

NVIDIA VIDEO TECHNOLOGIES

Abhijit Patait, 3/26/2018



AGENDA

NVIDIA Video Technologies Overview
Video Codec SDK Updates
Perf/Quality Optimization
Benchmarks
Roadmap

NVIDIA VIDEO TECHNOLOGIES

VIDEO CODEC SDK

A comprehensive set of APIs for GPU-accelerated video encode and decode

NVENCODE API for video encode acceleration

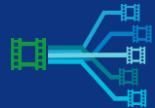
NVDECODE API for video & JPEG decode acceleration (formerly called NVCUVID API)

Independent of CUDA/3D cores on GPU for pre-/post-processing

Gamestream



Cloud transcoding



Remote desktop & visualization



Intelligent video analytics



Video archiving



Video editing



NVIDIA VIDEO TECHNOLOGIES

SOFTWARE



Easy access to GPU
video acceleration

DeepStream SDK

cuDNN, TensorRT,
cuBLAS, cuSPARSE

VIDEO CODEC SDK

Video Encode and Decode for Windows and Linux
CUDA, DirectX, OpenGL interoperability

CUDA TOOLKIT

APIs, libraries, tools, samples

HARDWARE

NVENC

Video encode



NVDEC

Video decode

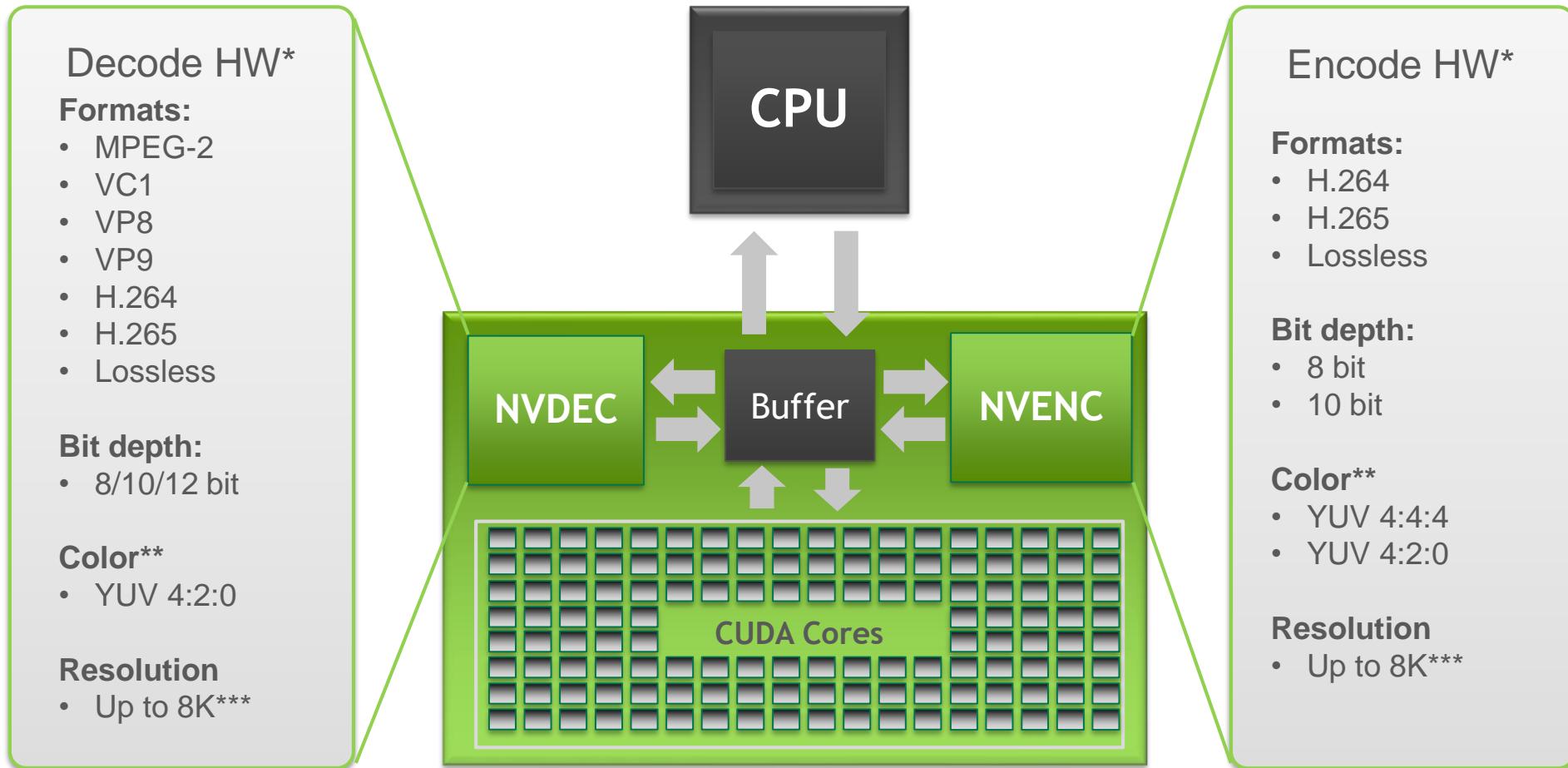


CUDA

High-performance
computing on GPU



NVIDIA GPU VIDEO CAPABILITIES



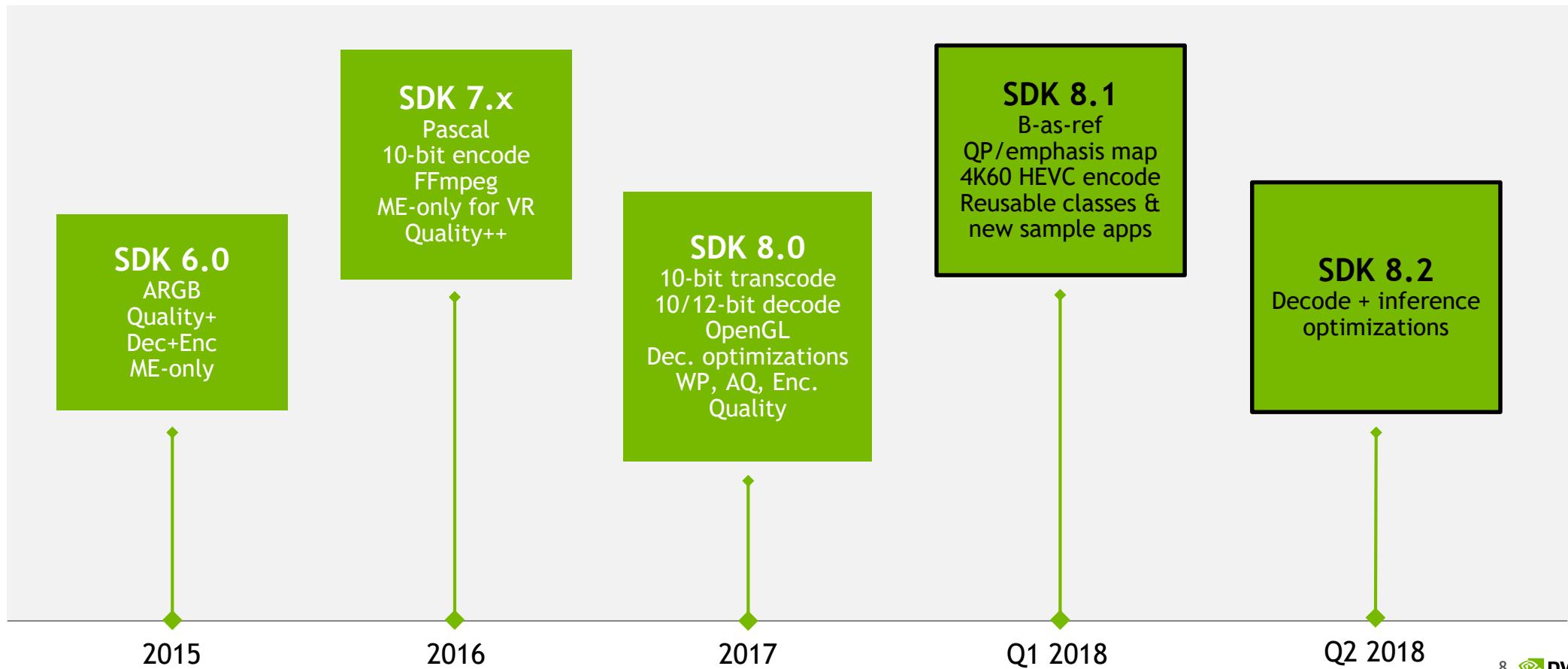
* See support diagram for previous NVIDIA HW generations

