NVIDIA Trusted Computing Solutions

Release Notes

RN-12620-001_v01 | December 2025

Document History

RN-12620-001_v01

Version	Date	Description of Change
01	December 2025	R590 TRD1 general availability (GA) release

Table of Contents

Overview

This release consists of the NVIDIA® CUDA® Toolkit version 13.1, which is paired with the NVIDIA Data Center GPU Drivers version 590.48.01.

The following features are supported in this software release:

- Single GPU Passthrough (SPT CC)
- Hopper Multiple GPU Passthrough with Protected PCIe (PPCIe)
- Blackwell Multiple GPU Passthrough (MPT CC) **NEW**

For more information about these features, refer to Feature Summary.

For additional information, refer to the following documentation:

- *Secure AI Compatibility Matrix* for more information about supported platforms, drivers, firmware, and Secure AI modes
- *Secure AI Deployment Guide* for details about setting up the confidential computing system
- *Secure AI Operations Guide* for best practices and CUDA developer considerations

Before you deploy workloads, NVIDIA recommends that users use good practices, such as performing regular attestations.

Feature Summary

This section provides information about the CC features in this release. For more details on Secure AI modes, security posture, and capabilities, please refer to the *NVIDIA Secure AI with Blackwell and Hopper GPUs* whitepaper.

Single GPU Passthrough (SPT CC)

NVIDIA® Trusted Computing support for NVIDIA Hopper™ GPUs was first introduced with the Hopper Single GPU Passthrough with a Bounce Buffer (SPT CC) mode. SPT CC mode support for Blackwell™ GPUs was also added in subsequent releases. In this mode, one GPU can be passed through for each Confidential VM (CVM). Multiple CVMs may be hosted on the same node, each with one GPU passed through. A bounce buffer stages encrypted data transfers between the GPU device and the CVM.

Following are the hardware SKUs that support SPT CC mode:

- NVIDIA H100 PCIe
- NVIDIA H800 PCIe
- NVIDIA H100NVL
- NVIDIA H800NVL
- NVIDIA H200NVL
- HGX H100 8-GPU 80GB (Air Cooled)
- HGX H100 4-GPU 80GB HBM3 (Partner Cooled)
- HGX H800 8-GPU 80GB (Air Cooled)
- HGX H800 8-GPU 80GB (Partner Cooled)
- HGX H100 4-GPU 64GB HBM2e (Partner Cooled)
- HGX H100 8-GPU 96GB (Air Cooled)
- HGX H100 4-GPU 94GB HBM2e (Partner Cooled)
- HGX H20A HBM3 96GB 8-GPU (Air Cooled)
- HGX H20 141GB HBM3e 8-GPU (Air Cooled)
- HGX H200 8-GPU 141GB (Air Cooled)
- HGX B200, 8-GPU, SXM6 180GB HBM3e, AC
- RTX PRO 6000 Blackwell Server Edition
- HGX B200-850, 8-GPU, SXM6 180GB HBM3e, AC

Refer to the *Intel TDX - Confidential Computing Deployment Guide and AMD SNP - Confidential Computing Deployment Guide* for more information.

Table 1. Component Versions to Enable SPT CC Mode

Component	Version
VBIOS	H100/H200: Hopper FW 1.9.0 [96.00.DA.00.XX] H20A: 96.00.DA.00.10 HGX Blackwell FW 1.3.0 [97.00.D9.00.xx] RTX 6000 FW 1.3 [98.02.9E.00.01] Refer to the *Secure AI Compatibility Matrix* for more information about supported platforms, drivers, firmware, and Secure AI modes
Host OS Kernel	Intel TDX Kernel 6.14 + (Vendor fork) AMD SEV/SEV-SNP 6.11+
Guest OS	Ubuntu 24.04
`gpu_admin.py`	The main branch is github.com/nvidia/nvtrust. GPU tools - v2025.04.07
Attestation SDK Local GPU Verifier	Version 2.6.0 or later

Hopper Multiple GPU Passthrough (Protected PCIe)

Trusted Computing support in the PPCIe mode is available with Hopper-architecture GPUs and Intel® CPUs with TDX/AMD CPUs with SEV/SEV-SNP technology in an Ubuntu KVM/QEMU environment.

In the PPCIe mode, multiple Hopper-architecture GPUs interconnected by NVSwitch or NVLink can be passed through to one CVM. As in the SPT CC mode, a bounce buffer is used to stage encrypted data transfers between the GPU device and CVM over the PCI Express bus. In this mode, GPU-GPU communications over the NVLink or NVSwitch interconnect are not encrypted.

