GPU TECHNOLOGY CONFERENCE
April 4-7, 2016 | Silicon Valley

HIGH PERFORMANCE VIDEO ENCODING WITH NVIDIA GPUS

Abhijit Patait
Eric Young

April 4th, 2016

PRESENTED BY
NVIDIA.

AGENDA

NVIDIA GPU Video Technologies

Video Hardware Capabilities

Video Software Overview

Common Use Cases for Video

Performance and Quality Tuning

New Directions

SDK Links

NVIDIA GPU VIDEO TECHNOLOGIES

NVIDIA VIDEO TECHNOLOGIES

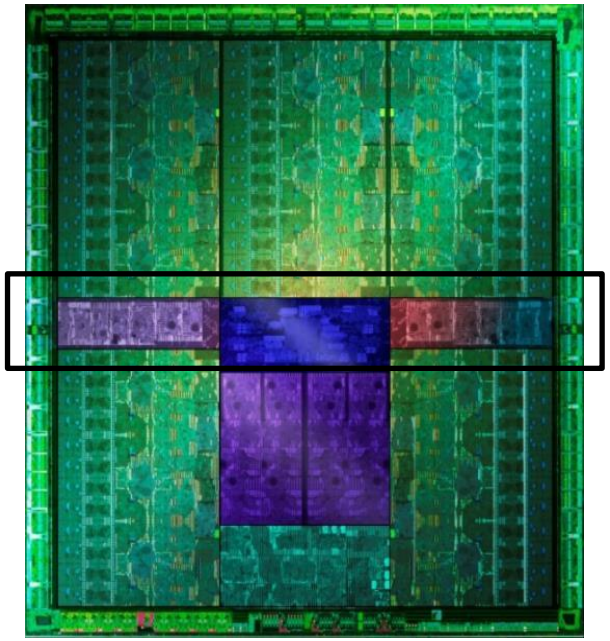
- Dedicated hardware for encode & decode
- Linux, Windows, FFMPEG

MPEG2

H.264 MPEG-4/AVC

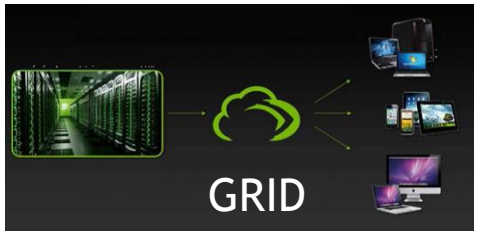
HEVC H.265 - HIGH EFFICIENCY VIDEO CODING

FFmpeg



NVIDIA VIDEO TECHNOLOGIES EVOLUTION

Low-latency Streaming



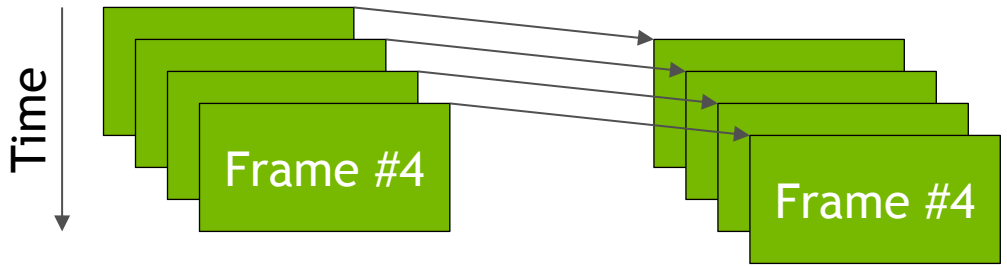
Cloud transcoding

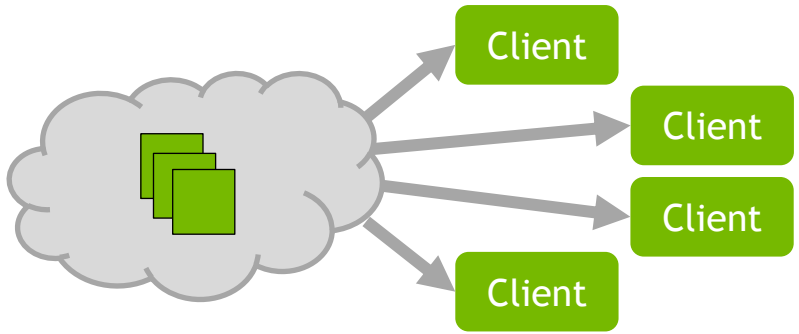
- Social media
- Live streaming
- Video-on-demand

GPU VIDEO ENCODE

Benefits

- Low power
- Low latency
- High performance and scalability
- Automatic benefit from improvements in hardware
- Linux, Windows, C/C++, FFMPEG support





VIDEO HARDWARE CAPABILITIES

NVIDIA GPU VIDEO HARDWARE

NVDEC

- Video decoder
- MPEG-2, VC-1, H.264, HEVC
- Fermi, Kepler, Maxwell, and future GPUs

