New High-speed Professional Video Compression using CUDA

Jan Weigner, CTO
jan@cinegy.com
Cinegy GmbH
Executive Summary

This presentation is about Cinegy‘s DANIEL2 GPU image and video codec which was developed specifically for maximum performance using NVIDIA‘s CUDA GPU technology. DANIEL2 provides massive performance improvements for professional image and video processing applications over existing CPU-based approaches. DANIEL2 is a game changer for professional high-resolution image and video processing with a wide range of applications.
Why Yet Another Codec?

Speed, speed and speed.

Other benefits were welcome side effects, but maximum performance was the key goal in designing a professional video encoder/decoder (codec) specifically for use NVIDIA GPUs.

A GPU-based codec is inevitable.
This audience, if any, should know why.
Target markets

- Film & Broadcast
- GIS
- Medical
- Defense
- Gaming
- Large scale video walls
- Visualization
- VR & AR
- Professional Photography
- Video over IP / KVMoIP
- ... many more
Driving Factors

Resolution: SD → HD → UHD → 8K → 16K ?

Higher Frame Rates: 30 fps → 60 fps → 120 fps

Dynamic Range: SDR → HDR

Precision: 8 bit → 10 bit → 12 bit → 16 bit
Driving Factors in Numbers

Resolution
- SD 5x
- HD 4x
- UHD 4x
- 8K 4x
- 16K ?

Higher Frame Rates
- 30 fps 2x
- 60 fps 2x
- 120 fps

Dynamic Range
- SDR
- HDR Move from 8 to 10 bit plus log profile

Precision
- 8 bit +25%
- 10 bit +20%
- 12 bit +33%
- 16 bit
### Driving Factors in Numbers

<table>
<thead>
<tr>
<th>Format</th>
<th>Bandwidth</th>
<th>Frame Rate</th>
<th>Technology</th>
<th>Bit Depth</th>
</tr>
</thead>
<tbody>
<tr>
<td>SD</td>
<td>270M</td>
<td>30fps</td>
<td>SDR</td>
<td>8 bit</td>
</tr>
<tr>
<td>HD</td>
<td>3G</td>
<td>60 fps</td>
<td>SDR</td>
<td>8 bit</td>
</tr>
<tr>
<td>UHD</td>
<td>12G</td>
<td>60 fps</td>
<td>HDR</td>
<td>10 bit</td>
</tr>
<tr>
<td>8K</td>
<td>48G</td>
<td>60 fps</td>
<td>HDR</td>
<td>10 bit</td>
</tr>
</tbody>
</table>
Driving Force: Tokyo 2020

- NHK will broadcast the 2020 Olympics in 8K. Test broadcasts started in 2016. Plans are to rollout full 8K service by 2018. NHK ultimate goal is 8K @ 120 fps
- SHARP 8K 75" TVs go on sale this month in China ~ $8000
- Dell’s 8K 32" monitor is out since March ~ $3899
- RED’s first Weapon 8K digital cinema camera is soon two years old, then came Helium, now MONSTRO – 3rd gen ~ from $79500
- Sony Alpha 7R II DSLR has a 42M pixel sensor ~ $2999
But 8K is just another step on the way ...

Lytro Cinema camera
755 RAW Megapixels
Up to 300 fps

Canon CCD
250 Megapixel sensor
BUT

... with current codecs and PC hardware there are a number of bottlenecks that make going beyond 4K problematic. At least if the goal is to do it with COTS PC hardware.
The Bottlenecks

Bandwidth
- Storage speed
- RAM speed
- PCIe bus

Compute
- # of compute cores
Bandwidth Bottlenecks

• HDD performance and network I/O used to be a considerable bottleneck. With PCIe SSDs and 40GB Ethernet the PCIe bus is the bigger obstacle.

• Using compression reduces the bandwidth required and allows scaling the number of streams that can be handled.

• The CPU RAM speed is improving slowly but even with the latest Intel / AMD CPUs is miles away from high-end GPUs.

