HIGH PERFORMANCE VIDEO ENCODING WITH NVIDIA GPUS

Abhijit Patait
Eric Young

April 4th, 2016
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**
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

<table>
<thead>
<tr>
<th>KEPLER (GK107, GK104)</th>
<th>MAXWELL GEN 1 (GM107)</th>
<th>MAXWELL GEN 2 (GM200, GM204, GM206)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H.264 only</td>
<td>H.264 only</td>
<td>H.264 and HEVC/H.265</td>
</tr>
<tr>
<td>~240 fps 2-pass encoding @ 720p</td>
<td>~500 fps 2-pass encoding @ 720p</td>
<td>~900 fps 2-pass encoding @ 720p</td>
</tr>
<tr>
<td>GRID K340/K520, K1/K2, Quadro K5000, Tesla K10/K20, GeForce GTX 680</td>
<td>Maxwell-based GRID &amp; Quadro products</td>
<td>Tesla M4, M40, M6, M60, Quadro M4000, M5000, M6000, GeForce GTX 960, 980, Titan X</td>
</tr>
<tr>
<td>NV Encode SDK 1.0-5.0</td>
<td>NV Encode SDK 4.0+</td>
<td>NV Encode SDK 5.0 Video Codec SDK 6.0+</td>
</tr>
</tbody>
</table>
## DECODE CAPABILITIES

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

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

### What’s New

<table>
<thead>
<tr>
<th>Feature</th>
<th>SDK release</th>
<th>Why</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video SDK = encode + decode</td>
<td>6.0</td>
<td>Transcoding</td>
</tr>
<tr>
<td>Quality++</td>
<td>6.0</td>
<td>Streaming, Transcoding, Broadcast, Video production</td>
</tr>
<tr>
<td>RGB inputs</td>
<td>6.0</td>
<td>Capture RGB + encode</td>
</tr>
<tr>
<td>Motion estimation only mode</td>
<td>6.0</td>
<td>Hardware assisted motion estimation for custom encoders, Image stabilization</td>
</tr>
<tr>
<td>Adaptive quantization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaptive B-frames</td>
<td>7.0</td>
<td>Improved perceptual quality - Available in May 2016</td>
</tr>
<tr>
<td>Adaptive GOP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Look-ahead</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# ROADMAP

<table>
<thead>
<tr>
<th>Q2’15</th>
<th>Q3’15</th>
<th>Q4’15</th>
<th>Q1’16</th>
<th>Q2’16</th>
<th>Q3’16</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NVENC SDK 5.0</strong></td>
<td>Video SDK 6.0</td>
<td>Future...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• HEVC</td>
<td>• ME-only (H.264)</td>
<td>• Quality++</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Maxwell Gen 2</td>
<td>• HEVC Quality+</td>
<td>• HEVC 10-bit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• H.264 4:4:4</td>
<td>• RGB inputs</td>
<td>• HEVC 4:4:4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>• H.264 lossless</td>
<td>• HEVC AQ</td>
<td>• HEVC lossless</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**GM204, GM206**

Maxwell Gen 2

**Pascal**

Future...

• Quality++
• HEVC 10-bit
• HEVC 4:4:4
• HEVC lossless
• ME-only (HEVC)
• 4K HEVC 60 fps
• 8K HEVC
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

---

**Simplified Diagram:**

- Apps
- Remote Graphics Stack
- Network
- H.264 or raw streams
- NVENC
- NVIFR
- NvFBC
- Render Target
- Front Buffer
- Framebuffer
- 3D Graphics commands
- Tesla, GRID, or Quadro GPU

---

**Technology:**

- H.264
- HEVC
- GEFORCE NOW
- NVIDIA NVENC
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 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
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)

\[ \text{NV_HW_ENC_PRESET_LOW_LATENCY_HQ} \]
\[ \text{NV_HW_ENC_PARAMS_RC_2_PASS_QUALITY} \]

• Video Bitrate settings for low latency

\[ \text{dwVBVBufferSize} = \frac{\text{dwAvgBitRate}}{(\text{dwFrameRateNum} / \text{dwFrameRateDen})} \]
\[ \text{dwVBVInitialDelay} = \text{dwVBVBufferSize} \]

• Video Bitrate settings for higher quality

\[ K = 4; \]
\[ \text{dwVBVBufferSize} = K \times \frac{\text{dwAvgBitRate}}{(\text{dwFrameRateNum} / \text{dwFrameRateDen})} \]
\[ \text{dwVBVInitialDelay} = \text{dwVBVBufferSize} \]
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

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

*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 INFEERENCE

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/](https://research.facebook.com/blog/c3d-generic-features-for-video-analysis/)
SDK LINKS
Since Kepler dGPU have had Fixed-Function Decoder and Encoder blocks

NVENC - NVIDIA Video Encoder

NVDEC - NVIDIA Video Decoder

Samples and documentation

FFMPEG + NVENC

- NVENC added 1/2015
- NVRESIZE added 8/2015
  - CUDA Context sharing and Zero-Copy
- NVDEC added 1/2016
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