NVENC

- Video encoder
- H.264, HEVC
- Kepler, Maxwell, and future GPUs

ENCODE CAPABILITIES

KEPLER (GK107, GK104)	MAXWELL GEN 1 (GM107)	MAXWELL GEN 2 (GM200, GM204, GM206)
H.264 only	H.264 only	H.264 and HEVC/H.265
Standard 4:2:0, Planar 4:4:4 & proprietary 4:4:4	Standard 4:2:0, 4:4:4 and H.264 lossless encoding	Standard 4:2:0, 4:4:4 and H.264 lossless encoding
~240 fps 2-pass encoding @ 720p	~500 fps 2-pass encoding @ 720p	~900 fps 2-pass encoding @ 720p
GRID K340/K520, K1/K2, Quadro K5000, Tesla K10/K20, GeForce GTX 680	Maxwell-based GRID & Quadro products	Tesla M4, M40, M6, M60, Quadro M4000, M5000, M6000, GeForce GTX 960, 980, Titan X
NV Encode SDK 1.0-5.0	NV Encode SDK 4.0+	NV Encode SDK 5.0 Video Codec SDK 6.0+

DECODE CAPABILITIES

KEPLER (GK107, GK104)	MAXWELL 1 (GM107, GM204, GM200)	MAXWELL 2 (GM206)
MPEG-2, MPEG-4, H.264	MPEG-2, MPEG-4, H.264, HEVC with CUDA acceleration	MPEG-2, MPEG-4, H.264 HEVC/H.265 fully in hardware
H.264: ~200 fps at 1080p; 1 stream of 4K@30	H.264: ~540 fps at 1080p 4 streams of 4K@30	H.264: ~540 fps at 1080p 4 streams of 4K@30
H.265: Not supported	H.265: Not supported	H.265: ~500 fps at 1080p 4 streams of 4K@30
Video Codec SDK 5.0+	Video Codec SDK 5.0+	Video Codec SDK 5.0+
4096 × 4096	4096 × 4096	4096 × 4096

VIDEO SOFTWARE OVERVIEW

NVIDIA VIDEO TECHNOLOGIES – PRE-2016

VIDEO DECODE/PLAYBACK

DXVA for Windows
VDPAU for Linux

NVENC SDK

Hardware encoder API
Windows, Linux
CUDA, DirectX interoperability

NVCUVID VIDEO DECODING

Windows, Linux,
CUDA interoperability

GRID/CAPTURE SDK, MFT

Use-case specific APIs

NVIDIA VIDEO TECHNOLOGIES – 2016++

VIDEO CODEC SDK

- Flexibility
- API for encode + decode
- Windows, Linux
- CUDA, DirectX, OpenGL interoperability
- High performance transcode
- Current: Video Codec SDK 6.0

FFMPEG SUPPORT*

- Hardware acceleration for most popular video and audio framework
- Leverages FFmpeg's Audio codec, stream muxing, and RTP protocols.
- Windows, Linux
- Wide adoption

