NVENC – NVIDIA HARDWARE VIDEO ENCODER

NVENC_DA-06209-001_v04| July 2014

Application Note

DOCUMENT CHANGE HISTORY

NVENC_DA-06209-001_v04

Version	Date	Authors	Description of Change
01	January 30,2012	AP/CC	Initial release
02	September 24, 2012	AP	Updated for NVENC SDK release 2.0
03	April 10, 2013	AP	Updated for Monterey SDK 2.0.0 update
04	August 4, 2013	AP	Updated for NVENC SDK release 3.0
05	June 17, 2014	SM/AP	Updated for NVENC SDK release 4.0

TABLE OF CONTENTS

NVIDIA Hardware Video Encoder (NVENC) 5
1. Introduction 5
2. NVENC Capabilities 6
   2.1 Block Diagram 8
   2.2 Performance 9
3. Programming NVENC 10
4. Performance NUMBERS 11

LIST OF FIGURES

Figure 1. NVENC hardware block diagram 8

LIST OF TABLES

Table 1. NVENC Hardware Capabilities 6
Table 2. Additional NVENC Hardware Capabilities in Maxwell GPUs 6
Table 3. Additional Software Features in SDK4.0 7
Table 4 Comparison between NVENC SDK and GRID SDK Capabilities 10

NVIDIA HARDWARE VIDEO ENCODER (NVENC)

1. INTRODUCTION

NVIDIA's latest generation of GPUs based on the Kepler and Maxwell architecture, contain a hardware-based H.264 video encoder (henceforth referred to as NVENC). This document provides information about the capabilities of the hardware encoder, along with some relevant data about encoding quality and performance.

Before Kepler GPUs, the only GPU based solution for video encoding was to do encoding using CUDA. One of the disadvantages of the CUDA-based encoder is that it uses a combination of the CPU and GPU's graphics engine for encoding, taking away processing power from other tasks that can be performed on the CPU and GPU's graphics engine. This approach also increased overall system power consumption.

NVENC, being dedicated H.264 hardware on the GPU chip, does not use the GPU's graphics engine and hence uses much less power compared to the CUDA-based encoder. It also leaves the CPU and GPU graphics engine to perform other tasks. The hardware is optimized to provide excellent quality at high performance, enabling a wide range of applications that require video encoding capabilities. The later versions of NVENC present the Maxwell class of GPUs further improve the encoding performance and also provide several additional features.

As explained in Section 3, NVENC hardware's encoding capabilities can be accessed via NVENC API and GRID API. Although there is some overlap in the functionality provided by these two SDK's, they are designed for slightly different use-cases (explained further in Section 3).

2. NVENC CAPABILITIES

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

Table 1. NVENC Hardware Capabilities

Feature	What it Provides
Supported codec	H.264
H.264 base, main, high profiles	Wide range of use-cases
Up to 8x HD encode (1080p @ 240 fps)	Faster than real-time encoding
Flexible ME, QP maps	Customizable quality, (ROI) region of interest encoding
YUV 4:2:0 and planar 4:4:4 support	High-quality encoding with and without chroma subsampling
MVC	Full-resolution stereo encode
Up to 4096 × 4096 in hardware	High resolution encode
API	NVENC SDK (Flexible API available on Windows and Linux)

The first Generation Maxwell GPUs support all Kepler NVENC features along with the following additional features below.

Table 2. Additional NVENC Hardware Capabilities in Maxwell GPUs

Additional Feature in first Generation Maxwell GPUs	What it provides
H.264 Lossless Encoding	The input YUV content can be encoded as lossless. This can be useful for the use case where it is desirable to have compression without any loss of quality compared to the source input.
H.264 Regular YUV 4:4:4	Maxwell hardware can encode Regular YUV 444 content. This avoids the side-effect of chroma sub-sampling, such as loss of detail in small pitched text or sharp edges.
Enhanced Performance	Encoding Performance is greatly improved. Section 4 compares the performance numbers Of Maxwell NVENC with respect to Kepler Generation of NVENC.
Enhanced two pass encoding	There are scenarios where NVENC SW stack supports two pass encoding. The first generation Maxwell hardware provides hardware architectural improvements which improve performance significantly with the same or better visual quality.
Enhanced Quality	Maxwell Hardware provides improvements in motion estimation logic, improving overall video quality.

