NVIDIA VIDEO ENCODER 5.0

NVENC_DA-06209-001_v06| November 2014

Application Note

DOCUMENT CHANGE HISTORY

NVENC_DA-06209-001_v06

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

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

NVIDIA HARDWARE VIDEO ENCODER 5.0

1. INTRODUCTION

NVIDIA GPUs - beginning with the Kepler generation - contain a hardware based encoder (also referred to as NVENC hence forth in the document) which provides fully accelerated hardware based video encoding.

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

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 and related encoding features that were supported on earlier Kepler and first generation Maxwell GPUs. The second generation Maxwell GPUs also provide significant improvement in encoding performance in comparison to earlier generations of NVENC. This improvement in performance is due to improvements in architecture within NVENC hardware. In order to support more number of simultaneous encoding sessions an extra NVENC had been added on certain variants of the second generation of Maxwell GPUs. The hardware capabilities available in NVENC are exposed through APIs referred to as "Encode APIs" or "NVENC API" in the document.

This document provides information about the capabilities of the hardware encoder and features exposed though Encode APIs.

2. NVENC CAPABILITIES

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

Table 1. NVENC Hardware Codec Capabilities

Feature	What it provides	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.	×	×	✓

At high level, the second generation Maxwell GPUs support several additional encoding features in addition to the features supported on Kepler and first generation Maxwell GPUs which are summarized in Table 2 and Table 3.

Table 2. Additional NVENC Hardware Capabilities in Second generation Maxwell GPUs

Additional NVENC Hardware Feature in second Generation Maxwell GPUs	What it provides
H.265 Main profile	The input YUV 4:2:0 sequence can be encoded to generate a H.265 bit-stream.
Enhanced performance for H.264 encoding	Encoding performance is significantly improved for H.264 in comparison to first generation Maxwell and Kepler GPUs.
Additional NVENC engine	In order to support more number of concurrent encoder sessions, the number of encoding engines has been increased to "two" in certain variants of second generation Maxwell GPUs. The NVIDIA Software stack manages the load-balancing between the two NVENC engines which ensures that applications do not need any changes to their software stack to benefit from the extra NVENC engine.

Table 3. Major H.265 features supported in Second Generation Maxwell GPUs

Features supported	What it provides
Max supported CTB size of 32x32	
Dynamic Reconfiguration of parameters on fly	This enables Clients to change encoding parameters without destroying and re-initializing the encoding session.
Asymmetric Motion Partitioning	Captures asymmetric motion.
Static and dynamic slice modes	Clients can specify the slice size in terms of CTBs or max number of bytes to be present in a slice respectively.
All Intra modes	All Intra modes mandated in H.265 specification.
Intra-refresh, invalidation of reference pictures and force IDR frame	Error resilience features useful in streaming scenarios, when feedback loop between server and client is available.

In addition to supporting features from earlier NVENC SDK releases, NVENC SDK 5.0 adds support for H.265 encoding on second-generation Maxwell GPU hardware (e.g. GTX 980, Quadro M6000 etc.). NVENC SDK 5.0 also contains several other features and improvements from the previous SDK release, some of which are highlighted in the following table.

Table 4. Additional features supported in NVIDIA Encoder SDK 5.0

Additional Software features	What it provides
Software support for H.265 Main profile	APIs exposed to access NVENC for YUV 4:2:0 H.265 encoding. Several low latency and Error Resilience features (helpful for handling and recovering from error conditions in streaming scenarios) are exposed through Encode APIs.
Support for 2 NVENC sessions in GeForce and low-end Quadro hardware	The current SDK package allows up to two simultaneous encode sessions per system for low-end Quadro and GeForce cards. If the system contains any low-end hardware (even in conjunction with other high-end hardware), only two encoding sessions will be permitted.
Several bug fixes and quality improvements	There have been several bug fixes and quality improvements since the last SDK release which increases the encoded quality and stability of the entire software stack.

3. 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 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). The region of interest encoding has been made available using the "QP delta map" where in the Quantization parameters derived from the Rate Control algorithm can be tweaked using the QP delta map.