FFmpeg

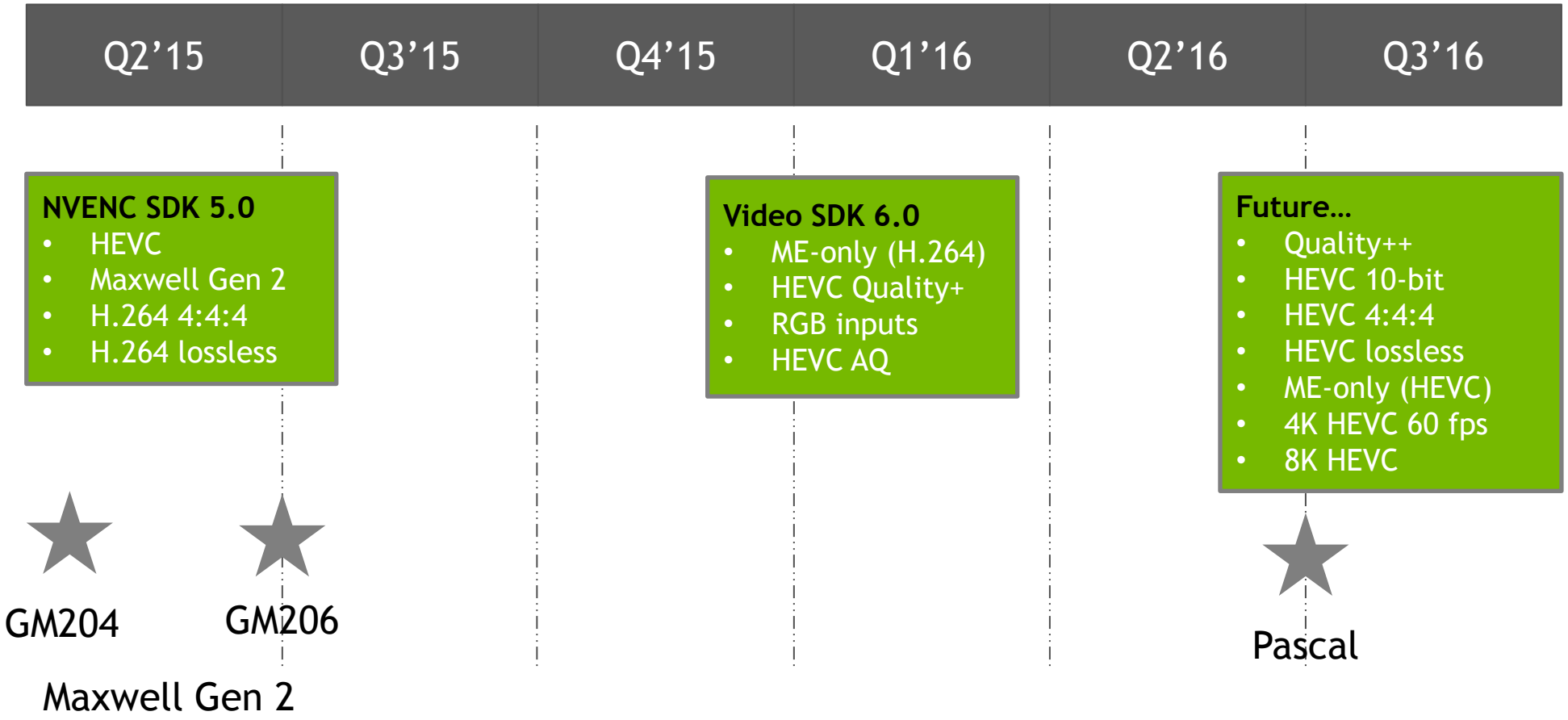
*To get access to the latest FFmpeg repository with NVENC support, please contact your NVIDIA relationship manager.

VIDEO CODEC SDK FEATURES

What's New

Feature	SDK release	Why
Video SDK = encode + decode	6.0	Transcoding
Quality++	6.0	Streaming, Transcoding, Broadcast, Video production
RGB inputs	6.0	Capture RGB + encode
Motion estimation only mode	6.0	Hardware assisted motion estimation for custom encoders, Image stabilization
Adaptive quantization	7.0	Improved perceptual quality – Available in May 2016
Adaptive B-frames		
Adaptive GOP		
Look-ahead		