Table 3. Additional Software Features in SDK4.0

Additional Software features in	What it provides
Software support for all the new Maxwell features mentioned in the table above	APIs exposed to use the new features added in Maxwell NVENC.
Adaptive Quantization	This is a SW feature that defines which quantization parameters to be used and changed within a row. The regular NVENC rate control is row-based. This feature helps in situations where there is a change in textures within a row.
Intra Refresh	This feature can be used to generate waves of rows of Intra macro blocks. This is useful to gradually recover from errors that may have happened on client side.
Advanced Rate control	The Rate control algorithm provides enhanced Quality in comparison to earlier SDK Releases.
Support for 2 NVENC sessions in GeForce and Low end Quadro Hardware	The current SDK package allows 2 NVENC sessions for Low end Quadro and Geforce cards on Windows OS only.
Several bug fixes from past SDK release	

The NVENC hardware is designed to accept YUV (NV12) picture data and output a H.264 elementary encoded bit-stream, as per the specified settings. The hardware itself provides the ability to control the range of encoding parameters from software, some of which are exposed via the software API's in NVENC SDK (refer to Section 3). Every GPU from NVIDIA's Kepler and Maxwell family has a separate NVENC engine that is independent of the graphics engine. The NVENC engine runs at the same clock speed, and its performance is independent of the graphics performance.

2.1 Block Diagram

Figure 1 shows the block diagram of NVENC. 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 encoding. The rate control algorithm is implemented in 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 NVENC APIs. The hardware also provides capability to use external motion estimation engine and custom quantization parameter maps (for ROI "region of interest" encoding). These features, however, are currently not exposed in the software APIs and will be available in future releases of the SDK.

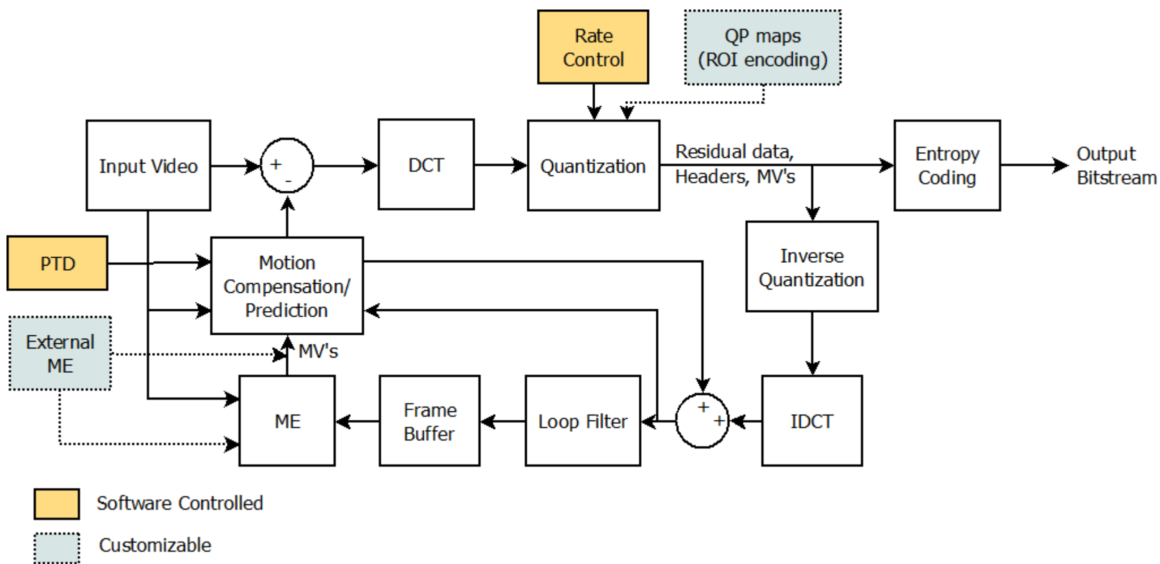

Figure 1. NVENC hardware block diagram

2.2 Performance

The Maxwell NVENC hardware doubles the encoder performance as compared to Kepler NVENC. Maxwell NVENC can support up to 16x real-time HD video encoding (1x HD = 1080p @ 30 fps). This means that the hardware can encode up to 480 frames per second of 1920 × 1080 progressive video in highest performance mode (HP preset). The application can trade performance for encoded picture quality.