** 4:2:2 is not natively supported on HW

*** Support is codec dependent

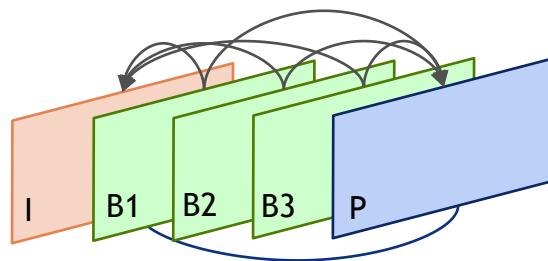
VIDEO CODEC SDK UPDATE

VIDEO CODEC SDK UPDATE

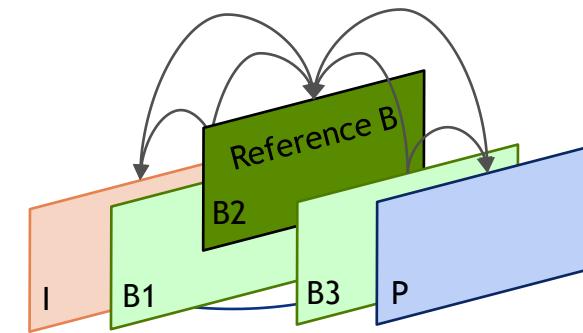


B-FRAMES AS REFERENCE

Non-ref B-frames



B-frames as reference



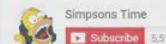
- Improved visual quality - up to 0.6 dB PSNR (BD-PSNR = 0.3 dB)
- Negligible performance penalty
- Ensure decoder support



the simpsons



The Simpsons Funniest Moments #27*HD*(Bart in Future)



Simpsons Time

Subscribe

5,511

[+ Add to](#) [Share](#) [More](#)

727,285 views

[Like](#) 4,552 [Dislike](#) 541

Published on Sep 17, 2016

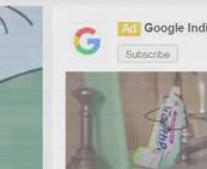
Category People & Blogs
License Standard YouTube License

COMMENTS • 300



Add a public comment...

Top comments ▾

Tabbedder 18 hours ago
The Simpsons Funny MomentsReply • 5 [Like](#)Jacklyn Garcia 9 hours ago
Saloure a good day please

Visit Advertiser's Site

Up next

Autoplay The Simpsons Funniest
Moments #28*HD*(Beer Made)
Simpsons Time
100,382 views • NEW
4:52The Simpsons Funniest
Moments #26*HD*(Kill The)
Simpsons Time
42,893 views • NEW
4:24The Simpsons Funniest
Moments #29*HD*(Crazy Smile)
Simpsons Time
90,375 views • NEW
5:50The Simpsons funniest moments
Season 27 episode 2
Nam Le
8,978 views • NEW
26:45The Simpsons funniest moments
Season 27 episode 1
Nam Le
18,340 views • NEW
24:51The Simpsons Funniest
Moments #30*HD*(Soap In The)
Simpsons Time
14,918 views • NEW
4:21The Simpsons Funniest
Moments Season 27 Part 2
melisa
6,595 views
26:13Infinity War - Avengers vs
Thanos Empire
Recommended for you
23:25Spiderman repairs Disney car
cartoon for childrens with
CAR FUN FUN
Recommended for you • NEW
10:22

WITHOUT B-AS-REF

1080p @3 Mbps

WITH B-AS-REF
1080p @3 Mbps

The Simpsons Funniest Moments #27*HD*(Bart in Future)

Simpsons Time [Subscribe](#) 5,511

Published on Sep 17, 2016

Category People & Blogs
License Standard YouTube License

Comments • 300

Add a public comment...

Top comments

Tabbender 18 hours ago
The Simpsons Funny Moments

Reply • 5

Jacklyn Garcia 9 hours ago
Salours a good day please

727,285 views

1,452 541

Ad Google India

Subscribe

Up next Autoplay

The Simpsons Funniest Moments #28*HD*(Beer Made)
Simpsons Time 100,382 views NEW 4:52

The Simpsons Funniest Moments #26*HD*(Kill The)
Simpsons Time 42,893 views NEW 4:24

The Simpsons Funniest Moments #29*HD*(Crazy Smile)
Simpsons Time 90,375 views NEW 5:50

The Simpsons funniest moments Season 27 episode 2
Nam Le 8,978 views NEW 26:45

The Simpsons funniest moments Season 27 episode 1
Nam Le 18,340 views NEW 24:51

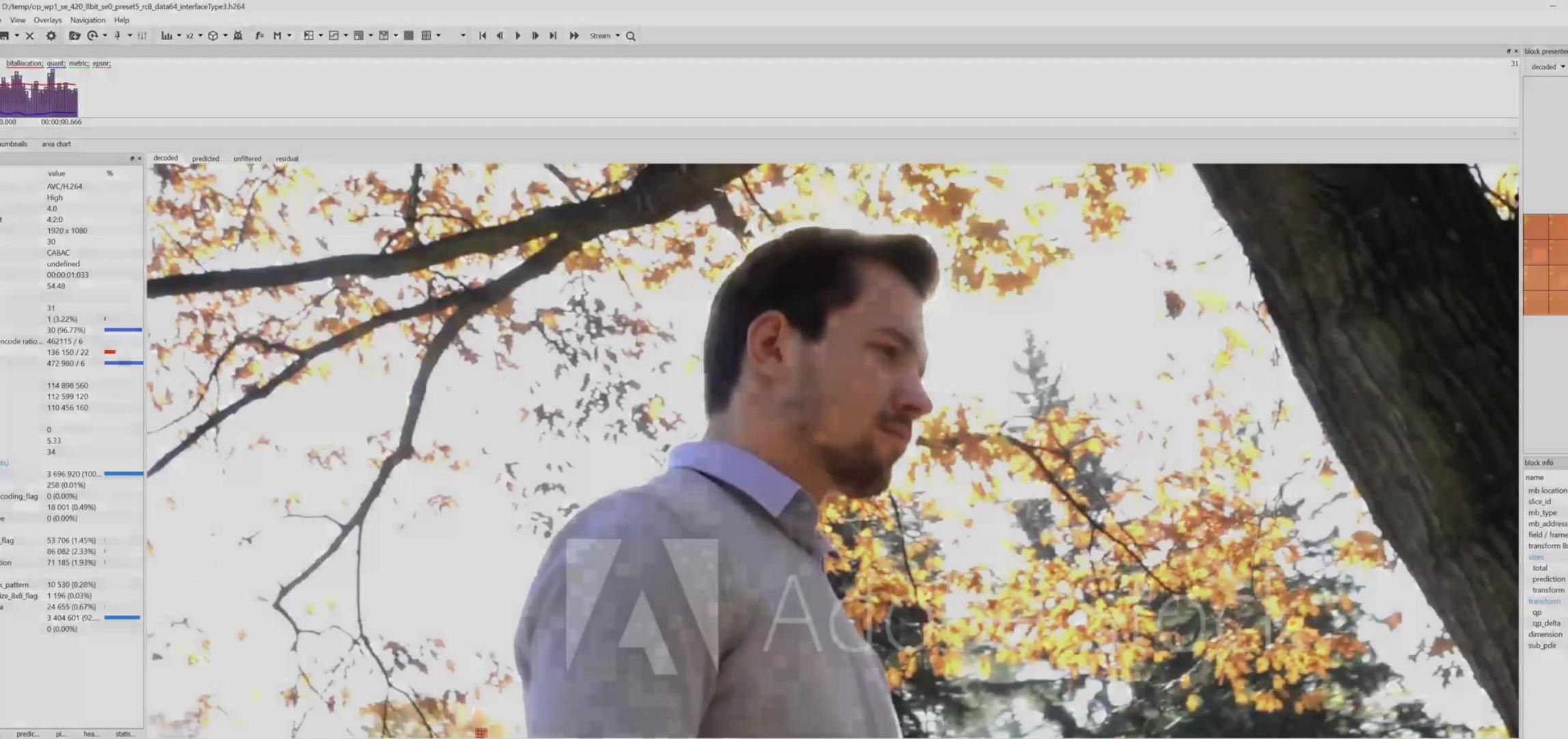
The Simpsons Funniest Moments #30*HD*(Soap In The)
Simpsons Time 14,918 views NEW 4:21

