NVIDIA TensorRT

Support Matrix | NVIDIA Docs

TRM-09025-001 _v8.5.1 | November 2022

Table of Contents

Chapter 1. Features for Platforms and Software

This section lists the supported NVIDIA® TensorRT™ features based on which platform and software.

Table 1. List of Supported Features per Platform

	Linux x86-64	Windows x64	Linux ppc64le	Linux AArch64
	8.5.x	8.5.x	8.5.x	8.5.x
Supported NVIDIA CUDA® versions	11.8 11.7 update 1[1] 11.6 update 2[2] 11.5 update 2[3] 11.4 update 4[4] 11.3 update 1[5]	11.8 11.7 update 1[9] 11.6 update 2[10] 11.5 update 2[11] 11.4 update 4[12] 11.3 update 1[13]	11.8	11.8 11.4

[1] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.

[9] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.

[2] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.

[10] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.

[3] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.

[11] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.

[4] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.

[12] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.

[5] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.

[13] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.

	Linux x86-64	Windows x64	Linux ppc64le	Linux AArch64
	8.5.x	8.5.x	8.5.x	8.5.x
	11.2 update 2[6] 11.1 update 1[7] 11.0 update 1[8] 10.2	11.2 update 2[14] 11.1 update 1[15] 11.0 update 1[16] 10.2		
Supported cuBLAS versions	11.11.3.6 11.10.3.66 11.9.2.110 11.7.4.6 11.6.5.2 11.5.1.109 11.4.1.1043 11.3.0.106 11.2.0.252 10.2.3.254	11.11.3.6 11.10.3.66 11.9.2.110 11.7.4.6 11.6.5.2 11.5.1.109 11.4.1.1043 11.3.0.106 11.2.0.252 10.2.3.254	11.11.3.6	11.11.3.6 11.6.5.2
Supported cuDNN versions	cuDNN 8.6.0	cuDNN 8.6.0	cuDNN 8.6.0	cuDNN 8.6.0
TensorRT Python API	Yes	Yes	Yes	Yes
NvUffParser	Yes	Yes	Yes	Yes
NvOnnxParser	Yes	Yes	Yes	Yes
Loops	Yes	Yes	Yes	Yes

Note:

- Serialized engines are not portable across platforms or TensorRT versions.

[6] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.
[14] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.
[7] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.
[15] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.
[8] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.
[16] These CUDA versions are supported using a single build, built with CUDA toolkit 11.8. It is compatible with all CUDA 11.x versions and only requires driver 450.x.

- Refer to the minimum compatible driver versions in the *NVIDIA CUDA Release Notes* for specific NVIDIA Driver versions.

Chapter 2. Hardware and Precision

The following table lists NVIDIA hardware and which precision modes that each hardware supports. TensorRT supports all NVIDIA hardware with capability SM 5.0 or higher. It also lists the availability of DLA on this hardware. Refer to the following tables for the specifics.

Note: Support for CUDA compute capability version 3.0 has been removed. Support for CUDA compute capability versions below 5.0 may be removed in a future release and is now deprecated.

Table 2. Supported Hardware

CUDA Compute Capability	Example Device	TF32	FP32	FP16	INT8	FP16 Tensor Cores	INT8 Tensor Cores	DLA
9.0	NVIDIA H100	Yes	Yes	Yes	Yes	Yes	Yes	No
8.9	NVIDIA RTX 4090	Yes	Yes	Yes	Yes	Yes	Yes	No
8.7	NVIDIA DRIVE AGX Orin™	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8.6	NVIDIA A10	Yes	Yes	Yes	Yes	Yes	Yes	No
8.0	NVIDIA A100/ GA100 GPU	Yes	Yes	Yes	Yes	Yes	Yes	No
7.5	NVIDIA T4	No	Yes	Yes	Yes	Yes	Yes	No

CUDA Compute Capability	Example Device	TF32	FP32	FP16	INT8	FP16 Tensor Cores	INT8 Tensor Cores	DLA
7.2	Jetson AGX Xavier	No	Yes	Yes	Yes	Yes	Yes	Yes
7.0	NVIDIA V100	No	Yes	Yes	Yes	Yes	No	No
6.1	NVIDIA P4	No	Yes	Yes	Yes	No	No	No
6.0	NVIDIA P100	No	Yes	Yes	No	No	No	No
5.2	NVIDIA M4	No	Yes	No	No	No	No	No
5.0	Quadro K2200	No	Yes	No	No	No	No	No

Deprecated Hardware

Table 3. List of Supported Precision Mode per Hardware

CUDA Compute Capability	Example Device	FP32	FP16	INT8	FP16 Tensor Cores	INT8 Tensor Cores	DLA
3.7	NVIDIA K80	Yes	No	No	No	No	No
3.5	NVIDIA K40	Yes	No	No	No	No	No

Removed Hardware

Table 4. List of Supported Precision Mode per Hardware

CUDA Compute Capability	Example Device	FP32	FP16	INT8	FP16 Tensor Cores	INT8 Tensor Cores	DLA
3.0	NVIDIA K10	Yes	No	No	No	No	No

Chapter 3. Compute Capability Per Platform

The section lists the supported compute capability based on platform.