Following are the hardware SKUs that support PPCIe mode:

- HGX H100 8-GPU 80GB (Air Cooled)
- HGX H800 8-GPU 80GB (Air Cooled)
- HGX H100 8-GPU 96GB (Air Cooled)
- HGX H20A HBM3 96GB 8-GPU (Air Cooled)
- HGX H200 8-GPU 141GB (Air Cooled)

Table 2. Component Versions to Enable PPCIe

Component	Version
HGX Hopper FW bundle	HGX Hopper FW 1.9.0 [96.00.DA.00.XX] Refer to the *Secure AI Compatibility Matrix* for more information about supported platforms, drivers, firmware, and Secure AI modes
Host OS Kernel	Intel TDX Kernel 6.14+ (Vendor fork) AMD SEV/SEV-SNP 6.11+
Guest OS	Ubuntu 24.04
gpu_admin.py	The main branch is github.com/nvidia/nvtrust. GPU tools - v2025.04.07
PPCIE Verifier	Version 1.6.0 or later

Blackwell Multiple GPU Passthrough (MPT CC)

This release introduces support for Blackwell Multiple GPU Passthrough with a Bounce Buffer (MPT CC) mode. In this mode, up to eight GPUs can be passed through for each Confidential VM (CVM). A bounce buffer stages encrypted data transfers between the GPU device and CVM. GPUs that are part of the same CVM can communicate peer-to-peer over encrypted NVLink connections.

Following are the hardware SKUs that support MPT CC mode:

- HGX B200, 8-GPU, SXM6 180GB HBM3e, AC
- HGX B200-850, 8-GPU, SXM6 180GB HBM3e, AC

Table 3. Component Versions to Enable MPT CC Mode for Blackwell GPUs

Component	Version
VBIOS	HGX Blackwell FW 1.3.0 [97.00.D9.00.xx] Refer to the *Secure AI Compatibility Matrix* for more information about supported platforms, drivers, firmware, and Secure AI modes
Host OS Kernel	Intel TDX Kernel 6.14+ (Vendor fork) AMD SEV/SEV-SNP 6.11+
Guest OS	Ubuntu 24.04
`gpu_admin.py`	The main branch is github.com/nvidia/nvtrust. GPU tools - v2025.04.07
Attestation SDK Local GPU Verifier	Version 2.6.3 or later

Known Issues

- The key rotation feature is not supported with PPCIe.
  A sophisticated attacker with physical or logical superuser access to the system can act as a passive adversary to capture the ciphertext and execute an attempt to break it or the key.

  Workaround
  Review the latest research on the effects of extreme AES key usage and cryptographic wear out to determine your requirements for an attacker advantage. To create a new set of encryption keys in PPCIe mode, you must terminate and launch your CVMs again.

- In systems with multiple GPUs interconnected by NVLink Switch interconnects, the driver registry key `NVreg_RegistryDwords="RmNvswitchGpioDetect=2"` must be used if any of the virtual machines has only a single GPU passed through with CC-mode set to OFF.

  Workaround
  None.

- With Blackwell CC Modes, the number of decoder sessions is restricted.
  Because of a software bug, you cannot use more than 190 sessions when a single CUDA context is shared among all decoder sessions or more than 28 sessions when one CUDA context per decoder session is created.

  Workaround
  None. This issue will be fixed in a future release.

- Performance is degraded with NCCL 2.26.2 and Protected PCIe.
  Applications might encounter up to 10% slowdown with NCCL 2.26.2.

  Workaround
  Use NCCL 2.26.3 or a newer version with the fix.

- IV exhaustion crashes the application in PPCIe mode.
  The H100 CC modes use a 96-bit deterministic IV for each virtual copy engine that is used to transfer data between the GPU and CPU. When this IV space is exhausted, transfers fail to complete.

  Workaround
  Rotate the keys often in supported modes. If the keys are not rotated often, restart the CVM.

- The GPU-Ready bit is always set when devtools mode is enabled, which may lead to confusion with the requirement to set the GPU-Ready bit when running MPT with devtools mode enabled.

  Workaround
  When in full CC-on modes, the driver does not accept any workloads until after the Attestation SDK, or the users, manually enable a GPU-Ready bit.

  

  Note: This bit is already enabled in devtools mode.

  Users should use best practices by attesting the GPU before performing any work. The GPUs booted in devtools mode are clearly identified, and attestation fails.

- With HGX Hopper firmware 1.6.0, there is an increased risk of the GPU or NVLink Switch falling off the PCIe bus during DC power cycling. This issue was resolved in the 1.7.0 firmware release.

  Workaround
  Reboot the system to bring the missing devices back on the PCIe bus. To avoid this issue, use HGX Hopper firmware 1.8.0 or newer.

- NVIDIA Performance Primitives (NPP) might not work.
  NPP uses optimized coding to extract the maximum performance from commonly used transforms and calculations as part of the leverage pinned host memory, which is not supported in CC.

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, the NVIDIA logo are trademarks and/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.

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

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