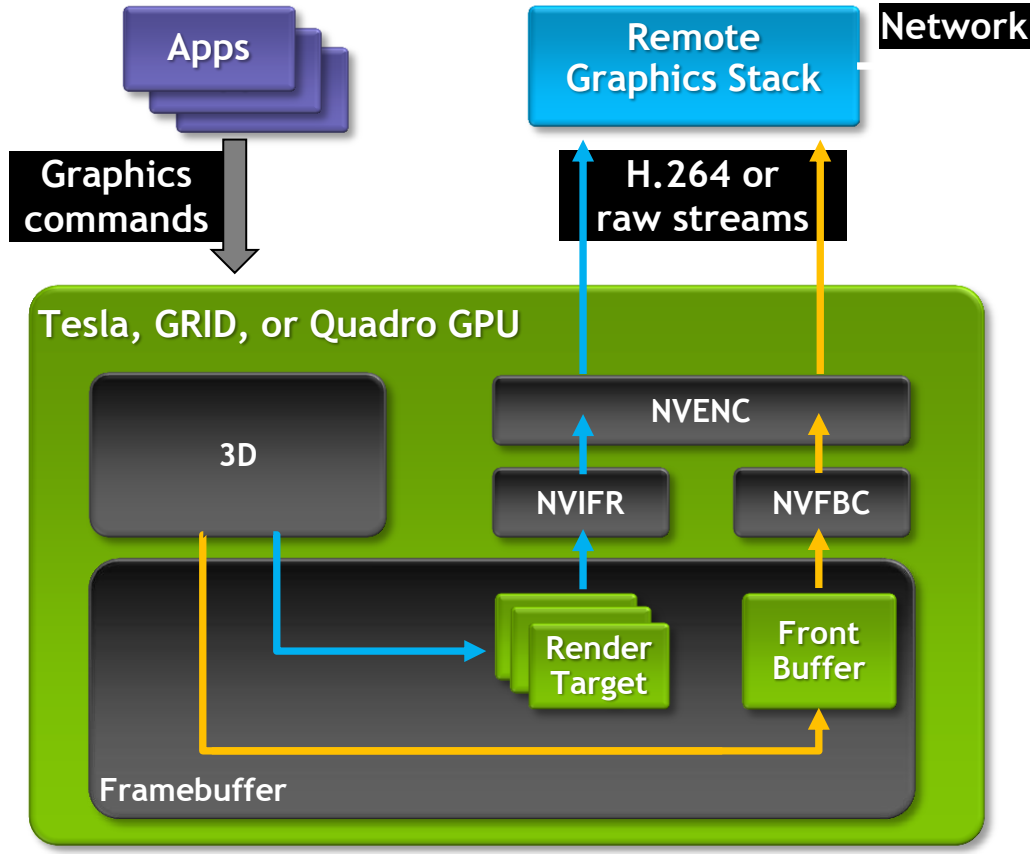
ROADMAP



COMMON USE CASES FOR VIDEO

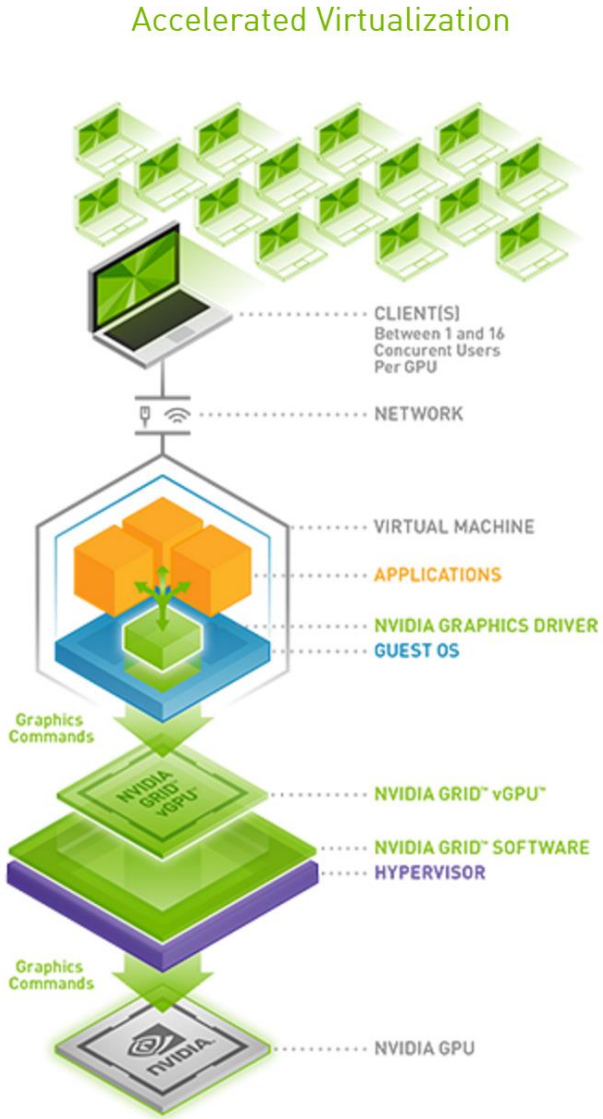
CAPTURE + ENCODE

- Capture Desktop (NvFBC) and RenderTargets (NvIFR)
- Low Latency, low CPU overhead
- Fully offloads H.264 and HEVC with NVENC
- High density of users per GPU
- Streaming Games and Enterprise Apps



STREAM APPLICATIONS

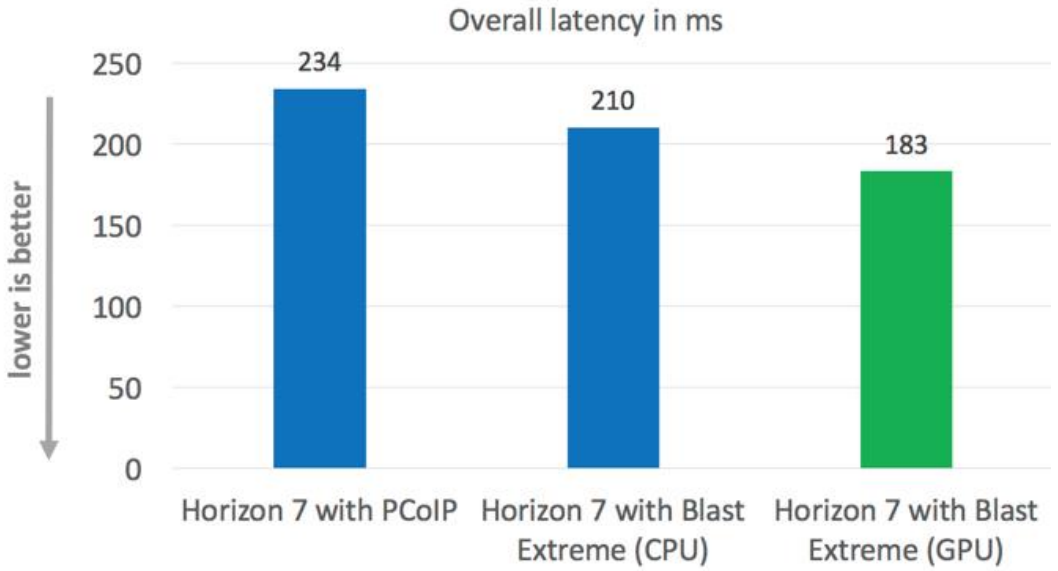
- Streaming software
  - VMware Horizon Blast Extreme
  - Nice Desktop Cloud Visualization
- Capture SDK + Encode SDK
  - Capture (NvFBC and NvIFR)
  - Encode with NvENC (H.264 and HEVC)
  - Supported in Virtualized environments
    - GPU direct attached mode
    - vGPU mode (shared GPU)