The Simpsons Funniest Moments Season 27 Part 2
melissa 6,595 views 26:13

Infinity War - Avengers vs Thanos
Thanos Empire Recommended for you 23:25

Spiderman repairs Disney car cartoon for childrens with
CAR FUN FUN Recommended for you NEW 10:22

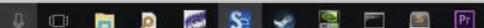
The Simpsons Funniest



WITHOUT B-AS-REF

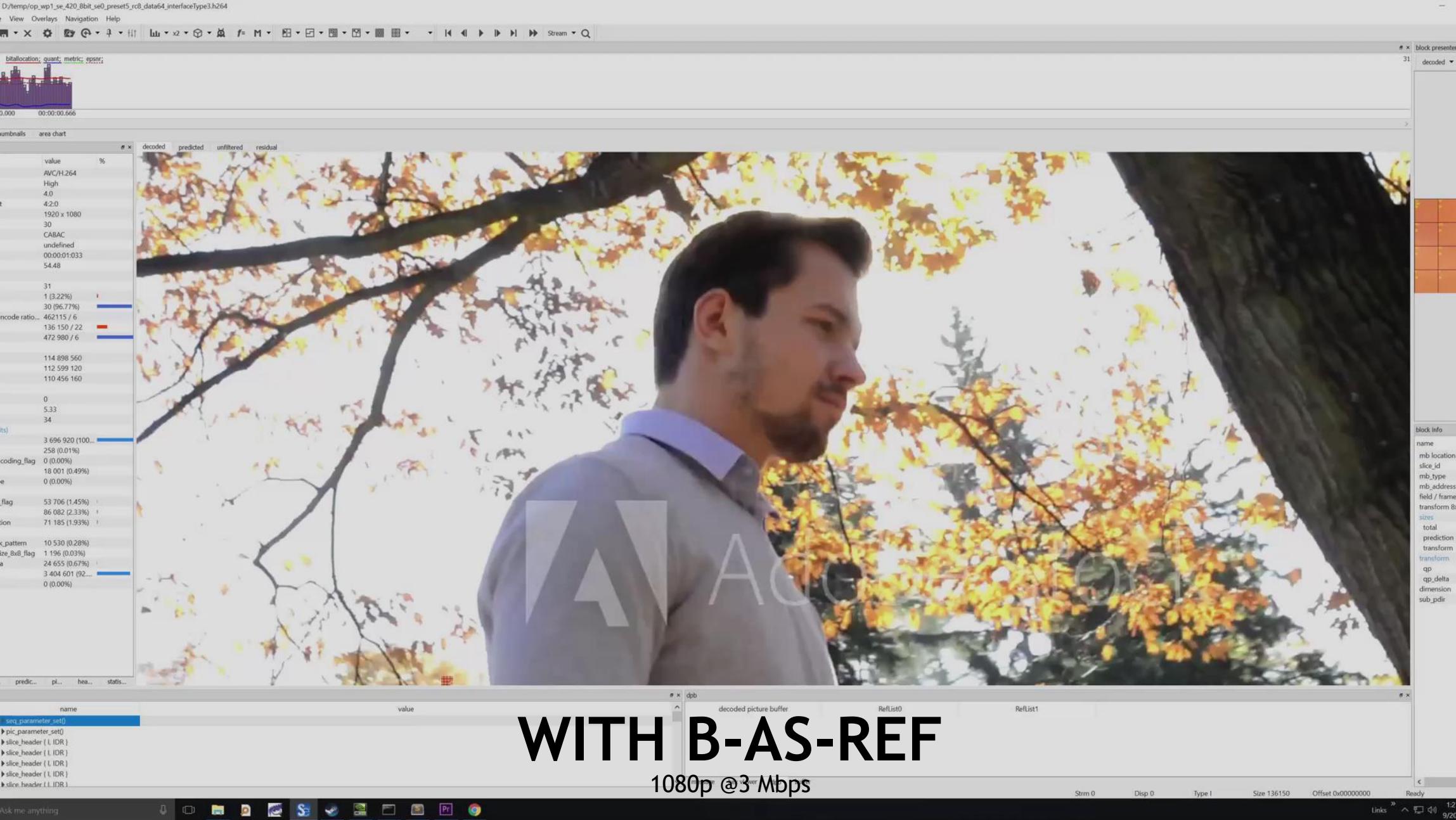
1080p @3 Mbps

Ask me anything



Strm 0 Disp 0 Type I Size 136150 Offset 0x00000000 Ready

Links > 1/27 9/20

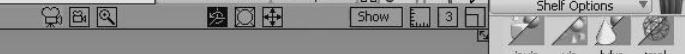


DESKTOP CONTENT ENCODING

Challenges in Preserving Details

Problem

- Desktop content is challenging to encode
- Thin-line text, wireframes, high-detail textures
- If severely bitrate constrained, recovery is difficult without IDR.
- QP modulation requires knowledge of complexity
 - Rate control in NVENC firmware

