NVIDIA VIDEO CODEC SDK - DECODER

Application Note

Table of Contents

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. NVDEC 支持远超实时速率的解码，因此除视频播放外，它也适用于转码场景。

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).

本文档仅重点说明当前视频编解码 SDK 软件包相较于之前版本的变更.

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	Blackwell GPUs
VC1 Simple, Main & Advanced profiles	Y	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	Blackwell GPUs
MPEG4 Simple and Advanced Simple Profiles	Y	Y	Y	Y	Y	Y	Y
MPEG2 Simple & Main profiles	Y	Y	Y	Y	Y	Y	Y
H.264 Baseline, Main, High Profiles	Y	Y	Y	Y	Y	Y	Y
VP8	N	Y	Y[1]	Y	Y	Y	Y
HEVC Main and Main 10 Profile[1]	N	Y[1]	Y	Y	Y	Y	Y
VP9 Profile 0[1]	N	Y[1]	Y	Y	Y	Y	Y
8192x8192 Decoding support (HEVC&VP9 only)	N	N	Y[1]	Y	Y	Y	Y
Multiple NVDECs[2]	N	N	N	N	Y	Y	Y
HEVC 444 decoding	N	N	N	N	Y	Y	Y
AV1 Main Profile decoding	N	N	N	N	N	Y	Y
8192x8192 Decoding support (H264)	N	N	N	N	N	N	Y
H264 High10/ High422 profiles	N	N	N	N	N	N	Y
HEVC main 422 10/12 profiles	N	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 性能

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.

NVDEC 原生支持多个硬件解码上下文，且上下文切换开销极小。因此，在不超出硬件性能限制与可用内存的前提下，应用程序可同时解码多个视频。

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.

硬件与软件会为每个解码会话维护上下文，使得多个解码会话可并行运行，且上下文切换开销极低。表 2 提供了基于 Maxwell、Pascal、Turing 与 Ampere 架构的显卡中，NVDEC 针对 AV1、HEVC、VP9 及 H.264 编码码流的解码性能参考数据。不同类别显卡（如 Quadro、Tesla）的性能存在差异，且每种硬件的性能均随时钟频率（几乎）呈线性增长。

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	H.264	694
	VP9	846
	HEVC	810
	HEVC Main10	789
Turing	H.264	771
	VP9	932
	VP9 10 bit	925
	HEVC	1316
	HEVC Main10	1158
Ampere	H.264	748
	VP9	1075
	VP9 10 bit	1120
	HEVC	1415
	HEVC Main10	1299
	AV1	790
Ada	H.264	903
	VP9	1290
	VP9 10 bit	1342
	HEVC	1641
	HEVC Main10	1520
	AV1	1018
Blackwell	H.264	2172
	VP9	1445
	VP9 10 bit	1498
	HEVC	1872
	HEVC Main10	1818
	AV1	1119

- All the measurement is done on the highest video clocks as reported by nvidia-smi (i.e. 1544 MHz, 1860 MHz, 1665 MHz, 2160 MHz, 2362 MHz for Pascal, Turing, Ampere, Ada, and Blackwell respectively). The performance should scale according to the video clocks as reported by nvidia-smi on target GPU. Information on nvidia-smi can be found at https://developer.nvidia.com/nvidia-system-management-interface .
- Resolution/Input format: 1920x1080/YUV 4:2:0

所有性能测试均在 nvidia-smi 报告的最高视频时钟频率下进行（即 Pascal、Turing、Ampere、Ada 与 Blackwell 显卡的频率分别为 1544 MHz、1860 MHz、1665 MHz、2160 MHz、2362 MHz）。目标显卡的性能应随 nvidia-smi 报告的视频时钟频率成比例变化

- Software: Windows 11, Video Codec SDK v13.0
- 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.

Hopper 与 GA100 显卡搭载的 NVDEC 采用与 Turing 架构相同的设计，因此 Hopper 与 GA100 显卡的解码性能与 Turing 显卡一致，并随时钟频率成比例变化。若需查看显卡支持的时钟频率，可使用 NVIDIA 驱动附带的工具 nvidia-smi。

While Maxwell, Pascal, and Volta generation of GPUs had one NVDEC engine per chip, some of GPUs based on Turing, Ampere, Ada, Hopper and Blackwell architecture have multiple NVDEC engines per chip. GH100 and GB100 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.

Maxwell、Pascal 与 Volta 系列显卡每块芯片仅配备一个 NVDEC 引擎，而部分基于 Turing、Ampere、Ada、Hopper 与 Blackwell 架构的显卡每块芯片配备多个 NVDEC 引擎（例如 GH100 与 GB100 显卡配备 8 个 NVDEC 引擎）。这一设计提升了显卡的总解码吞吐量。NVIDIA 驱动会负责芯片上多个 NVDEC 引擎之间的负载均衡，因此应用程序无需编写特殊代码即可利用多解码器优势，且能自动借助高端显卡硬件提供的更高解码容量。

表 2 中列出的解码性能为单个 NVDEC 引擎的性能。因此，若某款 Quadro 或 Tesla 显卡配备 2 个 NVDEC 引擎，可将表 2 中对应数值乘以每块芯片的 NVDEC 引擎数量，得到总最大性能（该计算方式仅适用于同时运行多个解码会话的场景）。需注意，无论显卡配备多少个 NVDEC 引擎，单个解码会话的性能均无法超过单个 NVDEC 引擎的性能。所有 GeForce 系列产品均仅配备一个 NVDEC 引擎。

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.

FFmpeg 是最主流的多媒体转码工具，广泛用于视频与音频转码。

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.

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