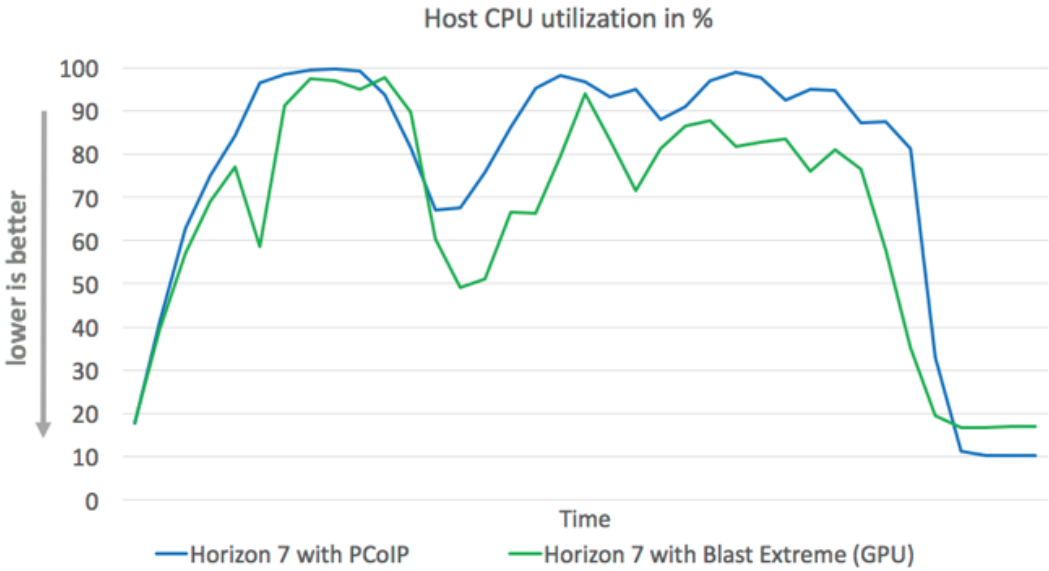


PERFORMANCE STUDY

vmware®

- VMWare Horizon Blast Extreme + GPU
  - 37% better performance (fps)
  - 21% lower latency
  - 19% reduction in bandwidth
  - 16% reduction in CPU utilization
  - 18% increase in number of users





LIVE VIDEO TRANSCODING

- **Higher number video streams per GPU server**
  - 1 stream to *N* streams (multi-resolution)
  - Fewer servers needed, higher density, lower TCO
  - Requires Lower bitrate (B-Frames)
- **Live Transcoding User Generated Content**
  - Live video broadcasts, presidential debates, concerts
  - Broadcasting from mobile device
  - Live game streaming events

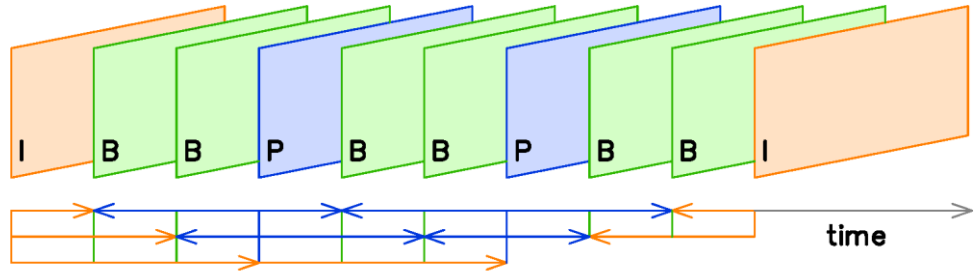
ArcSoft®

WOWZA™ media systems

telestream

TRANSCODE FOR ARCHIVING

- High density of streams per GPU servers
  - Lower TCO, lower latency
  - 1 stream to *N* streams (multi-resolution)
- Archiving
  - HQ archiving for non-live video streaming
  - Quality is and low bitrate are the most important (I, B, and P support)
  - Cost per stream



VIDEO CONFERENCING

- Live video conferencing
- Video transcoding (1 to *N* streams)
- Screen sharing for meetings
- Video enhancements
  - Video stabilization
  - Frame rate up sampling
- High quality, low bitrate

