDestroyDnMat, cusparseDestroy with NULL argument could cause segmentation fault on Windows.

- Deprecated Features
  - cusparseConstrainedGeMM has been deprecated in favor of cusparseSDDMM.
cuSparseCsrmvEx has been deprecated in favor of cuSparseSpMV.

COO Array of Structure (CooAoS) format has been deprecated including 
cuSparseCreateCooAoS, cuSparseCooAoSGet, and its support for cuSparseSpMV.

2.5.8. cuSPARSE: Release 11.2

- Known Issues
  - cuSparseXdense2csr provides incorrect results for some matrix sizes.

2.5.9. cuSPARSE: Release 11.1 Update 1

- New Features
  - cuSparseSparseToDense
    - CSR, CSC, or COO conversion to dense representation
    - Support row-major and column-major layouts
    - Support all data types
    - Support 32-bit and 64-bit indices
    - Provide performance 3x higher than cuSparseXcsc2dense, cuSparseXcsr2dense
  - cuSparseDenseToSparse
    - Dense representation to CSR, CSC, or COO
    - Support row-major and column-major layouts
    - Support all data types
    - Support 32-bit and 64-bit indices
    - Provide performance 3x higher than cuSparseXcsc2dense, cuSparseXcsr2dense
  - Known Issues
    - cuSparseXdense2csr provides incorrect results for some matrix sizes.
  - Deprecated Features
    - Legacy conversion routines: cuSparseXcsc2dense, cuSparseXcsr2dense, 
cuSparseXdense2csc, cuSparseXdense2csr

2.5.10. cuSPARSE: Release 11.0

- New Features
- Added new Generic APIs for Axpby (cusparseAxpby), Scatter (cusparseScatter), Gather (cusparseGather), Givens rotation (cusparseRot). __nv_bfloat16/ __nv_bfloat162 data types and 64-bit indices are also supported.

- This release adds the following features for cusparseSpMM:
  - Support for row-major layout for cusparseSpMM for both CSR and COO format
  - Support for 64-bit indices
  - Support for __nv_bfloat16 and __nv_bfloat162 data types
  - Support for the following strided batch mode:
    - Ci=A⋅Bi
    - Ci=Ai⋅B
    - Ci=Ai⋅Bi

2.5.11. cuSPARSE: Release 11.0 RC

- **New Features**
  - Added new Generic APIs for Axpby (cusparseAxpby), Scatter (cusparseScatter), Gather (cusparseGather), Givens rotation (cusparseRot). __nv_bfloat16/ __nv_bfloat162 data types and 64-bit indices are also supported.
  - This release adds the following features for cusparseSpMM:
    - Support for row-major layout for cusparseSpMM for both CSR and COO format
    - Support for 64-bit indices
    - Support for __nv_bfloat16 and __nv_bfloat162 data types
    - Support for the following strided batch mode:
      - Ci=A⋅Bi
      - Ci=Ai⋅B
      - Ci=Ai⋅Bi
  - SpVV: added support for __nv_bfloat16.

- **Deprecated Features**
  The following functions have been removed:
  - cusparse<t>gemmi()
CUDA Libraries

NVIDIA CUDA Toolkit 11.5.1

2.6. Math Library

2.6.1. CUDA Math: Release 11.5

Deprecations

The following undocumented CUDA Math APIs are deprecated and will be removed in a future release. Please consider switching to similar intrinsic APIs documented here: https://docs.nvidia.com/cuda/cuda-math-api/index.html

__device__ int mulhi(const int a, const int b)
__device__ unsigned int mulhi(const unsigned int a, const unsigned int b)

__device__ unsigned int mulhi(const int a, const unsigned int b)

__device__ unsigned int mulhi(const unsigned int a, const int b)

__device__ long long int mul64hi(const long long int a, const long long int b)

__device__ unsigned long long int mul64hi(const unsigned long long int a, const unsigned long long int b)

__device__ int float_as_int(const float a)

__device__ float int_as_float(const int a)

__device__ unsigned int float_as_uint(const float a)

__device__ float uint_as_float(const unsigned int a)

__device__ float saturate(const float a)

__device__ int mul24(const int a, const int b)

__device__ unsigned int umul24(const unsigned int a, const unsigned int b)

__device__ int float2int(const float a, const enum cudaRoundMode mode = cudaRoundZero)

__device__ unsigned int float2uint(const float a, const enum cudaRoundMode mode = cudaRoundZero)

__device__ float int2float(const int a, const enum cudaRoundMode mode = cudaRoundNearest)

__device__ float uint2float(const unsigned int a, const enum cudaRoundMode mode = cudaRoundNearest)

2.6.2. CUDA Math: Release 11.4

Beginning in 2022, the NVIDIA Math Libraries official hardware support will follow an N-2 policy, where N is an x100 series GPU.