• The massive CPU L2/L3 caches help reducing the pain to some extend.
The Evil PCIe Bus

What was once the least problem in terms of system performance has become the main bottleneck. We will still have to deal with PCIe 3.0 for at least two years before PCIe 4.0 will start to ripple through the PC eco system (CPUs, chipsets, motherboards, graphics cards, I/O cards etc.). By the time PCIe 4.0 materializes 8K will be common place and we will pray for the arrival of PCIe 5.0.
The PC System Bottlenecks

- **Core i9 XXXX Processor**
  - 44x PCIe lanes
  - 4 channel DDR4 RAM (~90GB/s)
- **PCIe SSD**
- **40G NIC**
- **GB NIC**
- **PCIe x16** NVIDIA GPU (~12GB/s)
- **PCIe x16** RAM (~500GB/s)
- **X299 Chipset**
- **DMI**
- **USB**
- **SATA**
The PC System Bottlenecks

- PCIe SSD
- 40G NIC
- Core i9 XXXX Processor
- RAM (4 chan DDR4 ~90GB/s)
- PCIe x16 (NVIDIA GPU ~12GB/s)
- PCIe x16 (~500GB/s)
- X299 Chipset
- GB NIC
- USB
- SATA
The PCIe Bottleneck

The PCIe 3.0 bus has a theoretical limit of around 32GB/s bi-directional ~ 16GB/s read or write. In reality much less - 10-12GB/s when pushing it.

This shows that uncompressed 8K with above parameters is likely to fail due to PCIe bus saturation when trying to push more than one stream. In case of 120 fps even one stream will be too much to handle on most machines.

Only when staying with 4:2:2 @ 60fps or 4:4:4 @ 30fps or less fps, is uncompressed playback of a single stream guaranteed.
Overcoming the PCIe Bottleneck

• There is only one way to overcome the PCIe bus bottleneck: stay in the compressed domain wherever and as long as you can.

• For those with quality concerns: use visually lossless or mathematically lossless compression modes.
CPU Bottleneck

• CPU performance as such is not a bottleneck, leaving costs and power consumption aspects aside. New AMD and Intel processors offer more processor cores than ever – for a price – but in terms of processing power they offer far less „bang per buck“ than GPUs. AVX2 optimization has helped our codecs more than anything else in the last years. Whether AVX512 is going to help equally much is yet to be seen.

• Production codecs such as Apple ProRes and AVID DNxHR can decode 8K streams even at 60fps in realtime given powerful enough CPUs.

• BUT this creates a high processor load and the PCIe bus bottleneck to the GPU remains. If the uncompressed image data still has to go to the GPU for display or further processing this creates needless traffic.
CPU Bottleneck

• The result is always the same – when wanting to decode more than one single stream of 8K (10bit @ 60fps) and display it, this is a challenge.
• If the codec in question then also uses 16bit writes to transfer color values of 10bit or higher into the GPU or video framebuffer, then even a single stream @ 60fps is a challenge.
• In any case CPU based codecs create or deal with the image data on the wrong side of the bus if this needs to be displayed or further processed using the GPU.
The almost exponential NVIDIA GPU performance growth already for years outperforms the x86 CPU speed gains. “Moore’s Law is Dead.”
GPU to the Rescue

- The PCIe bus and CPU bottleneck need to be circumvented.
- The video data must stay in the compressed domain going into the GPU for decoding there directly. -> The need for a pure GPU codec.
- The GPU must decode into the GPU memory for direct display or further processing inside the GPU.
- Distribution encoding for delivery also ideally happens inside the GPU. -> handover to NVENC
- The CPU is freed to do other tasks or can be smaller.
- This means less power consumption, less costs and higher speed.
Enter the Cinegy Daniel2 GPU Codec

- The Daniel2 is the logical evolution of the CPU-based Daniel1 codec.
- Sharing only the name with its predecessor, the design of Daniel2 is totally GPU oriented and not following standard design pattern such as JPEG, MJPEG, JPEG2000, H.263, H.264 etc.
- The Daniel2 design is radically different and architected to scale across all available GPU cores and use the abundant GPU RAM bandwidth.
- The design approach of Daniel2 pragmatically makes the most of the GPU's abilities and is not an academic, theoretical exercise.
- It is based on many years of deep understanding of the inner workings of the GPU architecture and applying this to the codec design.
Cinegy DANIEL2 - Positioning