PERFORMANCE AND QUALITY TUNING

RECOMMENDED SETTINGS

Remote Graphics

- NVENC has video presets for latency (I and P frames only)

```
NV_HW_ENC_PRESET_LOW_LATENCY_HQ

NV_HW_ENC_PARAMS_RC_2_PASS_QUALITY
```

- Video Bitrate settings for low latency

```
dwVBVBufferSize  = dwAvgBitRate / (dwFrameRateNum/dwFrameRateDen)

dwVBVInitialDelay = dwVBVBufferSize
```

- Video Bitrate settings for higher quality

K = 4;

```
dwVBVBufferSize  = K * dwAvgBitRate / (dwFrameRateNum/dwFrameRateDen)

dwVBVInitialDelay = dwVBVBufferSize
```

RECOMMENDED SETTINGS

Video Transcoding

- NVENC settings for video quality (I, B, P frames)

```
NV_ENC_PRESET_HQ_GUID
NV_ENC_PARAMS_RC_2_PASS_QUALITY
set B frames > 0 (EncodeConfig::numB)
```

- Video Bitrate settings for low latency

```
dwVBVBufferSize  = dwAvgBitRate / (dwFrameRateNum/dwFrameRateDen)

dwVBVInitialDelay = dwVBVBufferSize
```

- Video Bitrate settings for higher quality

K = 4;

```
dwVBVBufferSize  = K * dwAvgBitRate / (dwFrameRateNum/dwFrameRateDen)
dwVBVInitialDelay = dwVBVBufferSize
```

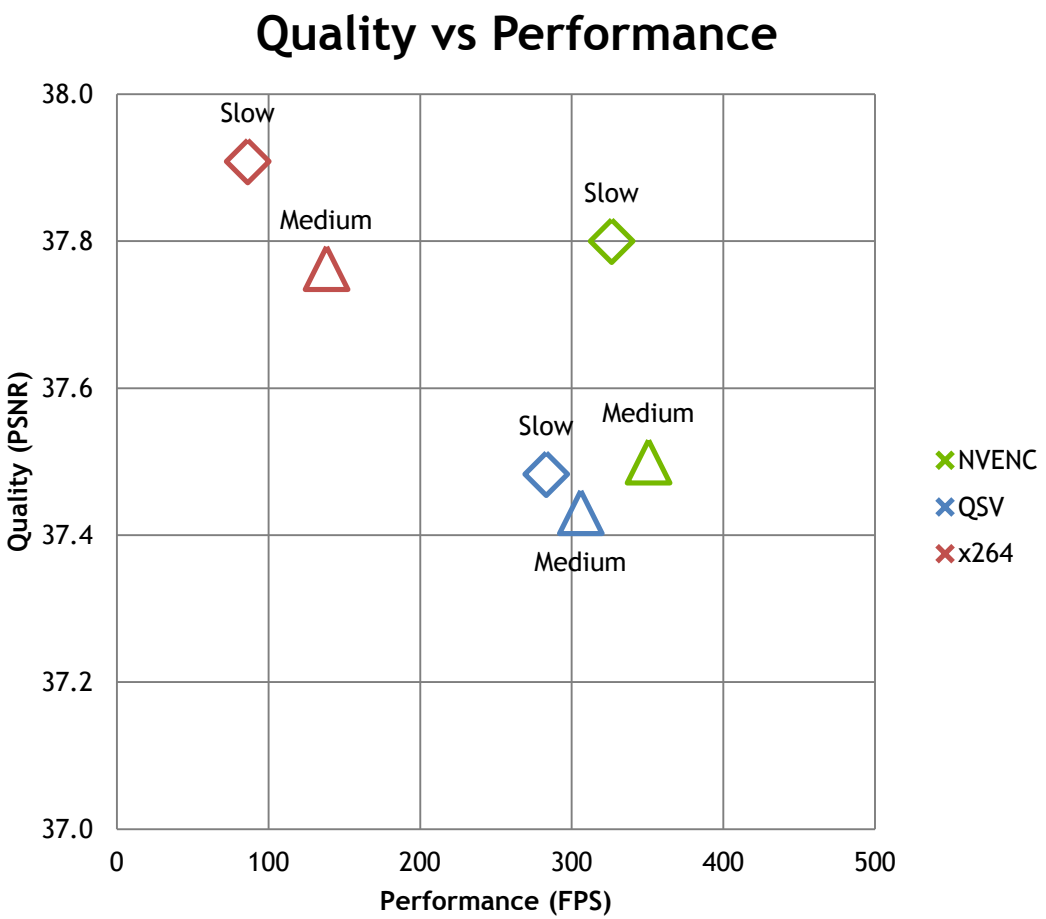
TESLA PERFORMANCE

	# NVDEC	# NVENC	# 1080P30 H.264 STREAMS*	# 1080P30 HEVC STREAMS*
Xeon E5 sw encode			2 (x264)	0.25-0.5 (x265)
Tesla M60 / 2xGM204	1+1	2+2	2 x (14+14) (870+870Mpixels/sec)	2 x (10+10) (622+622Mpixels/sec)
Tesla M6 / 1xGM204	1	2	14+14 (870+870Mpixels/sec)	10+10 (622+622Mpixels/sec)
Tesla M4 / 1xGM206	1	1	7 (435Mpixels/sec)	5 (311Mpixels/sec)