2.6.3. CUDA Math: Release 11.3

- Resolved Issues
Previous releases of CUDA were potentially delivering incorrect results in some Linux distributions for the following host Math APIs: sinpi, cospi, sincospi, sinpif, cospif, sincospif. If passed huge inputs like 7.3748776e+15 or 8258177.5 the results were not equal to 0 or 1. These have been corrected with this release.

2.6.4. CUDA Math: Release 11.1

- **New Features**
  - Added host support for half and nv_bfloat16 converts to/from integer types.
  - Added __hcmadd() device only API for fast half2 and nv_bfloat16 based complex multiply-accumulate.

2.6.5. CUDA Math: Release 11.0 Update 1

- **Resolved Issues**
  - nv_bfloat16 comparison functions could trigger a fault with misaligned addresses.
  - Performance improvements in half and nv_bfloat16 basic arithmetic implementations.

2.6.6. CUDA Math: Release 11.0 RC

- **New Features**
  - Add arithmetic support for __nv_bfloat16 floating-point data type with 8 bits of exponent, 7 explicit bits of mantissa.
  - Performance and accuracy improvements in single precision math functions: fmodf, expf, exp10f, sinhf, and coshf.

- **Resolved Issues**
  - Corrected documented maximum ulp error thresholds in erfcinvf and powf.
  - Improved cuda_fp16.h interoperability with Visual Studio C++ compiler.
  - Updated libdevice user guide and CUDA math API definitions for j1, jlf, fmod, fmodf, ilogb, and ilogbf math functions.

2.7. NVIDIA Performance Primitives (NPP)

2.7.1. NPP: Release 11.5

- **New Features**
New APIs added to compute Signed Anti-aliased Distance Transform using PBA, the anti-aliased Euclidean distance between pixel sites in images. This will improve the accuracy of distance transform.

- `nppiSignedDistanceTransformAbsPBA_xxxxx_C1R_Ctx()` - Input and output combination supports (xxxxxx) - 32f, 32f64f, 64f

New API for Absolute Manhattan distance transform; another method to improve the accuracy of distance transform using Manhattan distance transform between pixels.

- `nppiDistanceTransformAbsPBA_xxxxx_C1R_Ctx()` - Input and output combination supports (xxxxxx) - 8u16u, 8s16u, 16u16u, 16s16u, 8u32f, 8s32f, 16u32f, 16s32f, 8u64f, 8s64f, 16u64f, 16s64f, 32f64f, 64f

Resolved Issues

- Fixed an issue in `FilterMedian()` API with add interpolation when mask even size.
- Improved Contour function performance by parallelizing more of it and also improving quality.
- Resolved an issue with Alpha composition used to accumulate output buffers multiple times.
- Resolved an issue with `nppiLabelMarkersUF_8u32u_C1R` column processing incorrect results.

2.7.2. NPP: Release 11.4

New Features

- New API `FindContours`. `FindContours` can be explained simply as a curve joining all the continuous points (along the boundary), having the same color or intensity. The contours are a useful tool for shape analysis and object detection and recognition.

2.7.3. NPP: Release 11.3

New Features

- Added `nppiDistanceTransformPBA` functions.

2.7.4. NPP: Release 11.2 Update 2

New Features

- Added `nppiDistanceTransformPBA` functions.

2.7.5. NPP: Release 11.2 Update 1

New Features

New APIs added to compute Distance Transform using Parallel Banding Algorithm (PBA):
CUDA Libraries

- `nppiDistanceTransformPBA_xxxxx_C1R_Ctx()` — where `xxxxx` specifies the input and output combination: 8u16u, 8s16u, 16u16u, 16s16u, 8u32f, 8s32f, 16u32f, 16s32f
- `nppiSignedDistanceTransformPBA_32f_C1R_Ctx()`

Resolved Issues
- Fixed the issue in which Label Markers adds zero pixel as object region.

2.7.6. NPP: Release 11.0

- New Features
  - Batched Image Label Markers Compression that removes sparseness between marker label IDs output from LabelMarkers call.
  - Image Flood Fill functionality fills a connected region of an image with a specified new value.
  - Stability and performance fixes to Image Label Markers and Image Label Markers Compression.

2.7.7. NPP: Release 11.0 RC

- New Features
  - Batched Image Label Markers Compression that removes sparseness between marker label IDs output from LabelMarkers call.
  - Image Flood Fill functionality fills a connected region of an image with a specified new value.
  - Added batching support for `nppiLabelMarkersUF` functions.
  - Added the `nppiCompressMarkerLabelsUF_32u_C1IR` function.
  - Added `nppiSegmentWatershed` functions.
  - Added sample apps on GitHub demonstrating the use of NPP application managed stream contexts along with watershed segmentation and batched and compressed UF image label markers functions.
  - Added support for non-blocking streams.

