NVIDIA Datacenter Drivers

User Guide

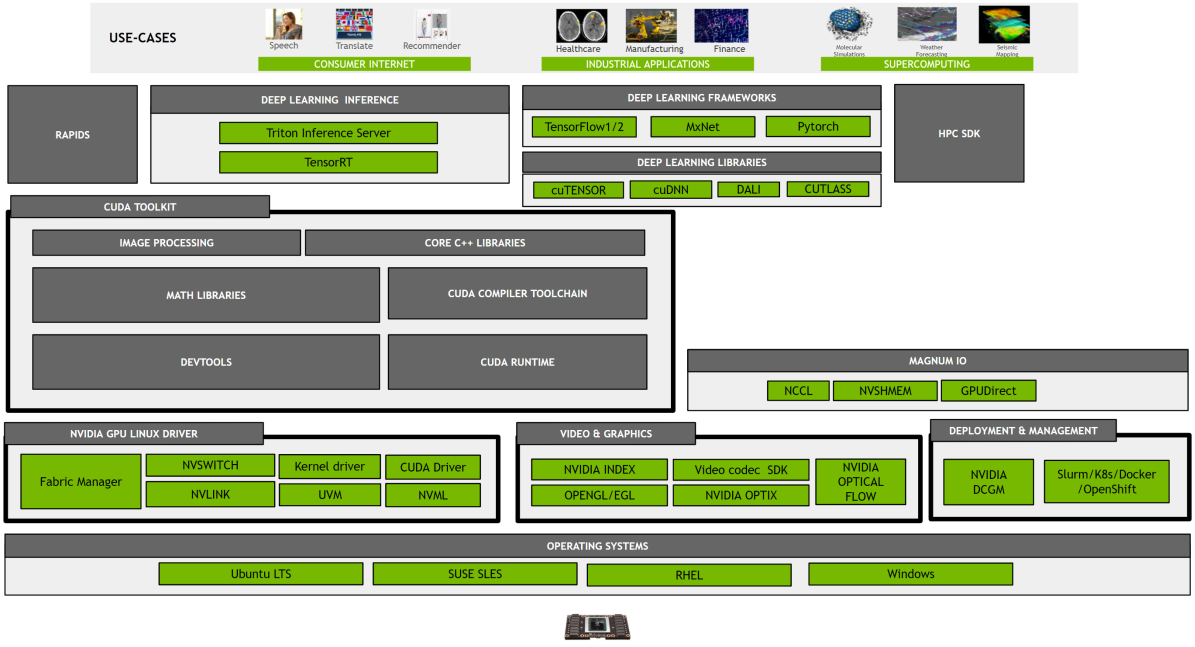
RN-08625 _v01 | February 2024

Table of Contents

Chapter 1. Introduction

The NVIDIA compute software “stack” consists of various software products in the system software or infrastructure that are required to bootstrap a system with NVIDIA GPUs and be able to run accelerated AI or HPC workloads. A software architecture diagram of CUDA and associated components is shown below for reference:

Figure 1. Overview of CUDA Toolkit and Associated Products



While NVIDIA provides a very rich software platform including SDKs, frameworks and applications, the focus of this document is on drivers, CUDA Toolkit and the Deep Learning libraries.

Chapter 2. Driver Lifecycle

2.1. Driver Branches

Starting in 2019, NVIDIA has introduced a new enterprise software lifecycle for datacenter GPU drivers.

New Feature Branch

Major feature release, indicated by a new branch X number. This is targeted towards early adopters who want to evaluate new features (e.g. new CUDA APIs). Note that these drivers may also be shipped along with CUDA Toolkit installer packages in some cases.

Release cadence: New driver branch is released approx. every quarter.

Production Branch

Branch that is qualified for use in production for enterprise/datacenter GPUs. Bug fixes and security updates are provided for up to 1 year.

Release cadence: Two driver branches are released per year (approx. every six months)

