NVIDIA Hopper Confidential Compute

Early Access 2 Release Notes

RN-11468-001 | August 2023

Table of Contents

Overview

The NVIDIA® Confidential Computing (CC) features of the NVIDIA H100 Tensor Core GPUs have been updated in this Early Access 2 (EA2) release.

The EA2 release consists of the CUDA Toolkit version 12.2 Update-2 paired with Tesla Recommended Driver version r535.104.05.

This EA software release features a complete software stack that targets a single NVIDIA H100 GPU in passthrough mode with a session key for encryption and authentication and basic use of the Developer Tools. Code and data will be confidential up to the limits of the NIST SP800-38D AES-GCM standard, after which the VM should be restarted.

NVIDIA recommends that users invoke good practices, such as testing only with data that does not have cryptographic confidentiality requirements (for example, synthetic non-production data, non-confidential test vectors, and so on), while using this release.

New Features

- TDX Support in non-container use cases has been introduced.
  Intel TDX CPUs include the features that are required for Confidential VMs. With the r535.104.05 driver, users of TDX machines will now be able to attach an H100 GPU to their CVM for confidential workloads

- GPU Operator now includes support for Confidential Kata Containers.
  The deployment guide for HCC now includes a tutorial on how to launch a sample Confidential Container via Kata. These instructions are for users who do not have a Kubernetes cluster configured.

  For full GPU Operator instructions for Confidential Containers, please refer to its documentation hub.

Known Issues

- A key rotation feature is missing.
  A sophisticated attacker with physical access or logical superuser access to the system might be able to act as a passive adversary to capture the ciphertext and execute an attempt to break the ciphertext or the key.

  Workaround
  Only data without cryptographic confidentiality requirements should be used with this EA release.

- IV rotation exhausts early.
  The H100 CC modes use a 96-bit deterministic IV for each virtual copy engine used to transfer data between the GPU and CPU. It is composed of a concatenation of 64-bits of `channel_counter` and 32-bits of `message_counter`. As a channel reaches the 32-bit message maximum, the NVIDIA driver will not automatically roll over to a new channel.

  Depending on the workload, this may result in either returned channel error codes or silent encryption failures resulting in plaintext transfers, the latter of which presents a security risk.

  Workaround
  Only data without cryptographic confidentiality requirements should be used with this

EA release. Workloads that rely heavily on many UVM page migrations will be most affected by this issue.

- Certain graphic applications will crash a guest virtual machine (VM).
  When trying to execute graphics interop kernels, the guest VM will crash.

  Workaround
  Graphics interop tests are not supported on Hopper CC. Resetting the GPU with a Physical Function Function Level Reset (PF-FLR) from the host will recover the GPU.

- Nonce incrementing on secure worklaunch does not roll over past 2^32 operations.
  After 2^32 operations, the RM API calls will fail to acquire a lock; the driver will appear to be hung.

  Workaround
  This work launch occurs upon the first transaction of a work-launch (passing of the buffer descriptor). Workloads with fragmented memory buffers will reach this limit much sooner than workloads with contiguous memory. In our worst-case scenario tests, this may happen after 12-hours. Resetting a seemingly frozen/hung VM after this time will reset the counters.

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

VESA DisplayPort
DisplayPort and DisplayPort Compliance Logo, DisplayPort Compliance Logo for Dual-mode Sources, and DisplayPort Compliance Logo for Active Cables are trademarks owned by the Video Electronics Standards Association in the United States and other countries.

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

Arm
Arm, AMBA, and ARM Powered are registered trademarks of Arm Limited. Cortex, MPCore, and Mali are trademarks of Arm Limited. All other brands or product names are the property of their respective holders. "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.

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

Copyright

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