- Resolved Issues
  - Stability and performance fixes to Image Label Markers and Image Label Markers Compression.
  - Improved quality of `nppiLabelMarkersUF` functions.
  - `nppiCompressMarkerLabelsUF_32u_C1IR` can now handle a huge number of labels generated by the `nppiLabelMarkersUF` function.

- Known Issues
The `nppiCopy` API is limited by CUDA thread for large image size. Maximum image limits is a minimum of 16 * 65,535 = 1,048,560 horizontal pixels of any data type and number of channels and 8 * 65,535 = 524,280 vertical pixels for a maximum total of 549,739,036,800 pixels.

2.8. **nvJPEG Library**

2.8.1. **nvJPEG: Release 11.5 Update 1**

- **Resolved Issues**
  - Fixed the issue in which `nvcuvid()` released uncompressed frames causing a memory leak.

2.8.2. **nvJPEG: Release 11.4**

- **Resolved Issues**
  - Additional subsampling added to solve the NVJPEG_CSS_2x4.

2.8.3. **nvJPEG: Release 11.2 Update 1**

- **New Features**
  - nvJPEG decoder added new APIs to support region of interest (ROI) based decoding for batched hardware decoder:
    - `nvjpegDecodeBatchedEx()`
    - `nvjpegDecodeBatchedSupportedEx()`

2.8.4. **nvJPEG: Release 11.1 Update 1**

- **New Features**
  - Added error handling capabilities for nonstandard JPEG images.

2.8.5. **nvJPEG: Release 11.0 Update 1**

- **Known Issues**
  - `NVJPEG_BACKEND_GPU_HYBRID` has an issue when handling bit-streams which have corruption in the scan.

2.8.6. **nvJPEG: Release 11.0**

- **New Features**
CUDA Libraries

- nvJPEG allows the user to allocate separate memory pools for each chroma subsampling format. This helps avoid memory re-allocation overhead. This can be controlled by passing the newly added flag `NVJPEG_FLAGS_ENABLE_MEMORY_POOLS` to the `nvjpegCreateEx` API.
- nvJPEG encoder now allow compressed bitstream on the GPU Memory.

2.8.7. nvJPEG: Release 11.0 RC

- **New Features**
  - nvJPEG allows the user to allocate separate memory pools for each chroma subsampling format. This helps avoid memory re-allocation overhead. This can be controlled by passing the newly added flag `NVJPEG_FLAGS_ENABLE_MEMORY_POOLS` to the `nvjpegCreateEx` API.
  - nvJPEG encoder now allow compressed bitstream on the GPU Memory.
  - Hardware accelerated decode is now supported on NVIDIA A100.
  - The nvJPEG decode API (nvjpegDecodeJpeg()) now has the flexibility to select the backend when creating `nvjpegJpegDecoder_t` object. The user has the option to call this API instead of making three separate calls to `nvjpegDecodeJpegHost()`, `nvjpegDecodeJpegTransferToDevice()`, and `nvjpegDecodeJpegDevice()`.

- **Known Issues**
  - `NVJPEG_BACKEND_GPU_HYBRID` has an issue when handling bit-streams which have corruption in the scan.

- **Deprecated Features**
  The following multiphase APIs have been removed:
  - nvjpegStatus_t NVJPEGAPI nvjpegDecodePhaseOne
  - nvjpegStatus_t NVJPEGAPI nvjpegDecodePhaseTwo
  - nvjpegStatus_t NVJPEGAPI nvjpegDecodePhaseThree
  - nvjpegStatus_t NVJPEGAPI nvjpegDecodeBatchedPhaseOne
  - nvjpegStatus_t NVJPEGAPI nvjpegDecodeBatchedPhaseTwo
Notice

This document is provided for information purposes only and shall not be regarded as a warranty of a certain functionality, condition, or quality of a product. NVIDIA Corporation ("NVIDIA") makes no representations or warranties, expressed or implied, as to the accuracy or completeness of the information contained in this document and assumes no responsibility for any errors contained herein. NVIDIA shall have no liability for the consequences or use of such information or for any infringement of patents or other rights of third parties that may result from its use. This document is not a commitment to develop, release, or deliver any Material (defined below), code, or functionality.

NVIDIA reserves the right to make corrections, modifications, enhancements, improvements, and any other changes to this document, at any time without notice.

Customer should obtain the latest relevant information before placing orders and should verify that such information is current and complete.

NVIDIA products are sold subject to the NVIDIA standard terms and conditions of sale supplied at the time of order acknowledgement, unless otherwise agreed in an individual sales agreement signed by authorized representatives of NVIDIA and customer ("Terms of Sale"). NVIDIA hereby expressly objects to applying any customer general terms and conditions with regards to the purchase of the NVIDIA product referenced in this document. No contractual obligations are formed either directly or indirectly by this document.

OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

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

© 2007-2021 NVIDIA Corporation & affiliates. All rights reserved.