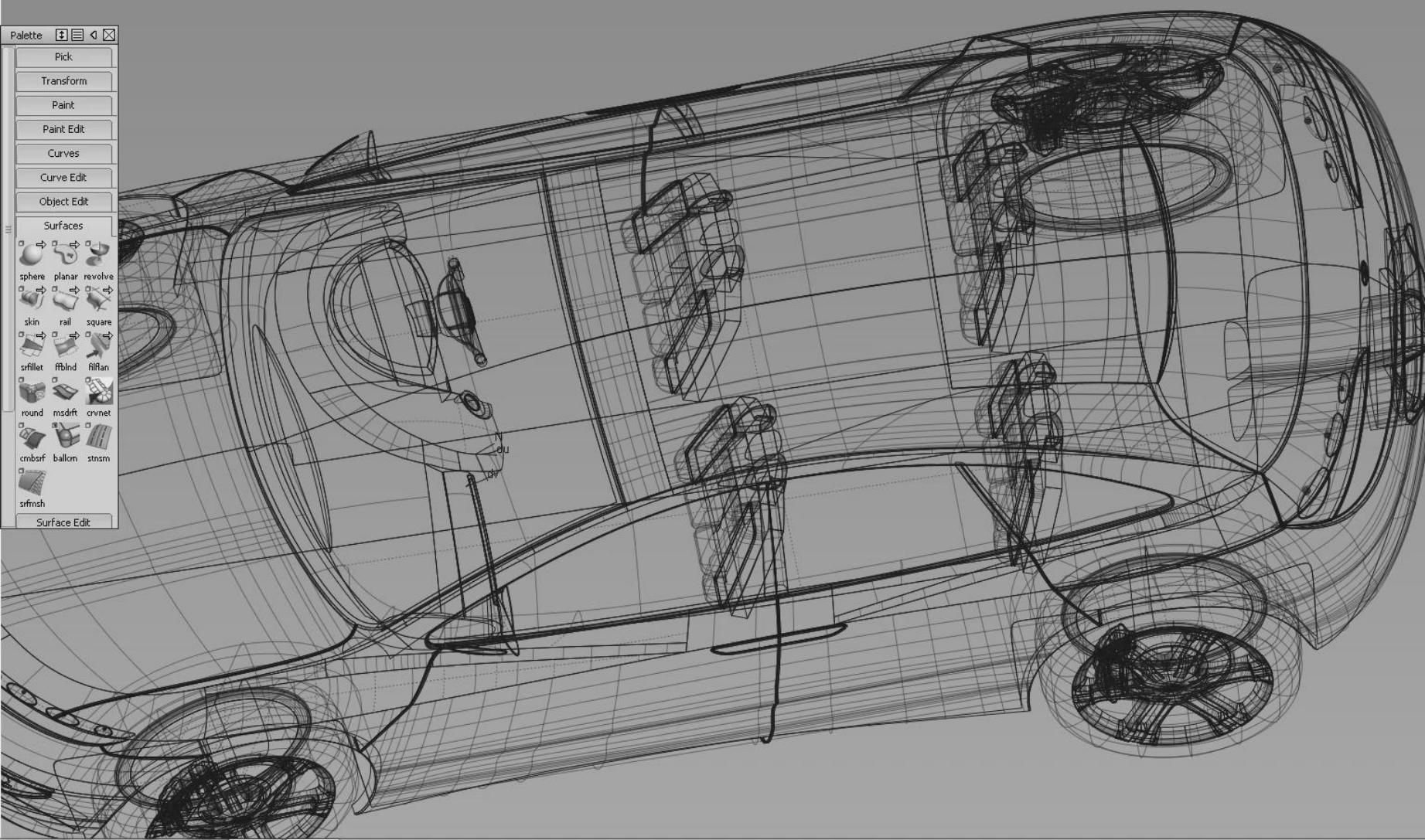


Original Image

Palette

- Pick
- Transform
- Paint
- Paint Edit
- Curves
- Curve Edit
- Object Edit
- Surfaces**
- sphere planar revolve
- skin rail square
- sfillet ffilan filfan
- round msdrift crvnet
- cmbmf balloon strmf
- sifmsh

Surface Edit



Degree Spans

Display

Deviation

Cv/Hull

Edit Points

Blend Points

Isoparm U V

Curvature U V

Transparency

Quality



Diagnostic Shade

shdnon mulcol rancol curevl

isong horver surevl usetex

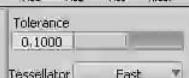
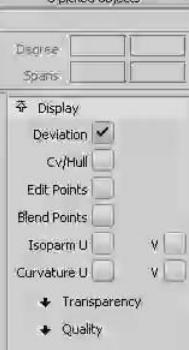
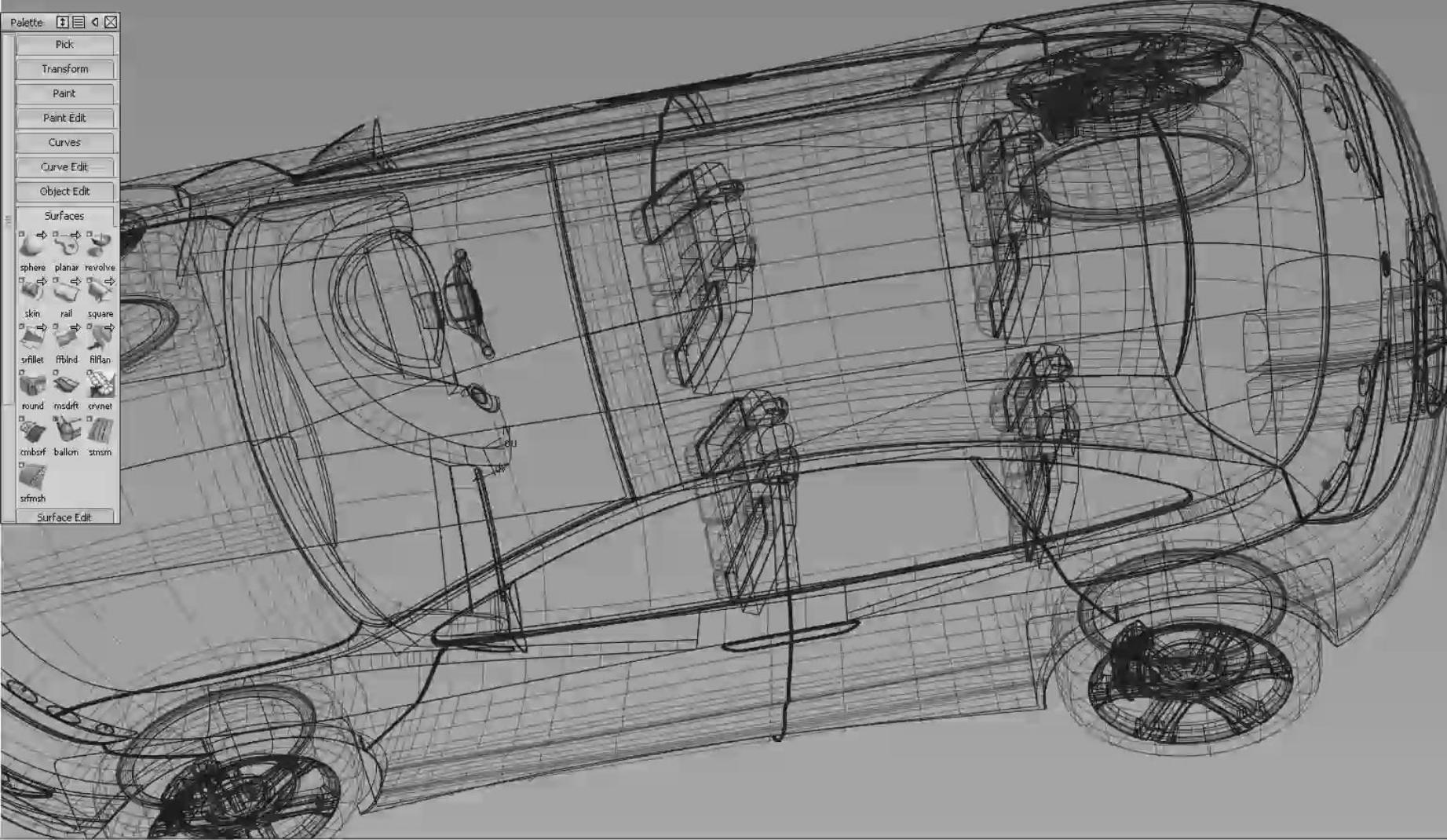
ltunnel clayao sophotes vred

vis1 vis2 vis3 filest

Tolerance 0.1000

Tessellator Fast

Encoded (& Decoded) Image



EMPHASIS MAP

Region of Interest Encoding

Solution

- Identify “high-detail” areas within the captured image (NVFBC)
- Provide feedback to encoder to treat these areas differently (NVENC)