Each Maxwell NVENC can do:

- *7x h.264 1080p30 Highest Quality with B-frames*
- *5x HEVC 1080p30 Highest Quality with no B-frames*

ENCODE PERF/QUALITY



- Quality
  - = x264
- Performance
  - Single NVENC is 3-4x vs x264

NEW DIRECTIONS

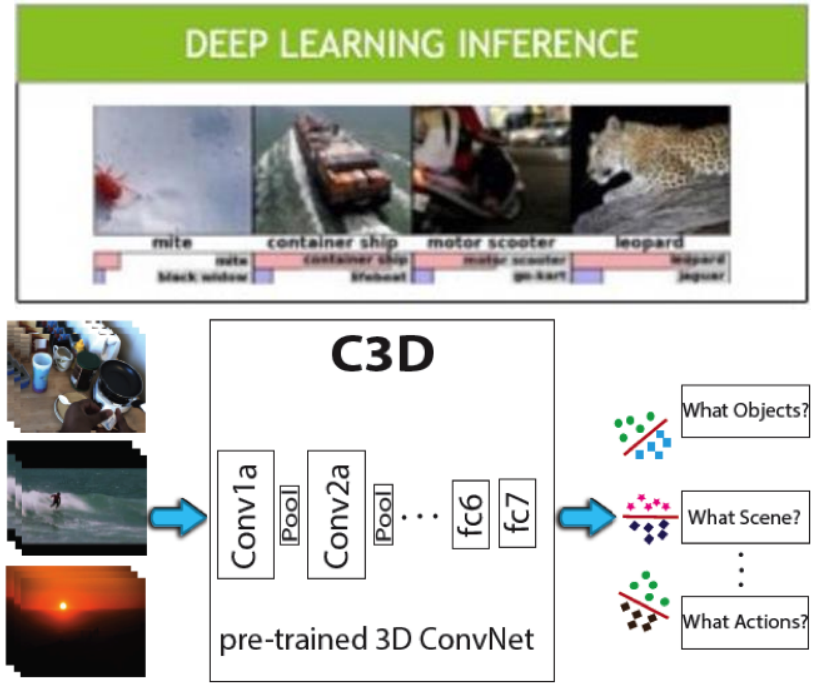
NEW USE CASES

- Standalone NVENC motion estimation mode
- Continued video quality improvements
  - Adaptive GOP, Adaptive B-frames, Adaptive Quantization
  - Temporal AQ
  - Frame look ahead
- Video Stabilization with compute
  - Use CUDA cores for image stabilization to remove video shakiness
  - Algorithm is well suited for GPU architectures
    - Takes advantage of texture cache
    - Scales on GPUs because of high level of parallelism\

DEEP LEARNING VIDEO INFERENCE

Using 3D ConvNet

- Video Analysis using pre-trained Convolution3D network (spatiotemporal signals)
- Use NVDEC to improve performance when running GPU inference
- https://research.facebook.com/blog/c3d-generic-features-for-video-analysis/



SDK LINKS

NVIDIA VIDEO CODEC SDK

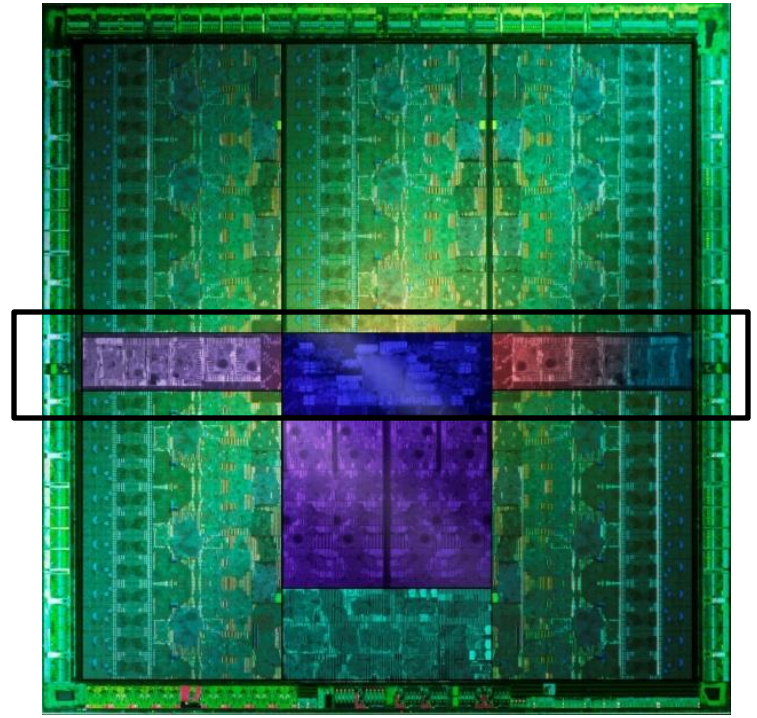
Since Kepler dGPU have had Fixed-Function Decoder and Encoder blocks