Table 5. Compute Capability per Platform

Platform	Compute capability
Linux x86-64	3.5, 3.7, 5.0, 5.2, 6.0, 6.1, 7.0, 7.5, 8.0[17], 8.6[18], 8.9[19], 9.0[20]
Windows 10 x64	3.5, 3.7, 5.0, 5.2, 6.0, 6.1, 7.0, 7.5, 8.0[21], 8.6[22], 8.9[23], 9.0[24]
CentOS 8.5 ppc64le	7.0, 7.5, 8.0, 8.6, 9.0
Ubuntu 20.04 SBSA	7.0, 7.5, 8.0, 8.6, 9.0
NVIDIA JetPack AArch64	7.2, 8.7

[17] Requires CUDA Toolkit 11.0 or newer and a TensorRT CUDA 11.x build.
[18] Requires CUDA Toolkit 11.1 or newer and a TensorRT CUDA 11.x build.
[19] Requires CUDA Toolkit 11.8 or newer and a TensorRT CUDA 11.x build.
[20] Requires CUDA Toolkit 11.8 or newer and a TensorRT CUDA 11.x build.
[21] Requires CUDA Toolkit 11.0 or newer and a TensorRT CUDA 11.x build.
[22] Requires CUDA Toolkit 11.1 or newer and a TensorRT CUDA 11.x build.
[23] Requires CUDA Toolkit 11.8 or newer and a TensorRT CUDA 11.x build.
[24] Requires CUDA Toolkit 11.8 or newer and a TensorRT CUDA 11.x build.

Chapter 4. Software Versions Per Platform

The section lists the supported software versions based on platform.

Table 6. List of Supported Platforms per Software Version

Platform	Compiler version	Python version
Ubuntu 18.04 x86-64	gcc 8.3.1	3.6
Ubuntu 20.04 x86-64	gcc 8.3.1	3.8
Ubuntu 22.04 x86-64	gcc 8.3.1	3.10
CentOS 7.9 x86-64	gcc 8.3.1	3.6
CentOS 8.5 x86-64	gcc 8.3.1	3.8
SLES 15 x86-64	gcc 8.3.1	Not Applicable
Windows 10 x64	MSVC 2017u8	Not Applicable
CentOS 8.5 ppc64le	Clang 14.0.6	3.8
Ubuntu 20.04 SBSA	gcc 8.4.0	3.8
NVIDIA JetPack AArch64	gcc 9.4.0	3.8

Note: Python versions supported when using Debian or RPM packages. When using Python wheel files, versions 3.6, 3.7, 3.8, 3.9, and 3.10 are supported.

Chapter 5. ONNX Operator Support

The ONNX operator support list for TensorRT can be found here.

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.

Arm

Arm, AMBA and Arm Powered are registered trademarks of Arm Limited. Cortex, MPCore and Mali are trademarks of Arm Limited. "Arm" is used to represent Arm Holdings plc; its operating company Arm Limited; and the regional subsidiaries Arm Inc.; Arm KK; Arm Korea Limited.; Arm Taiwan Limited; Arm France SAS; Arm Consulting (Shanghai) Co. Ltd.; Arm Germany GmbH; Arm Embedded Technologies Pvt. Ltd.; Arm Norway, AS and Arm Sweden AB.

HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

Blackberry/QNX

Trademarks, including but not limited to BLACKBERRY, EMBLEM Design, QNX, AVIAGE, MOMENTICS, NEUTRINO and QNX CAR are the trademarks or registered trademarks of BlackBerry Limited, used under license, and the exclusive rights to such trademarks are expressly reserved.

Google

Android, Android TV, Google Play and the Google Play logo are trademarks of Google, Inc.

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

Trademarks

NVIDIA, the NVIDIA logo, and BlueField, CUDA, DALI, DRIVE, Hopper, JetPack, Jetson AGX Xavier, Jetson Nano, Kepler, Maxwell, NGC, Nsight, Orin, Pascal, Quadro, Tegra, TensorRT, Triton, Turing and Volta are trademarks and/or registered trademarks of NVIDIA Corporation in the United 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
http://www.nvidia.com