DANIEL2 is aiming for the same markets as:

Apple ProRes  AVID DNxHD  SONY XAVC

CineForm  OpenEXR  TIFF
Cinegy Daniel2 GPU Codec Specs

- From 4:2:2 to 4:4:4:4 - YUV to RGBA
- 8 bit, 10 bit, 12 bit and 16 bit per component
- No resolution limitation other than RAM
- Intelligent alpha channel support
- Extremely low latency
- Region of Interest decoding
- Multi-generation re-compression
- Freely selectable compression ratio
- Adaptable VBR, CBR or CQ
- Lossy or lossless encoding
- Decode pipeline integrated scaler
- Ultra fast Nvidia GPU (CUDA) codec
- Multi GPU support
- Very fast CPU codec (e.g. for VMs)
- High-quality IP streaming via RTP
- 3D LUT based realtime color correction
- Integrated realtime effects pipeline
- MXF OP1A wrapper for edit while write
- Free Cinegy Player with DANIEL2 support
- Free Adobe CC import & export plugin
- Cinecoder Developer SDK
- Windows now, Linux and Mac soon
The quality is similar to Apple ProRes and AVID DNxHR while for now producing slightly bigger files.
Decoding Performance

8K 4:2:2
Decode to null:
Fastest mode

The DANIEL2 decoding performance allows to process multiple 8K streams in parallel and to perform additional processing in parallel.
Encoding Performance

With the GTX1070 or P4000 and upwards it shows that we start hitting the PCIe bus bandwidth limits and start to flatline.

8K 4:2:2
Official Cinescore results

![Graph showing Encoding Performance](image)
DEMO TIME

Part 1
Cinegy Player 3.0

- Windows 10 style player for Daniel2 video files (and other formats)
- Requires Nvidia Maxwell / Pascal GPU and Windows 64bit OS
- Native 8K output, or scaled output to 4K, full HD or smaller displays.
- Zoom-in, zoom-out, scan & pan while playing or while paused.
- JKL controls, single stepping, scrubbing.
- Realtime 3D LUT color correction and image effects.
- 16 or 24 bit audio playback.
- Detailed technical info and status display.
- Full screen output with 10 bit support or windowed playback.
- Multiple Cinegy Player can in parallel.
- Free download at www.daniel2.com
If time permits ... If not, come to our booth for a private demo - SL11116

FPS = 60.02
quant = 0 bit_depth = 0 chroma_format = 4 width = 7680 height = 4320
zoom = 16.88%
frames = 30689 cur_frame_num = 487
vertical sync - on
Mute - off
color - 0 | poster - 0 | BnW - 0 | neg - 0 | gamma - 0 | flip - 0
Current BitRate - 2189 Mbits/sec
If time permits ...
If not, come to our booth for a private demo.
DEMO TIME

Part 2
Cinegy DANIEL2 Adobe CC Plugin

• Free plugin for Adobe CC
• Import, edit, export DANIEL2 MXF files
• Integrates with Premiere, AfterEffects and Media Encoder
• Support for all modes and color spaces.
• Full Alpha Channel support
• 8K editing on a notebook
• Currently Windows only
Summary

• Going to 8K and beyond is possible today on inexpensive, commodity hardware including notebooks using Nvidia GPUs.

• A complete GPU centric redesign of codecs and effects / rendering pipelines is necessary to achieve this.

• The PCIe bus is and will remain the primary bottleneck requiring to stay in the compressed domain on the CPU side of the PCIe bus.

• The CPU becomes a I/O pump and is freed to perform other tasks.

• Handling 16K @ 60fps and more in realtime is possible today using high-end Nvidia GPUs.
Download it now:

WORLD'S FASTEST VIDEO CODEC

www.daniel2.com