NVENC – NVIDIA Video Encoder

NVDEC – NVIDIA Video Decoder

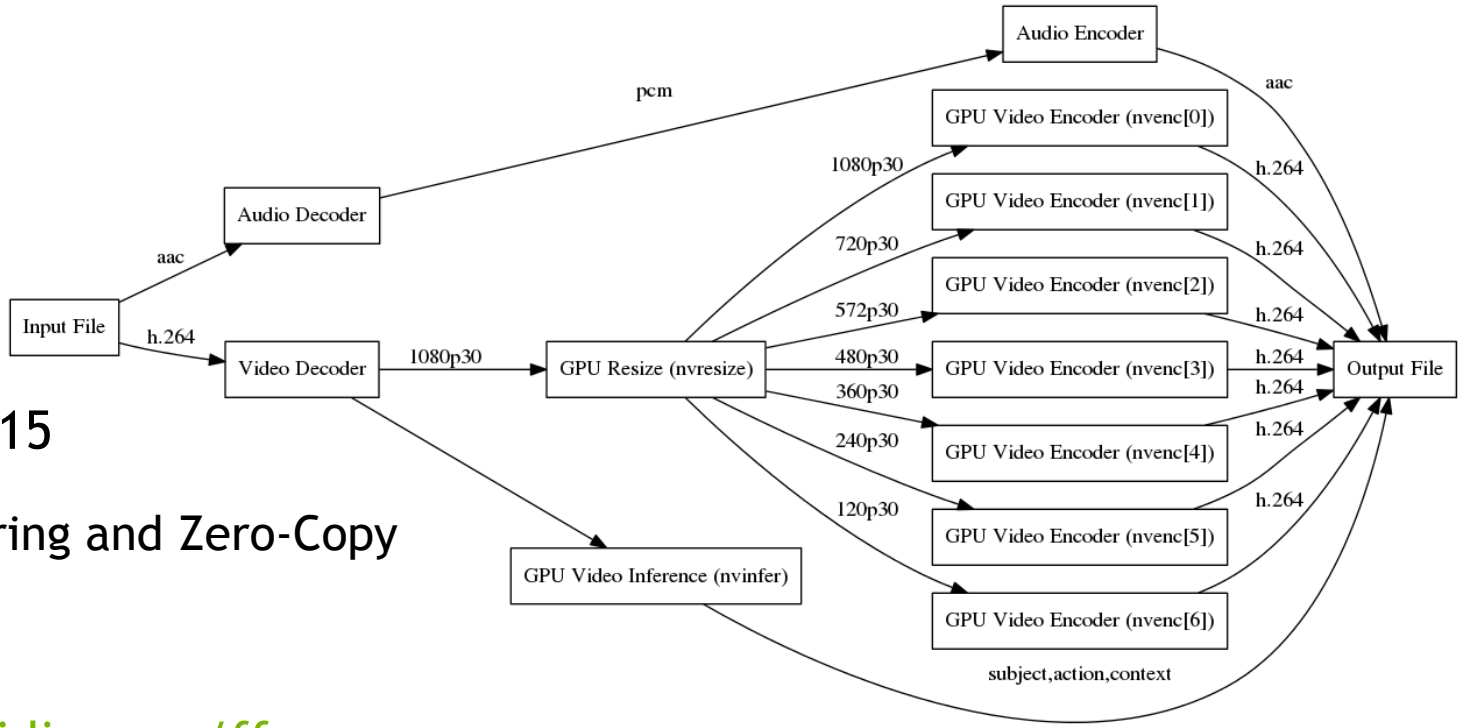
Samples and documentation

https://developer.nvidia.com/nvidia-video-codec-sdk



GM200

FFMPEG + NVENC



- NVENC added 1/2015
- NVRESIZE added 8/2015
  - CUDA Context sharing and Zero-Copy
- NVDEC added 1/2016
- https://developer.nvidia.com/ffmpeg

QUESTIONS?

Find us at GTC Hangouts

GTC Pod B - H6145A: Video and Image Processing
4/5 (Tuesday) @ 12:45 – 2pm

GTC Pod A - H6145B: Video and Image Processing
4/6 (Wednesday) @ 8:45am - 10am

Abhijit Patait apatait@nvidia.com
Eric Young eyoung@nvidia.com