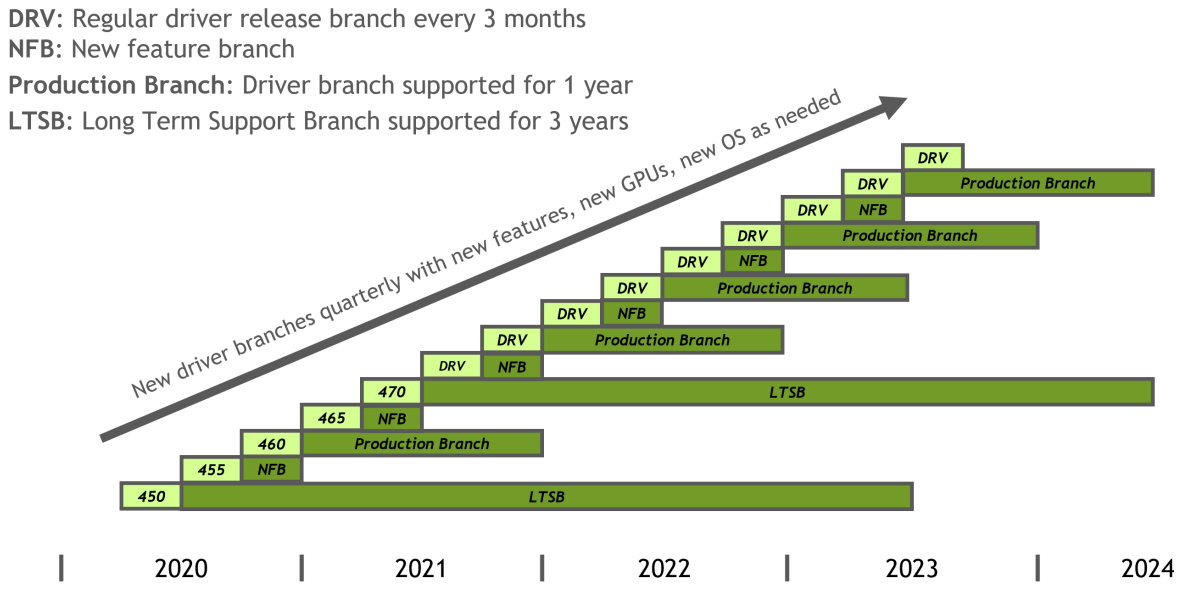
Note that during the lifetime of a production branch, quarterly bug fixes and security updates are released.

Long Term Support Branch

A production branch that will be supported and maintained for a much longer time than a normal production branch is supported. Every LTSB is a production branch, but not every production branch is an LTSB.

Customers who are looking for a longer cycle of support from their deployed branch will gain that support through LTSB releases. LTSB releases will receive bug updates and critical security updates, on a reasonable effort basis, through minor releases during the 3 years that they are supported.

Figure 2. Taxonomy of NVIDIA Driver Branches. (For illustration purposes only. See note below)



2.2. Comparison of Driver Branches

The table below summarizes the differences between the various driver branches.

Table 1. NVIDIA Driver Branches

	New Feature Branch (NFB)	Production Branch (PB)	Long Term Support Branch (LTSB)
Target Customers	Early adopters who want to evaluate new features	Use in production for enterprise/ datacenter GPUs	Use in production for enterprise/ datacenter GPUs and for customers looking for a longer cycle of support.
Major Release Cadence	At least once every 3 months	Twice a year. See also note below	At least once per hardware architecture. See also note below
Length of support	N/A	1 year	3 years
Minor release (bug updates and critical security updates)	N/A	Yes. Quarterly bug and security releases for 1 year.	Yes. Quarterly bug and security releases for 1 year.

Note:

General guidance only. The actual security update and release cadence can change at NVIDIA's discretion.

Chapter 3. Supported Drivers and CUDA Toolkit Versions

NVIDIA releases CUDA Toolkit and GPU drivers at different cadences. The NVIDIA datacenter GPU driver software lifecycle and terminology are available in the lifecycle section of this documentation.

The release information can be scraped by automation tools (for example `jq`) by parsing the release information: `releases.json`.

The table below lists the current support matrix for CUDA Toolkit and NVIDIA datacenter drivers.

Table 2. CUDA and Drivers

	R470	R535	R550
Branch Designation	Long Term Support Branch	Long Term Support Branch	Production Branch
End of Life	July 2024	June 2026	February 2025
Maximum CUDA Version Supported	CUDA 11.0+ This driver branch supports CUDA 11.x (through CUDA minor version compatibility).	CUDA 12.0+ This driver branch supports CUDA 12.x (through CUDA minor version compatibility).	CUDA 12.0+ This driver branch supports CUDA 12.x (through CUDA minor version compatibility).

Note: All other previous driver branches not listed in the table above (such as R525, R515, R510, R495, R465, R460, R455, R450, R440, R418, R410) are end of life.