EMPHASIS MAP

Region of Interest Encoding

Generated by NVFBC

5	5	4	5	3	2	1	0
5	5	5	3	3	2	2	0
5	5	4	4	2	1	0	0
3	2	4	3	2	1	1	2
1	1	0	3	2	4	0	0

Interpreted by NVENC as ΔQP

---	---	---	---	---	---	-	-
---	---	---	---	---	---	-	-
---	---	---	---	---	-	-	-
--	-	---	---	-	-	-	-
-	-	-	---	---	-	---	-

5 = High detail areas

0 = Low detail areas

Encoder translates to ΔQP

ΔQP depends on absolute QP

REDESIGNED SDK SAMPLES

Reusable Encoder/Decoder Classes

- Reusable base classes, easy-to-understand, end-user focused
- Sample apps re-designed
- **Encode base classes:** NvEncoderD3D9, NvEncoderD3D11, NvEncoderCUDA, NvEncoderD3GL
- **Decode base class:** NvDecoder
- Abstraction over low-level enc/dec APIs
 - `init()`, `run()`, `destroy()`
- FFmpeg demux

REDESIGNED SDK SAMPLES

Decode Applications

AppDec	Basic Decoding	AppDecLowLatency	Low-latency decode
AppDecD3D	Decode and Display using D3D9 and D3D11	AppDecMem	Decode from memory buffer
AppDecGL	Decode and Display using OpenGL	AppDecMultiInput	Use-case: Surveillance, multiple videos on screen
AppDecImageProvider	Decoding and Color Conversion to a specific format (BGRA, BGRA64)	AppDecPerf	Multi-threaded, perf measurement

REDESIGNED SDK SAMPLES

Encode Applications

AppEncCUDA	Encoding CUDA surfaces	AppEncLowLatency	Low-latency encode, intra-refresh, slices etc.
AppEncD3D9	Encoding using D3D9 surfaces	AppEncME	ME-only mode
AppEncD3D11	Encoding using D3D11 surfaces	AppEncPerf	App for Encoder performance measurement
AppEncDec	Encoding & decoding in different threads, HDR streaming	AppEncQual	Encoding & quality measurement (PSNR)

OPTIMIZATION STRATEGIES

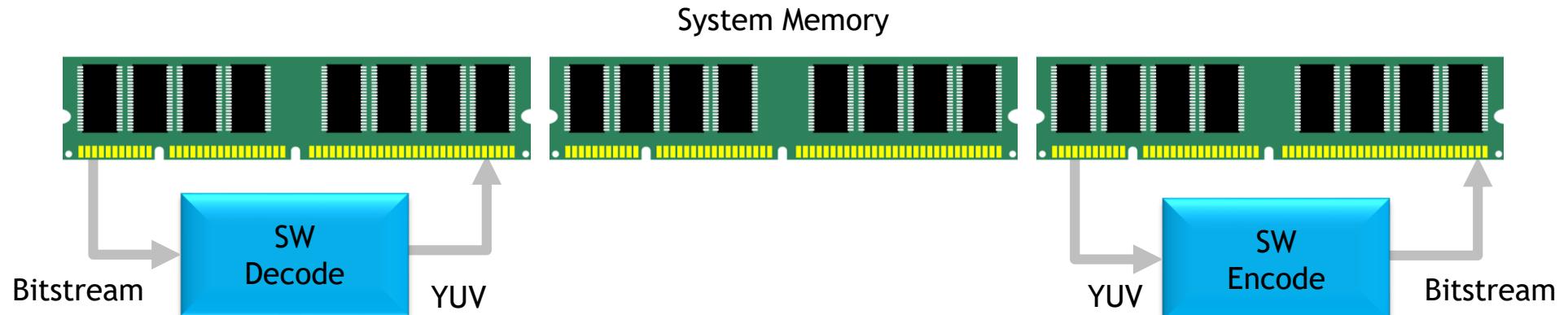
OPTIMIZATION STRATEGIES

General Guidelines

- Minimize PCIe transfers
 - Eliminate, if possible
 - Use CUDA for video pre-/post-processing
- Multiple threads/processes to balance enc/dec utilization
 - Monitor using nvidia-smi: `nvidia-smi dmon -s uc -i <GPU_index>`
 - Analyze using GPUView on Windows
- Minimize disk I/O
- Optimize encoder settings for quality/perf balance

SW TRANSCODE

```
ffmpeg -c:v h264 -i input.mp4 -c:a copy -c:v h264 -b:v 5M output.mp4
```

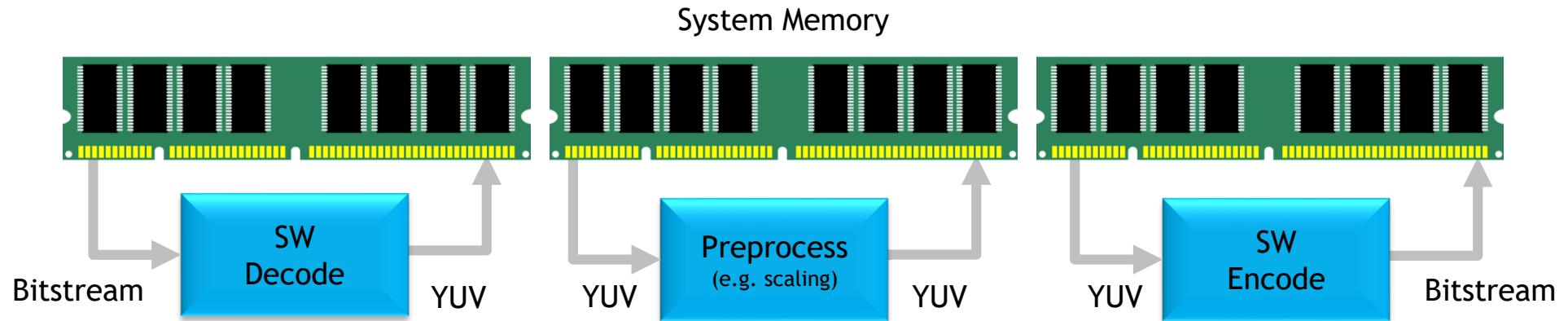


32 fps*

*1:2 transcode, fps per session
4 GHz Intel i7-6700K

SW TRANSCODE + SCALE

```
ffmpeg -c:v h264 -i input.mp4 -vf scale=1280:720 -c:a copy -c:v h264 -b:v 5M output.mp4
```

