NVIDIA VIDEO CODEC SDK - DECODER

Application Note

vNVDEC_DA-08097-001_v08 | March 2024

Table of Contents

Chapter 1. NVIDIA Hardware Video Decoder 1
1.1. Introduction 1
1.2. NVDEC Capabilities 1
1.3. NVDEC Performance 2
1.4. Programming NVDEC 4
1.5. FFmpeg Support 4

Chapter 1. NVIDIA Hardware Video Decoder

1.1. Introduction

NVIDIA GPUs contain a hardware-based decoder (referred to as NVDEC in this document) which provides fully accelerated hardware-based video decoding for several popular codecs. With complete decoding offloaded to NVDEC, the graphics engine and CPU are free for other operations.

NVDEC supports much faster than real-time decoding which makes it suitable for transcoding scenarios in addition to video playback.

The hardware capabilities available in NVDEC are exposed through APIs referred to as NVDECODE APIs in this document. This document provides information about the capabilities of the NVDEC engine and the features exposed through NVDECODE APIs. The current document highlights *only* the changes in the current video codec SDK package with respect to the previous SDK packages. To know about the features exposed in earlier SDKs please refer to the earlier SDK package(s).

1.2. NVDEC Capabilities

At a high level, Table 1 summarizes the capabilities of the NVDEC engine exposed through NVDECODE APIs.

Table 1. NVDEC Hardware Capabilities

Hardware Features	1st Gen Maxwell GPUs	2nd Gen Maxwell GPUs	Pascal GPUs	Volta GPUs	Turing/ GA100/ Hopper GPUs	GA10x[3] and Ada GPUs
VC1 Simple, Main & Advanced profiles	Y	Y	Y	Y	Y	Y

Hardware Features	1st Gen Maxwell GPUs	2nd Gen Maxwell GPUs	Pascal GPUs	Volta GPUs	Turing/ GA100/ Hopper GPUs	GA10x[3] and Ada GPUs
MPEG4 Simple and Advanced Simple Profiles	Y	Y	Y	Y	Y	Y
MPEG2 Simple & Main profiles	Y	Y	Y	Y	Y	Y
H.264 Baseline, Main, High Profiles	Y	Y	Y	Y	Y	Y
VP8	N	Y	Y[1]	Y	Y	Y
HEVC Main and Main 10 Profile[1]	N	Y[1]	Y	Y	Y	Y
VP9 Profile 0[1]	N	Y[1]	Y	Y	Y	Y
8192x8192 Decoding support (HEVC&VP9 only)	N	N	Y[1]	Y	Y	Y
Multiple NVDECs[2]	N	N	N	N	Y	Y
HEVC 444 decoding	N	N	N	N	Y	Y
AV1 Main Profile decoding	N	N	N	N	N	Y

- **Y**: Supported, **N**: Unsupported
- [1]: Present in select GPUs
- [2]: Present in select GPUs
- [3]: GA10x GPUs include all GPUs based on Ampere architecture except GA100

1.3. NVDEC Performance

NVDEC natively supports multiple hardware decoding contexts with negligible context-switching penalty. As a result, subject to the hardware performance limit and available memory, an application can decode multiple videos simultaneously.

The hardware and software maintain the context for each decoding session, allowing many simultaneous decoding sessions to run in parallel with minimal context switch penalty. Table 2 provides indicative data of the decoding performance of NVDEC in GPUs based on Maxwell, Pascal, Turing and Ampere architectures for AV1, HEVC, VP9, and H.264 encoded bitstreams. The performance varies across GPU classes (e.g. Quadro, Tesla), and scales (almost) linearly with the clock speeds for each hardware.

Table 2. NVDEC decoding performance (indicative)

GPU Architecture	Codec	Performance in frames/second
Pascal(GTX1060)	H.264	696
	VP9	835
	HEVC	803
	HEVC Main10	787
Turing (RTX8000)	H.264	719
	VP9	864
	VP9 10 bit	871
	HEVC	1247
	HEVC Main10	1145
Ampere (RTX3090)	H.264	742
	VP9	1069
	VP9 10 bit	1116
	HEVC	1419
	HEVC Main10	1323
	AV1	849
Ada (RTX4090)	H.264	883
	VP9	1265
	VP9 10 bit	1322
	HEVC	1666
	HEVC Main10	1549
	AV1	1005

- All the measurement is done on the highest video clocks as reported by nvidia-smi (i.e. 1129 MHz, 1683 MHz, 1755 MHz, 1770 MHz for M2000, P2000, RTX8000 and RTX3090 respectively). The performance should scale according to the video clocks as reported by nvidia-smi for other GPUs of every individual family. Information on nvidia-smi can be found at https://developer.nvidia.com/nvidia-system-management-interface .
- Resolution/Input format: 1920x1080/YUV 4:2:0
- Software: Windows 11, Video Codec SDK v12.2, NVIDIA display driver: 551.76
- Hopper and GA100 GPUs contain NVDEC with same architecture as Turing. As a result, the decoding performance on Hopper and GA100 GPUs is same as that of Turing GPUs, scaled by the clock speed. To view the clocks available on your GPU, please use the tool nvidia-smi included with the NVIDIA driver.

While Maxwell, Pascal, and Volta generation GPUs had one NVDEC engine per chip, some GPUs based on Turing, Ampere, Ada and Hopper architecture have multiple NVDEC engines per chip. GH100 has 8 NVDECs. This increases the aggregate decoding throughput of the GPU. The NVIDIA driver takes care of load balancing among multiple NVDEC engines on the chip so that applications don't require special code to take advantage of multiple decoders, and

automatically benefit from higher decoder capacity on higher-end GPU hardware. The decode performance listed in Table 2 is given per NVDEC engine. Thus, if a Quadro or Tesla GPU has 2 NVDECs, multiply the corresponding number in Table 2 by the number of NVDECs per chip to get aggregate maximum performance (applicable only when running multiple simultaneous decode sessions). Note that performance with a single decoding session cannot exceed performance per NVDEC, regardless of the number of NVDECs present on the GPU. All GeForce products consist of a single NVDEC.

1.4. Programming NVDEC

Refer to the SDK release notes for information regarding the required driver version.

Various capabilities of NVDEC are exposed to the application software via the NVIDIA proprietary application programming interface (NVDECODE APIs). Refer to the Video Decoder Programming guide for details on using these APIs.

For a complete list of GPUs supporting hardware accelerated decoding refer to https://developer.nvidia.com/nvidia-video-codec-sdk.

1.5. FFmpeg Support

FFmpeg is the most popular multimedia transcoding tool used extensively for video and audio transcoding.

The video hardware accelerators in NVIDIA GPUs can be effectively used with FFmpeg to significantly speed up the video decoding, encoding and end-to-end transcoding at very high performance.

Note that FFmpeg is open-source project and its usage is governed by specific licenses and terms and conditions.

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 acknowledgment, 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.

Trademarks

NVIDIA, the NVIDIA logo, and cuBLAS, CUDA, CUDA Toolkit, cuDNN, DALI, DIGITS, DGX, DGX-1, DGX-2, DGX Station, DLProf, GPU, Jetson, Kepler, Maxwell, NCCL, Nsight Compute, Nsight Systems, NVCaffe, NVIDIA Deep Learning SDK, NVIDIA Developer Program, NVIDIA GPU Cloud, NVLink, NVSHMEM, PerfWorks, Pascal, SDK Manager, Tegra, TensorRT, TensorRT Inference Server, Tesla, TF-TRT, Triton Inference Server, 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