NVIDIA VIDEO ENCODER APPLICATION NOTE

NVENC_DA-06209-001_v07| November 2015

Application Note

DOCUMENT CHANGE HISTORY

NVENC_DA-06209-001_v07

Version	Date	Authors	Description of Change	Highlights
01	Jan 30,2012	AP/CC	Initial release	Initial Support for Kepler NVENC
02	Sept 24, 2012	AP	Updated for NVENC SDK release 2.0	Additional features on Kepler NVENC
03	April 10, 2013	AP	Updated for Monterey SDK 2.0.0 update	Additional features on Kepler NVENC
04	Aug 4, 2013	AP	Updated for NVENC SDK release 3.0	New APIs added to SDK
05	June 17, 2014	SM/AP	Updated for NVENC SDK release 4.0	Software Support for First generation Maxwell GPUs
06	Nov 14, 2014	SM	Updated for NVENC SDK release 5.0	Software Support for Second generation Maxwell GPUs
07	Oct 10, 2015	SM	Updated for NVENC SDK Release 6.0	Support for additional software features

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NVIDIA HARDWARE VIDEO ENCODER

1. INTRODUCTION

NVIDIA GPUs - beginning with the Kepler generation - contain a hardware-based encoder (referred to as NVENC in this document) which provides fully-accelerated hardware-based video encoding, and is independent of graphics performance. With complete encoding (which is computationally complex) offloaded to NVENC, the graphics engine and the CPU are free for other operations. For example, in a game recording scenario, encoding being completely offloaded to NVENC makes the graphics engine bandwidth fully available for game rendering.

NVIDIA's latest generation of GPUs, based on the second-generation Maxwell architecture, support full hardware acceleration for High Efficiency Video coding (also known as HEVC or H.265) along with support for H.264 encoding.

With the current SDK, H.264 Motion Estimation only mode support has been added. Using this feature, the codecs not supported by NVENC can be accelerated by offloading Motion Estimation to NVENC and the rest of the encoding operations can be done using the CPU and/or CUDA.

The hardware capabilities available in NVENC are exposed through APIs referred to as NVENCODE APIs in the document.

This document provides information about the capabilities of the hardware encoder and features exposed through NVENCODE APIs. The current document *only* highlights the changes in the current NVENC SDK package with respect to the previous SDK packages. In order to know about the features exposed in earlier SDKs please refer to the earlier SDK package(s). The driver supporting NVENC SDK 6.0 is completely *backward compatible* with earlier SDKs, which means that applications compiled with earlier

NVENC header version(s) can be expected to work "as-is" with the driver supporting NVENC SDK 6.0 and beyond.

2. NVENC BLOCK DIAGRAM

Apart from the rate control and picture type decision, NVENC can perform all tasks that are a critical part of the end-to-end H.264 and H.265 encoding. The rate control algorithm is implemented in the GPU's firmware and controlled via the driver. From the application's perspective, rate control is a hardware function controlled via the parameters exposed in the NVENCODE APIs. The hardware also provides the ability to use external motion estimation engine and custom quantization parameter (QP) maps (for "Region of Interest" encoding). The region of interest encoding has been made available using the "QP delta map" where the quantization parameters derived from the rate control algorithm can be tweaked using the QP delta map.

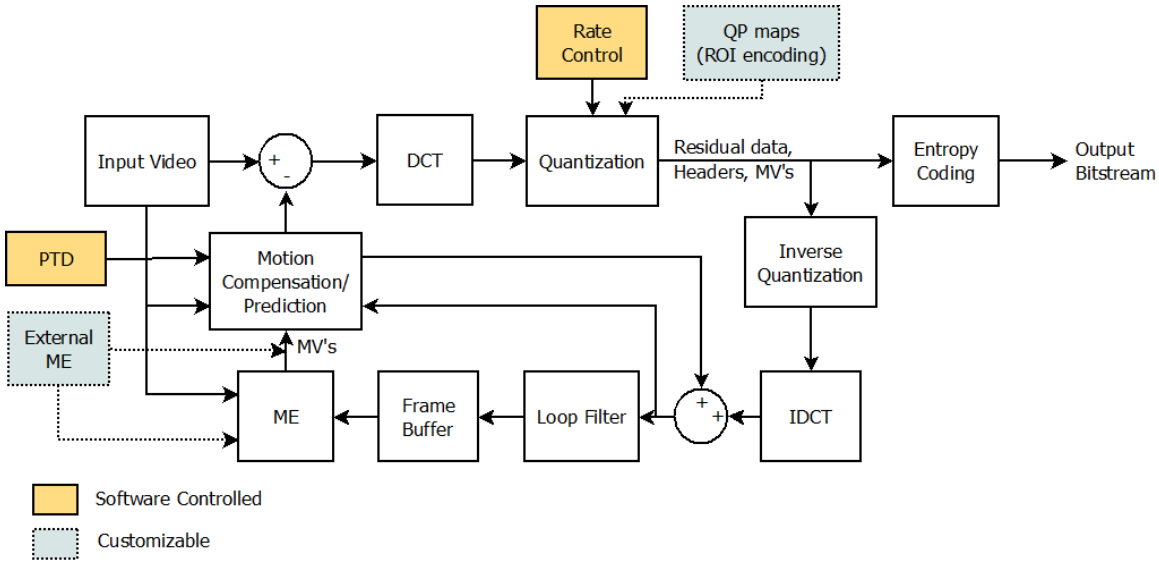

Figure 1. NVENC hardware block diagram

3. NVENC CAPABILITIES

At a high level, capabilities of the NVENC hardware are summarized in Table 1.