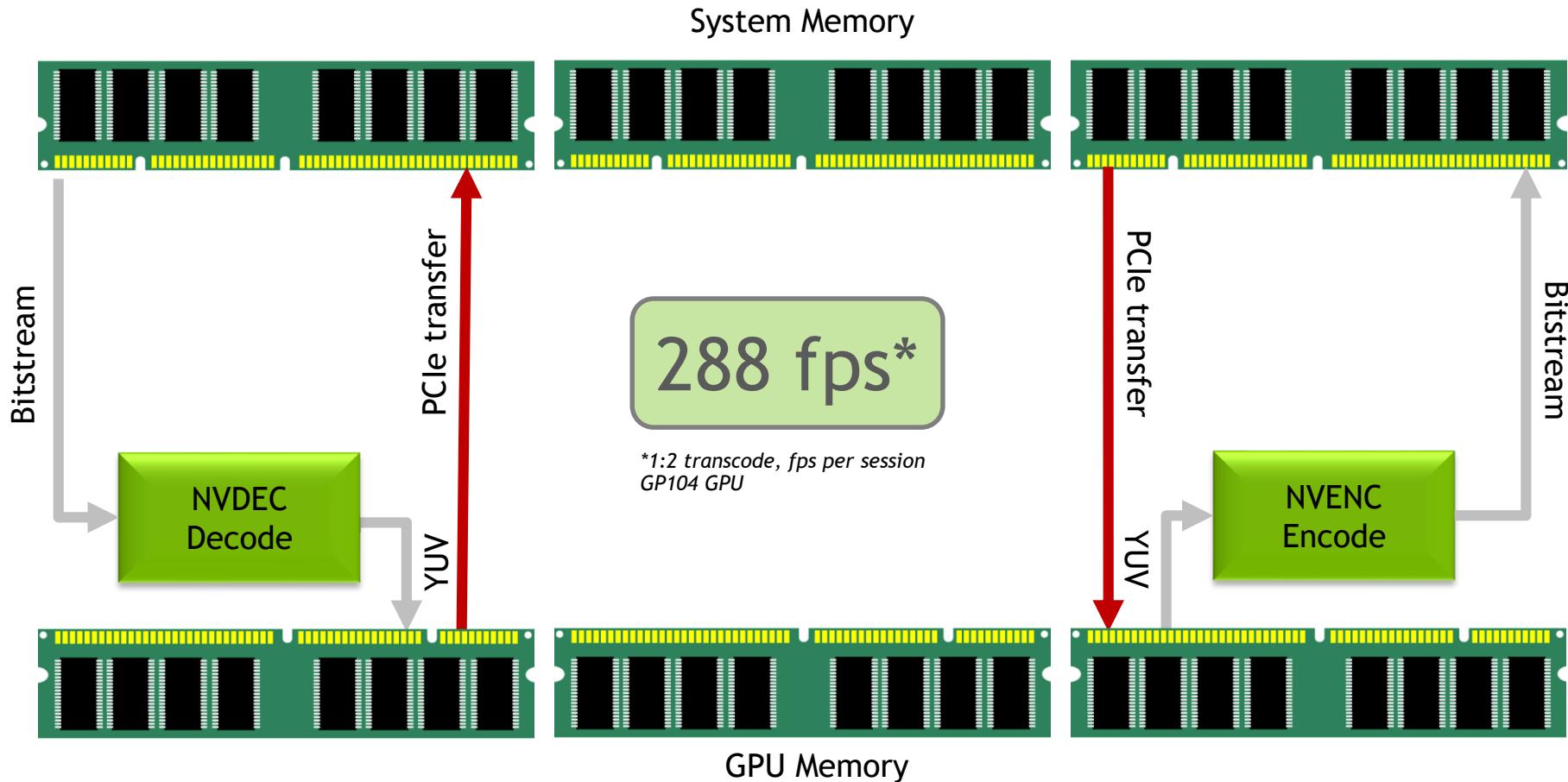


29 fps*

*1:2 transcode, fps per session
4 GHz Intel i7-6700K

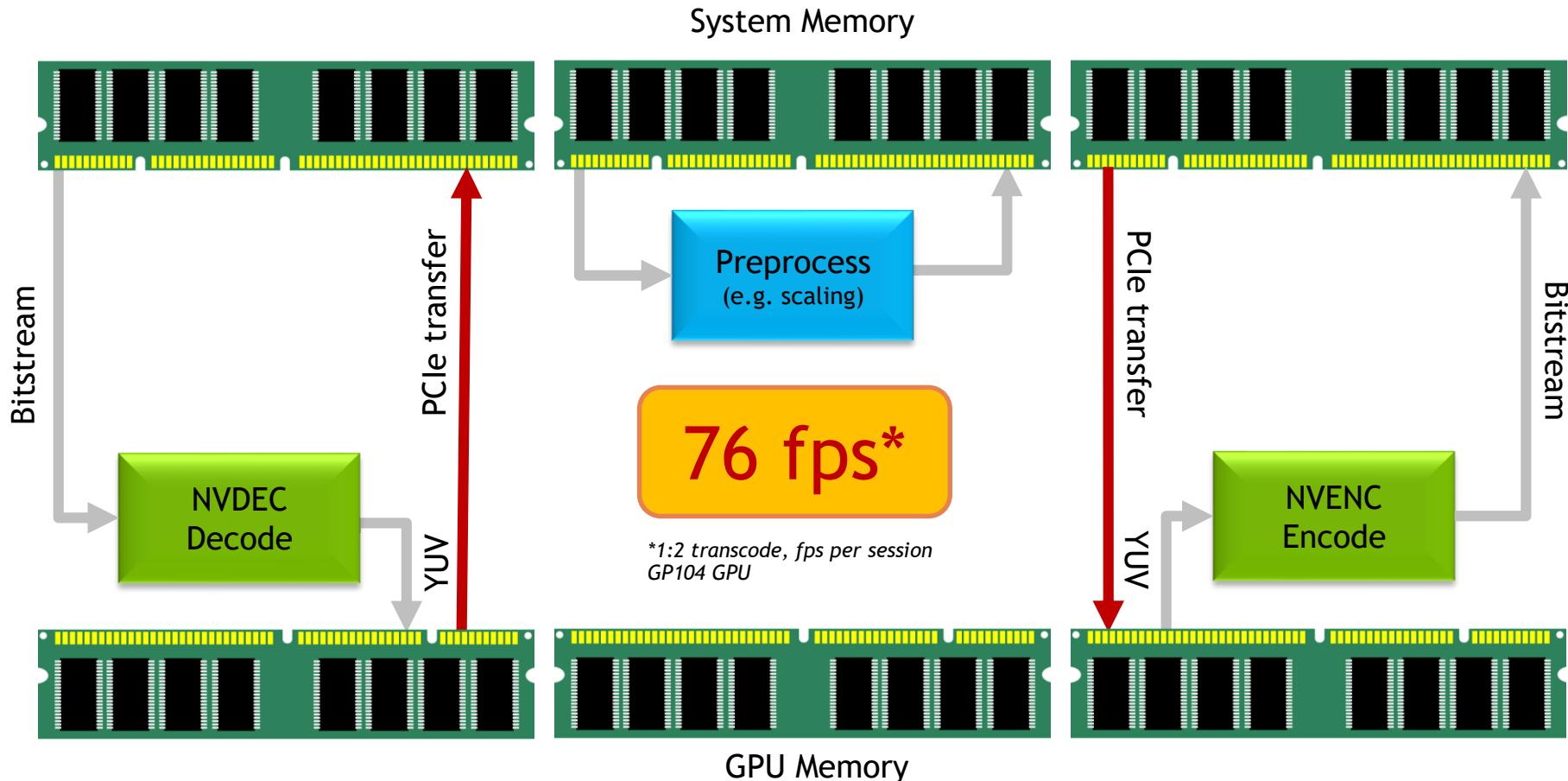
GPU UNOPTIMIZED TRANSCODE

```
ffmpeg -vsync 0 -c:v h264_cuvid -i input.mp4 -c:a copy -c:v h264_nvenc -b:v 5M output.mp4
```