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. For all GeForce hardware and some low-end Quadro hardware, the number of simultaneous encoding sessions is limited to 2.

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. For example, the presets with the prefix *LOW_LATENCY* are useful for applications that require very low-latency encoding (e.g. real-time streaming or remote interactive applications). Similarly, 2-pass rate control modes help the encoder to gather statistics of the frame to be encoded before actually encoding it in the second pass, thereby resulting in optimal bit-utilization within the frame and consequently, higher encoding quality.

Note that the encoder performance is a function of several parameters. Refer to Section 4 which provides 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. The performance does not vary if using motion video or synthetically-generated content (e.g. gameplay, desktop). But video quality and latency requirements for different types of content may be significantly different. This can affect the overall encoding performance either positively or negatively which is determined based on the NVENC parameter settings.

3. PROGRAMMING NVENC

Various capabilities of NVENC are exposed to the application software via the NVIDIA proprietary application programming interface (API). There are two API's available to use NVENC encoding capabilities:

1. NVENC SDK – Useful for direct encoding applications such as video conferencing, transcoding, video editing, archiving etc.

2. GRID SDK – Useful for screen capture + encoding use-cases such as cloud gaming, streaming etc.

Table 4 Comparison between NVENC SDK and GRID SDK Capabilities

Direct Encode - NVENC SDK	**Capture + Encode - GRID SDK**
No capture – H.264 encode only	Capture + H.264 encode
Use cases: Transcoding, archiving, video conferencing, video editing, camera capture and encoding	Use cases: Low-latency applications such as cloud gaming, streaming where a single API performs screen capture + encode in most optimized manner
Linux and Windows	Linux and Windows
Access to exhaustive encoder settings and fine-grained control	Limited encoder settings, applicable to only low-latency streaming use-cases
Available via NVIDIA developer zone at https://developer.nvidia.com/nvidia-video-codec-sdk	Available under license from NVIDIA
Works on GeForce, Quadro, Tesla, and GRID boards. For low end Quadro and GeForce boards two sessions of NVENC are allowed.	Works on Quadro and GRID boards. For low end Quadro and GeForce boards two sessions of NVENC are allowed.

4. PERFORMANCE NUMBERS

Resolution	Preset	RC Mode	FPS- Maxwell	FPS - Kepler
YUV420, 1080p	Low Latency HP	NV_ENC_PARAMS_RC_CBR2	360.88	283.493
		NV_ENC_PARAMS_RC_CBR	331.922	275.82
		NV_ENC_PARAMS_RC_2_PASS_FRAMESIZE_CAP	252.323	146.346
		NV_ENC_PARAMS_RC_VBR	324.263	270.051
		NV_ENC_PARAMS_RC_2_PASS_QUALITY	249.076	144.927

Resolution	Preset	RC Mode	FPS- Maxwell	FPS - Kepler
YUV420, 1080p	Low Latency HQ	NV_ENC_PARAMS_RC_CBR2	216.213	109.848
		NV_ENC_PARAMS_RC_CBR	210.807	103.269
		NV_ENC_PARAMS_RC_2_PASS_FRAMESIZE_CAP	148.316	88.62
		NV_ENC_PARAMS_RC_VBR	209.395	102.951
		NV_ENC_PARAMS_RC_2_PASS_QUALITY	145.013	87.278

Resolution	Preset	RC Mode	FPS- Maxwell	FPS - Kepler
YUV420, 1080p	High Performance	NV_ENC_PARAMS_RC_CBR2	514.362	493.197
		NV_ENC_PARAMS_RC_CBR	489.396	492.28
		NV_ENC_PARAMS_RC_2_PASS_FRAMESIZE_CAP	334.289	128.639
		NV_ENC_PARAMS_RC_VBR	496.093	492.5
		NV_ENC_PARAMS_RC_2_PASS_QUALITY	333.648	127.449