Table 1. NVENC Hardware Capabilities

Feature	Description	Kepler GPUs	First generation Maxwell GPUs	Second generation Maxwell GPUs
H.264 Base, Main, High Profiles	YUV 4:2:0 Encoding.	✓	✓	✓
H.264 4:4:4 and Lossless	Regular YUV 4:4:4 and Lossless Encoding.	×	✓	✓
H.265 Main Profile	YUV 4:2:0 Encoding.	×	×	✓
H264 Motion Estimation (ME) only Mode	Capability to provide Macroblock level motion vectors and intra/inter modes	×	✓	✓
Support for ARGB Input	Capability to accept RGB input for limited color spaces.	✓	✓	✓

Table 2. New features in the NVENC API for NVIDIA Video Codec SDK 6.0

Additional Software features	Description
H.264 Motion Estimation(ME) Only Mode	NVENC APIs are exposed to get the motion vectors and mode information from NVENC. This feature would be useful in cases where ONLY motion vector (and/or mode decision) information are needed, such as for motion compensated filtering or for offloading motion estimation and mode decision evaluation for codecs not supported by NVENC.
H.265 Adaptive Quantization	Support for H.265 adaptive quantization. Adaptive Quantization (AQ) helps improve visual quality.
Support for RGB input format	Support for accepting input in RGB format for limited number of color spaces.
Licensing Policy for NVEncodeAPI	There has been a change in policy from the earlier SDK. Please refer section 4.
Support for VUI and SEI parameters for H.265	This feature provides the capability of embedding custom SEI and VUI messages for H.265.
Bug fixes and quality improvements	There have been several bug fixes and quality improvements since the last SDK release.

4. NVENC LICENSING POLICY

The drivers supporting NVENC SDK 6.0 (R358 and above) have changes in the licensing policy for systems having a mix of qualified and non-qualified hardware. For the purposes of this discussion, non-qualified hardware is defined as any GeForce GPUs or low-end Quadro GPUs (for a complete list, refer to https://developer.nvidia.com/nvidia-video-codec-sdk).

Following is the difference between licensing policies staring R358 driver(s) and the previous driver(s). In driver(s) earlier to R358, two simultaneous encoding sessions per system were allowed irrespective of the number of non-qualified GPUs present in the system. For example, on systems with one Quadro K4000 card and two GeForce cards, only two simultaneous encode sessions could be run, regardless of which GPU is running the encode session. Starting R358, the restriction of two simultaneous encoding sessions will apply *only* to the *combined number of sessions executed on all non-qualified cards*. In the example above, the application can run *N* simultaneous encode sessions on Quadro K4000 card (where *N* is defined by the encoder/memory/hardware limitations) and two sessions on both GeForce cards combined. Thus the limit on the number of simultaneous encode sessions for such a system is *N* + 2.

5. NVENC PERFORMANCE

The second-generation Maxwell NVENC hardware improves standalone encoding performance compared to earlier NVENC versions. The application can trade performance for encoded picture quality.

While Kepler and first generation Maxwell GPUs had one NVENC engine, certain variants of the second generation Maxwell GPUs have two NVENC engines physically present on the silicon. That enables clients to support a greater number of concurrent encoding sessions. The underlying software implementation takes care of the load balancing between the two engines so that applications don't require changes in their own software stack to take advantage of both the engines.

NVENC hardware natively supports multiple hardware encoding contexts with negligible context-switching penalty. As a result, subject to the hardware performance limit and available memory, an application can encode multiple videos simultaneously. The hardware and software maintain the context for each encoding session, allowing a large number of simultaneous encoding sessions to run in parallel.

NVENC API exposes several presets and rate control modes for programming the hardware. A combination of these two parameters enables video encoding at varying quality and performance.

Note that the encoder performance is a function of several parameters. Table 3 provide an indicative data of NVENC performance on Kepler and Maxwell GPUs for different presets and rate control modes.

The hardware has been extensively tested and verified to yield the advertised performance at all settings.

Table 3. NVENC Encoding Performance

Preset	Rate control mode	H.264 (FPS)			H.265(FPS)
		Kepler	First Gen. Maxwell	Second Gen. Maxwell	Second Gen. Maxwell
High Performance	Constant QP	520	833	1111	526
	Single Pass	490	826	1111	552
	Dual Pass	247	515	653	400
High Quality	Constant QP	157	512	653	292
	Single Pass	157	502	641	347
	Dual Pass	99	280	352	133
Low latency High Performance	Constant QP	311	549	653	523
	Single Pass	296	549	653	549
	Dual Pass	161	375	448	400
Low latency High Quality	Constant QP	120	483	645	436
	Single Pass	120	462	632	460
	Dual Pass	87	270	349	208

FPS: Encoding speed in "Frames per second". Resolution/Format: 1280x720/ YUV 4:2:0

6. PROGRAMMING NVENC

Various capabilities of NVENC are exposed to the application software via the NVIDIA proprietary application programming interface (NVENC API). Please refer to the Vidoe Encoder NVENC Programming guide for details on using the APIs to accelerate video encoding with NVENC hardware.

Notice

Trademarks

Copyright