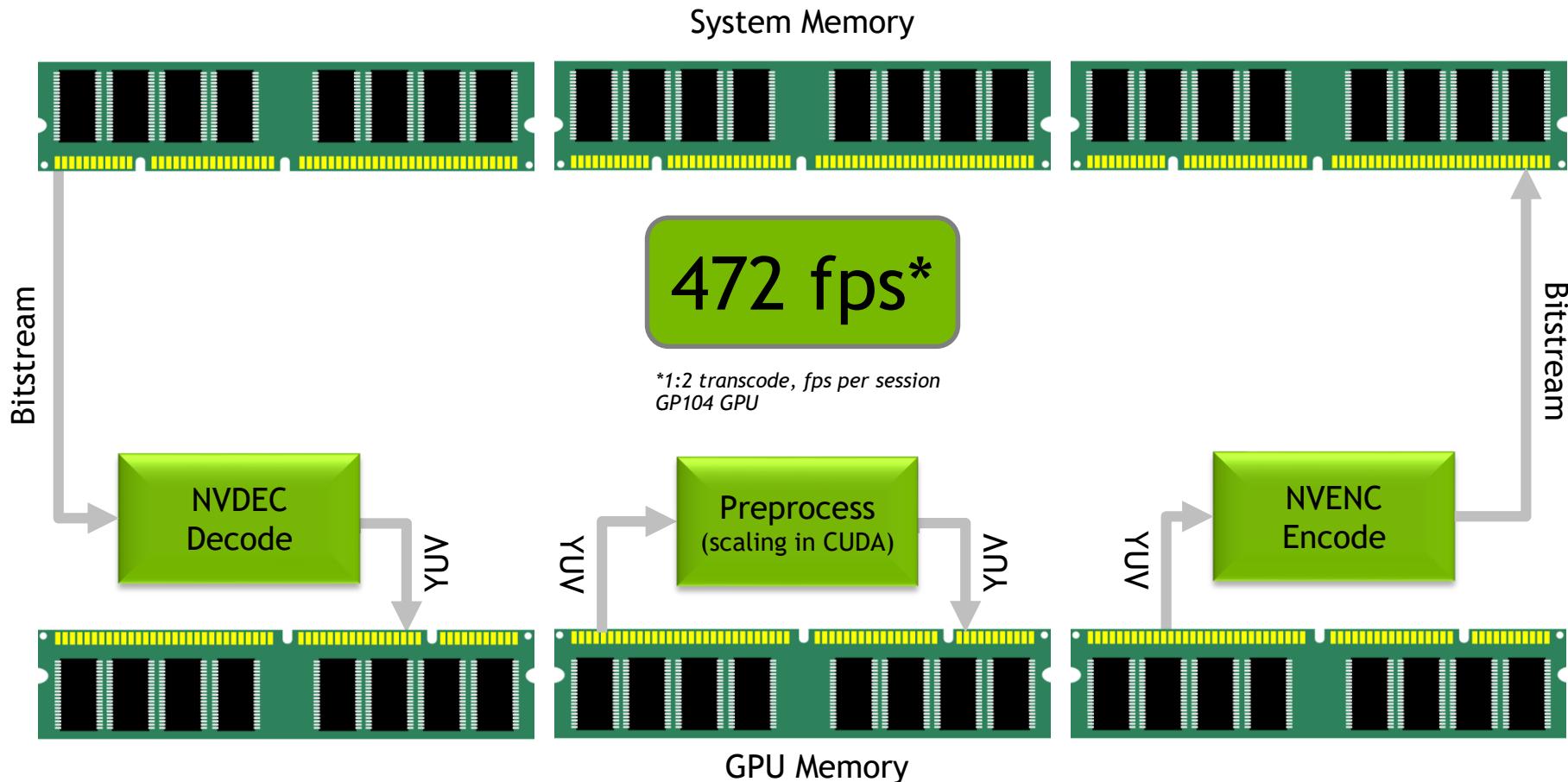
GPU UNOPTIMIZED TRANSCODE + CPU SCALE

```
ffmpeg -vsync 0 -c:v h264_cuvid -i input.mp4 -c:a copy -vf scale=1280:720 -c:v h264_nvenc -b:v 5M output.mp4
```



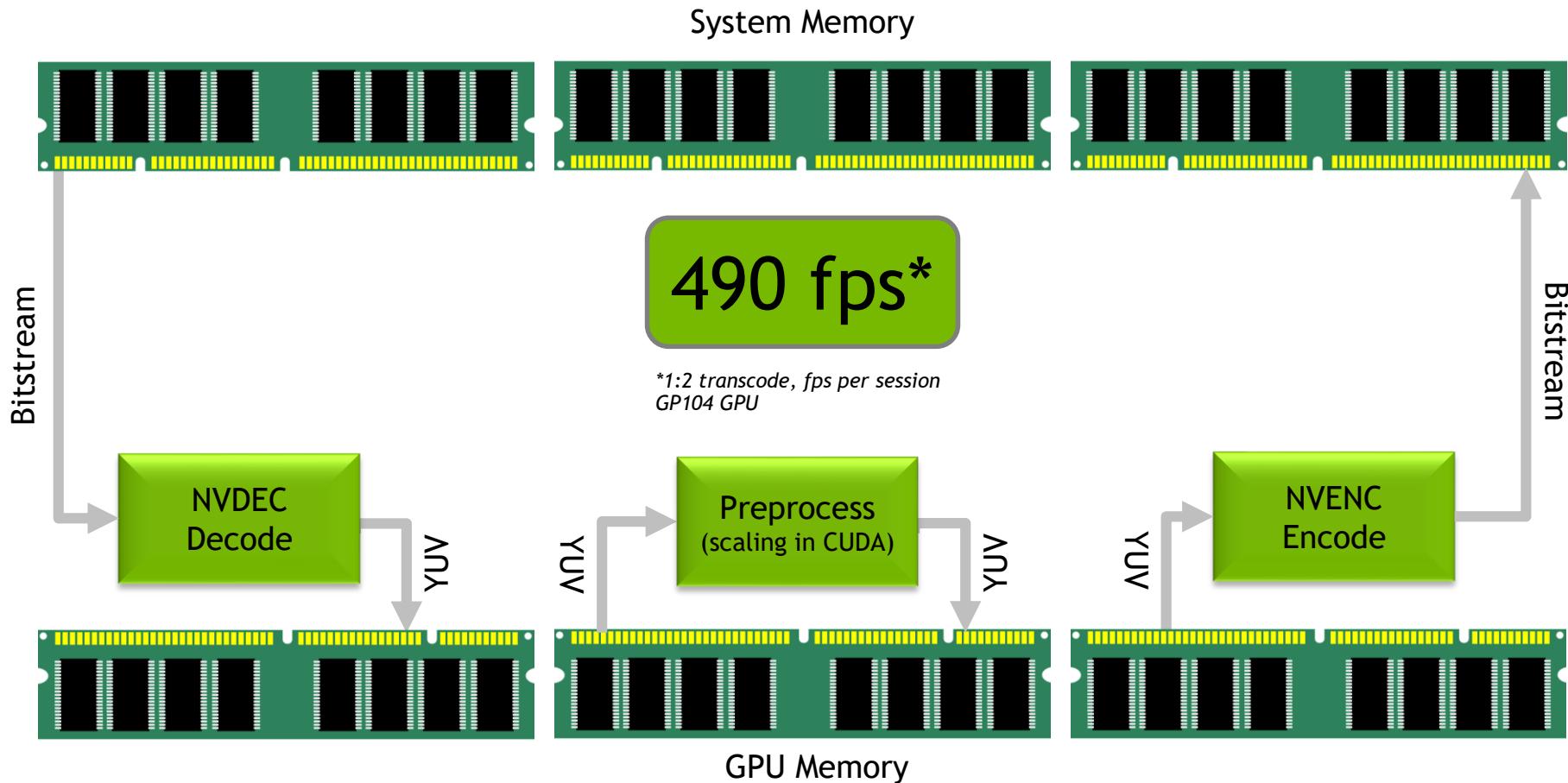
HIGH-PERF GPU OPTIMIZED TRANSCODE

```
ffmpeg -vsync 0 -hwaccel cuvid -c:v h264_cuvid -i input.mp4 -c:a copy -vf scale_npp=1280:720 -c:v h264_nvenc -b:v 5M output.mp4
```



HIGH-PERF GPU OPTIMIZED TRANSCODE

```
ffmpeg -vsync 0 -hwaccel cuvid -c:v h264_cuvid -resize 1280x720 -i input.mp4 -c:a copy -c:v h264_nvenc -b:v 5M output.mp4
```