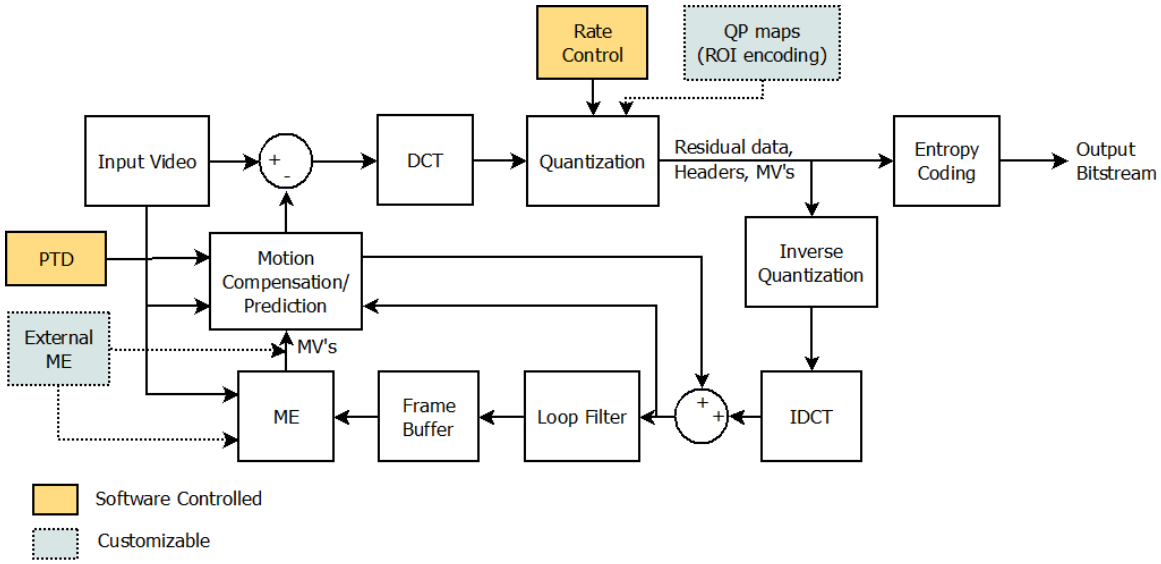

Figure 1. NVENC hardware block diagram

4. NVENC PERFORMANCE

The second-generation Maxwell NVENC hardware improves standalone encoding performance compared to first generation Maxwell NVENC and Kepler generation NVENC. The application can trade performance for encoded picture quality via the software API provided.

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 more 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 for taking 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. 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. Table 5 and Table 6 provide 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. 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 by the NVENC parameter settings.

Table 5. NVENC Encoding Performance – High Performance/High Quality Presets

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
	CBR	502	833	1123	534
	VBR	490	826	1111	552
	VBR MinQP	497	819	1111	540
	Two PassFrameSize	250	534	675	421
	Two PassQuality	250	523	680	421
	Two PassVBR	247	515	653	400
High Quality	Constant QP	157	512	653	292
	CBR	160	529	645	273
	VBR	157	502	641	347
	VBR MinQP	158	500	641	303
	Two PassFrameSize	101	300	387	167
	Two Pass Quality	101	301	390	167
	Two PassVBR	99	280	352	133

Table 6. NVENC Encoding Performance - Low Latency Presets

Preset	Rate control mode	H.264 (FPS)			H.265(FPS)
		Kepler	First Gen. Maxwell	Second Gen. Maxwell	Second Gen. Maxwell
Lowlatency High Performance	Constant QP	311	549	653	523
	CBR	310	564	666	531
	VBR	296	549	653	549
	VBR MinQP	301	549	636	540
	Two PassFrameSize	177	398	469	425
	Two PassQuality	177	392	467	423
	Two PassVBR	161	375	448	400
Lowlatency High Quality	Constant QP	120	483	645	436
	CBR	120	490	657	421
	VBR	120	462	632	460
	VBR MinQP	119	465	636	440
	Two PassFrameSize	89	288	469	232
	Two Pass Quality	89	287	467	232
	Two PassVBR	87	270	349	208

FPS: Encoding speed in "Frames per second"

Resolution/Format: 1280x720/ YUV 4:2:0

5. 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:

- NVENC SDK – Useful for direct encoding applications such as video conferencing, transcoding, video editing, archiving etc.
- GRID SDK – Useful for screen capture + encoding use-cases such as cloud gaming, streaming etc.

Table 7. Comparison between NVENC SDK and GRID SDK Capabilities

Standalone Encode – NVENC SDK	**Capture + Encode – GRID SDK**
No capture – H.264/H.265 encode only	Capture + H.264/H.265 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 cards. For GeForce and low end Quadro cards "two" encoding sessions "per-system" are allowed.	Works on Quadro, Tesla, and GRID boards only.

Notice

Trademarks

Copyright