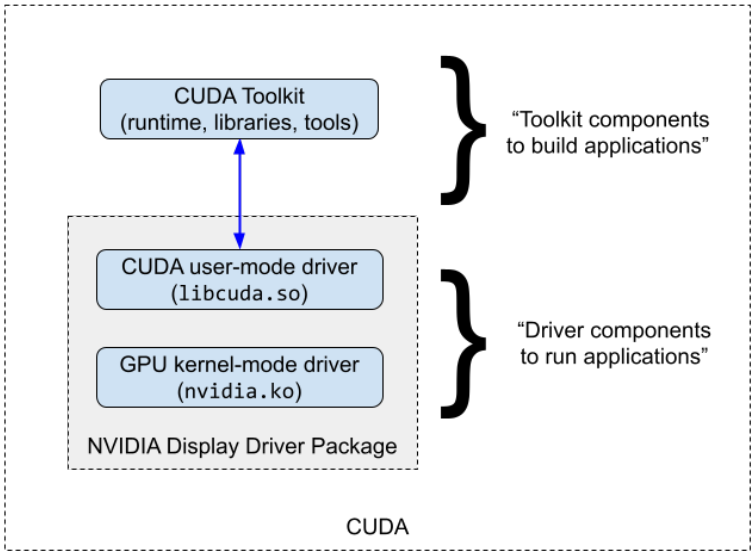
Chapter 4. Software Deployment Workflow

The CUDA software environment consists of three parts:

- CUDA Toolkit (libraries, runtime and tools) - User-mode SDK used to build CUDA applications
- CUDA driver - User-mode driver component used to run CUDA applications (e.g. libcuda.so on Linux systems)
- NVIDIA GPU device driver - Kernel-mode driver component for NVIDIA GPUs

On Linux systems, the CUDA driver and kernel mode components are delivered together in the NVIDIA display driver package. This is shown in the figure below.

Figure 3. CUDA



The CUDA Toolkit is generally optional when GPU nodes are only used to run applications (as opposed to develop applications) as the CUDA application typically packages (by statically or dynamically linking against) the CUDA runtime and libraries needed.

Typical Workflow

A typical suggested workflow for bootstrapping a GPU node in a cluster:

1. Install the NVIDIA drivers (do not install CUDA Toolkit as this brings in additional dependencies that may not be necessary or desired)
2. Install the CUDA Toolkit using meta-packages. This provides additional control over what is installed on the system.
3. Install other components such as cuDNN or TensorRT as desired depending on the application requirements and dependencies.

4.1. Datacenter Driver Installation

Note: The full content of this section is available at: https://docs.nvidia.com/datacenter/tesla/tesla-installation-notes/index.html.

NVIDIA drivers are available in three formats for use with Linux distributions:

- Runfile installers
- Package managers
- Containerized drivers

NVIDIA provides Linux distribution specific packages for drivers that can be used by customers to deploy drivers into a production environment. The links above provide detailed information and steps on how to install driver packages for supported Linux distributions, but a summary is provided below.

4.1.1. Installation Using Package Managers

Using package managers is the recommended method of installing drivers as this provides additional control over choice of driver branches, precompiled kernel modules, driver upgrades and additional dependencies such as Fabric Manager/NSCQ for NVSwitch systems.

On Ubuntu LTS

```
$ sudo apt-get -y install cuda-drivers-<branch-number>
```

Where the branch-number = the specific datacenter branch of interest (e.g. 450, 460)

On RHEL 8

```
$ sudo dnf module install nvidia-driver:<stream>/<profile>
```

For example, `nvidia-driver:latest-dkms/fm` will install the latest drivers and also install the Fabric Manager dependencies to bootstrap an NVSwitch system such as HGX A100.

For more information on the supported streams/profiles, refer to this section in the documentation.

4.2. CUDA Toolkit Installation

The CUDA Toolkit packages are modular and offer the user control over what components of the CUDA Toolkit are installed on the system. CUDA supports a number of meta-packages that are available here.

Since the `cuda` or `cuda-<release>` packages also install the drivers, these packages may not be appropriate for datacenter deployments.

Instead, other packages such as `cuda-toolkit-<release>` should be used as this package has no dependency on the driver. The following example only installs the CUDA Toolkit 11.4 packages and does not install the driver.

```
$ sudo apt-get -y install cuda-toolkit-11-4
```

Table 3. Supported CUDA Meta Packages

Meta-Package	Purpose
`cuda`	Installs all CUDA Toolkit and Driver packages. Handles upgrading to the next version of the `cuda` package when it's released.
`cuda-11-4`	Installs all CUDA Toolkit and Driver packages. Remains at version 11.4 until an additional version of CUDA is installed.
`cuda-toolkit-11-4`	Installs all CUDA Toolkit packages required to develop CUDA applications. Does not include the driver.
`cuda-tools-11-4`	Installs all CUDA command line and visual tools.
`cuda-runtime-11-4`	Installs all CUDA Toolkit packages required to run

Meta-Package	Purpose
	CUDA applications, as well as the Driver packages.
`cuda-compiler-11-4`	Installs all CUDA compiler packages.
`cuda-libraries-11-4`	Installs all runtime CUDA Library packages.
`cuda-libraries-dev-11-4`	Installs all development CUDA Library packages.
`cuda-drivers`	Installs all Driver packages. Handles upgrading to the next version of the Driver packages when they're released.