Resolution	Preset	RC Mode	FPS- Maxwell	FPS - Kepler
YUV420, 1080p	High Quality	NV_ENC_PARAMS_RC_CBR2	260.444	245.253
		NV_ENC_PARAMS_RC_CBR	254.064	236.071
		NV_ENC_PARAMS_RC_2_PASS_FRAMESIZE_CAP	100.982	58.682
		NV_ENC_PARAMS_RC_VBR	252.815	231.034
		NV_ENC_PARAMS_RC_2_PASS_QUALITY	98.869	57.372

Resolution	Preset	RC Mode	FPS- Maxwell	FPS - Kepler
YUV420, 1080p	Lossless HP	NA	291.779	NA
	Lossless HQ	NA	206.697	NA

Resolution	Preset	RC Mode	FPS- Maxwell	FPS - Kepler
Regular YUV444, 1080p	Low Latency HP	NV_ENC_PARAMS_RC_CBR2	131.86	NA
		NV_ENC_PARAMS_RC_CBR	128.304	NA
		NV_ENC_PARAMS_RC_2_PASS_FRAMESIZE_CAP	95.845	NA
		NV_ENC_PARAMS_RC_VBR	127.89	NA
		NV_ENC_PARAMS_RC_2_PASS_QUALITY	96.005	NA

Resolution	Preset	RC Mode	FPS- Maxwell	FPS - Kepler
Regular YUV444, 1080p	Low Latency HQ	NV_ENC_PARAMS_RC_CBR2	83.038	NA
		NV_ENC_PARAMS_RC_CBR	83.194	NA
		NV_ENC_PARAMS_RC_2_PASS_FRAMESIZE_CAP	63.913	NA
		NV_ENC_PARAMS_RC_VBR	80.49	NA
		NV_ENC_PARAMS_RC_2_PASS_QUALITY	63.83	NA

Resolution	Preset	RC Mode	FPS- Maxwell	FPS - Kepler
Regular YUV444, 1080p	High Performance	NV_ENC_PARAMS_RC_CBR2	155.185	NA
		NV_ENC_PARAMS_RC_CBR	157.029	NA
		NV_ENC_PARAMS_RC_2_PASS_FRAMESIZE_CAP	107.539	NA
		NV_ENC_PARAMS_RC_VBR	151.976	NA
		NV_ENC_PARAMS_RC_2_PASS_QUALITY	107.39	NA

Resolution	Preset	RC Mode	FPS- Maxwell	FPS - Kepler
YUV444, 1080p	High Quality	NV_ENC_PARAMS_RC_CBR2	92.378	NA
		NV_ENC_PARAMS_RC_CBR	91.787	NA
		NV_ENC_PARAMS_RC_2_PASS_FRAMESIZE_CAP	35.457	NA
		NV_ENC_PARAMS_RC_VBR	91.873	NA
		NV_ENC_PARAMS_RC_2_PASS_QUALITY	35.407	NA

Resolution	Preset	RC Mode	FPS- Maxwell	FPS - Kepler
YUV444, 1080p	Lossless HP	NA	117.861	NA
	Lossless Default	NA	72.618	NA

Notice

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.

Information furnished is believed to be accurate and reliable. However, NVIDIA Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication of otherwise under any patent rights of NVIDIA Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all other information previously supplied. NVIDIA Corporation products are not authorized as critical components in life support devices or systems without express written approval of NVIDIA Corporation.

HDMI

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

ROVI Compliance Statement

NVIDIA Products that support Rovi Corporation's Revision 7.1.L1 Anti-Copy Process (ACP) encoding technology can only be sold or distributed to buyers with a valid and existing authorization from ROVI to purchase and incorporate the device into buyer's products.

This device is protected by U.S. patent numbers 6,516,132; 5,583,936; 6,836,549; 7,050,698; and 7,492,896 and other intellectual property rights. The use of ROVI Corporation's copy protection technology in the device must be authorized by ROVI Corporation and is intended for home and other limited pay-per-view uses only, unless otherwise authorized in writing by ROVI Corporation. Reverse engineering or disassembly is prohibited.

OpenCL

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

Trademarks

NVIDIA, the NVIDIA logo, GeForce, and Quadro 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.

Copyright