4.3. cuDNN Installation

NVIDIA cuDNN can also be installed from the CUDA network repository using Linux package managers by using the `libcudnn` and `libcudnn-dev` packages. Some examples on supported Linux distributions are shown below:

Ubuntu LTS

```
$ CUDNN_VERSION=8.1.1.33 \
   && sudo apt-get -y install \
   libcudnn8=$CUDNN_VERSION-1+cuda11.2 libcudnn8-dev=$CUDNN_VERSION-1+cuda11.2
```

Chapter 5. Software Support Matrix

5.1. CUDA Toolkit, Driver and Architecture Matrix

The CUDA driver provides an API that is backwards compatible. Thus, new NVIDIA drivers will always work with (applications compiled with) an older CUDA toolkit. This behavior of CUDA is documented here. Each CUDA Toolkit however, requires a minimum version of the NVIDIA driver. Corollarily, when using tools such as `nvidia-smi`, the NVIDIA driver reports a maximum version of CUDA supported and thus is able to run applications built with CUDA Toolkits up to that version.

CUDA Toolkit and drivers may also deprecate and drop support for GPU architectures over the product life cycle of the CUDA Toolkit. See the `-arch` and `-gencode` options in the CUDA compiler (`nvcc`) toolchain documentation.

Table 4. CUDA and Architecture Matrix

Architecture	CUDA Capabilities	First CUDA Toolkit Support	Last CUDA Toolkit Support	Last Driver Support
Fermi	2.0	CUDA 3.0	CUDA 8.0	R390
Kepler	3.0 3.2	CUDA 6.0	CUDA 10.2	R470
Kepler	3.5 3.7	CUDA 6.0	CUDA 11.x	R470
Maxwell	5.0 5.2 5.3	CUDA 6.5	Ongoing	Ongoing
Pascal	6.0 6.1	CUDA 8.0	Ongoing	Ongoing
Volta	7.0	CUDA 9.0	Ongoing	Ongoing

Architecture	CUDA Capabilities	First CUDA Toolkit Support	Last CUDA Toolkit Support	Last Driver Support
Turing	7.5	CUDA 10.0	Ongoing	Ongoing
Ampere	8.0 8.6	CUDA 11.0	Ongoing	Ongoing
Ada	8.9	CUDA 11.8	Ongoing	Ongoing
Hopper	9.0	CUDA 11.8 CUDA 12.0	Ongoing	Ongoing

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.

NVIDIA products are not designed, authorized, or warranted to be suitable for use in medical, military, aircraft, space, or life support equipment, nor in applications where failure or malfunction of the NVIDIA product can reasonably be expected to result in personal injury, death, or property or environmental damage. NVIDIA accepts no liability for inclusion and/or use of NVIDIA products in such equipment or applications and therefore such inclusion and/or use is at customer’s own risk.

NVIDIA makes no representation or warranty that products based on this document will be suitable for any specified use. Testing of all parameters of each product is not necessarily performed by NVIDIA. It is customer’s sole responsibility to evaluate and determine the applicability of any information contained in this document, ensure the product is suitable and fit for the application planned by customer, and perform the necessary testing for the application in order to avoid a default of the application or the product. Weaknesses in customer’s product designs may affect the quality and reliability of the NVIDIA product and may result in additional or different conditions and/or requirements beyond those contained in this document. NVIDIA accepts no liability related to any default, damage, costs, or problem which may be based on or attributable to: (i) the use of the NVIDIA product in any manner that is contrary to this document or (ii) customer product designs.

No license, either expressed or implied, is granted under any NVIDIA patent right, copyright, or other NVIDIA intellectual property right under this document. Information published by NVIDIA regarding third-party products or services does not constitute a license from NVIDIA to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property rights of the third party, or a license from NVIDIA under the patents or other intellectual property rights of NVIDIA.

Reproduction of information in this document is permissible only if approved in advance by NVIDIA in writing, reproduced without alteration and in full compliance with all applicable export laws and regulations, and accompanied by all associated conditions, limitations, and notices.

THIS DOCUMENT AND 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. TO THE EXTENT NOT PROHIBITED BY LAW, IN NO EVENT WILL NVIDIA BE LIABLE FOR ANY DAMAGES, INCLUDING WITHOUT LIMITATION ANY DIRECT, INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES, HOWEVER CAUSED AND REGARDLESS OF THE THEORY OF LIABILITY, ARISING OUT OF ANY USE OF THIS DOCUMENT, EVEN IF NVIDIA HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. Notwithstanding any damages that customer might incur for any reason whatsoever, NVIDIA’s aggregate and cumulative liability towards customer for the products described herein shall be limited in accordance with the Terms of Sale for the product.

Trademarks

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

Copyright

NVIDIA Corporation | 2788 San Tomas Expressway, Santa Clara, CA 95051
https://www.nvidia.com