FFMPEG VIDEO TRANSCODING

Tips

- Look at FFmpeg users' guide in NVIDIA Video Codec SDK package
- Use `-hwaccel` keyword to keep entire transcode pipeline on GPU
- Run multiple 1: N transcode sessions to achieve $M:N$ transcode at high perf

CUDA FILTERS IN FFmpeg

- -resize option with NVDEC (e.g. `-c:v h264_cuvid -resize 1280x720 ...`)
- `scale_npp`: Built-in CUDA library filters
- Custom CUDA filter examples in FFmpeg
 - `scale_cuda`
 - `thumbnail_cuda`
 - Build your own using above as guide
- If you *must* use CPU and GPU filters, minimize PCIe x'fers

MIXING CPU & GPU FILTERS

Fade (CPU) + Scale (GPU)

Why doesn't this work?

```
ffmpeg.exe -y -c:v h264_cuvid -i input.264 -vf "fade,scale_npp=1280:720" -c:v h264_nvenc output.264
```

This works

```
ffmpeg.exe -y -c:v h264_cuvid -i input.264 -vf "fade,hwupload_cuda,scale_npp=1280:720" -c:v h264_nvenc  
output.264
```

MIXING CPU & GPU FILTERS

Scale (GPU) + Fade (CPU)

Why doesn't this work?

```
ffmpeg.exe -y -c:v h264_cuvid -i input.264 -vf "hwupload_cuda,scale_npp=1280:720,hwdownload,fade" -c:v h264_nvenc output.264
```

One solution

```
ffmpeg.exe -y -c:v h264_cuvid -i input.264 -vf  
"hwupload_cuda,scale_npp=1280:720,hwdownload,format=nv12,fade" -c:v h264_nvenc output.264
```

Optimal solution

```
ffmpeg.exe -y -hwaccel cuvid -c:v h264_cuvid -i input.264 -vf  
"scale_npp=1280:720,hwdownload,format=nv12,fade" -c:v h264_nvenc output.264
```

OPTIMIZATION TIPS

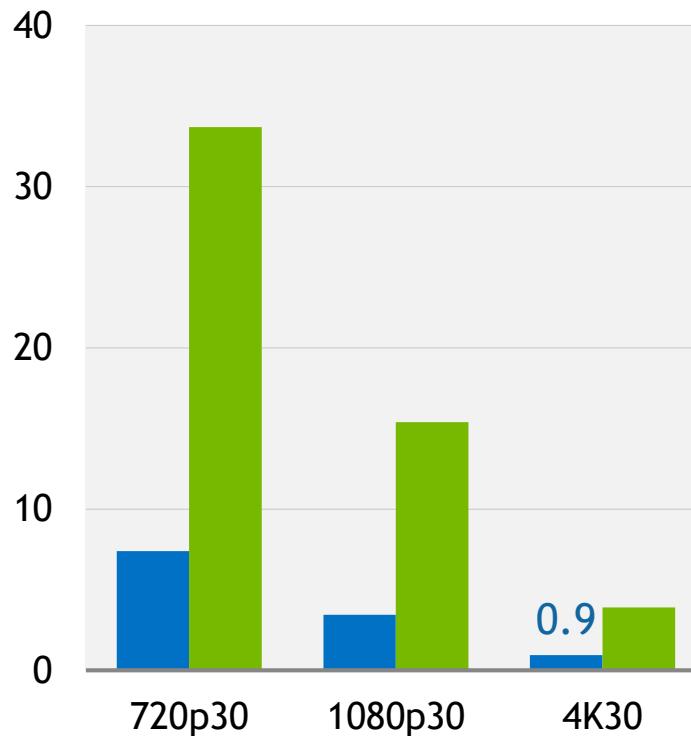
- Write your own CUDA filters
- Combine CUDA filters; e.g. scaling + color space conversion in a single filter
- For systems with multiple CPU sockets, avoid accesses to *local* sysmem of one CPU from another CPU. Find the local NUMA node and *localize* the storage *per CPU*.

BENCHMARKS

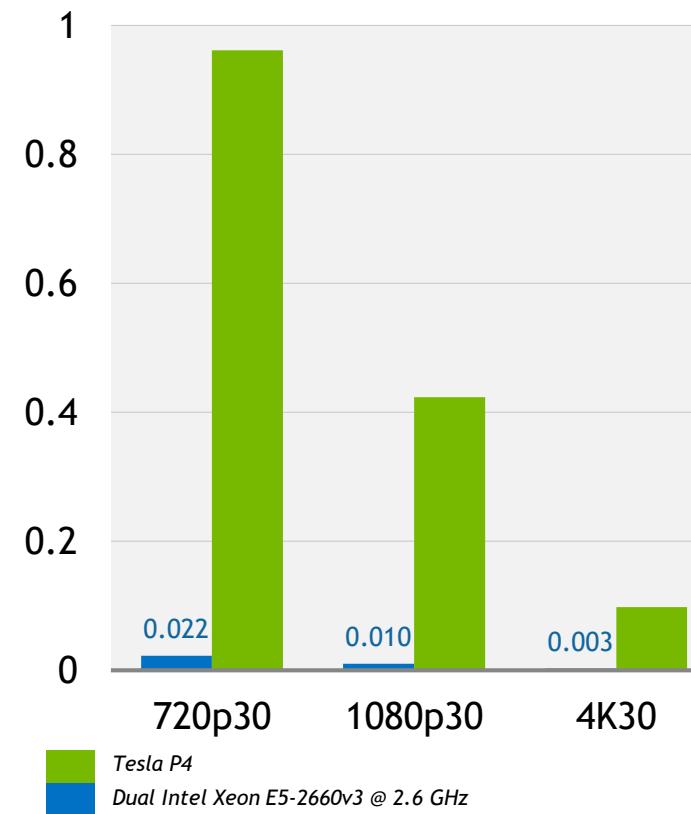
P4: 5X MORE H.264 ENCODE THAN 2S CPU SERVER

Up to 5x more throughput, up to 10x better efficiency at ~ quality

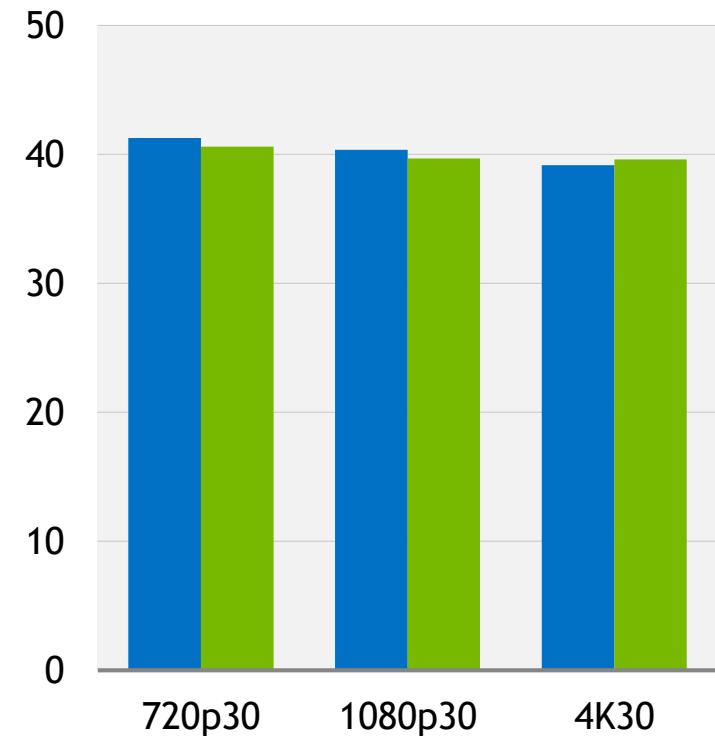
*H.264 hq Encode Throughput
(Streams)*



*H.264 hq Encode Efficiency
(Streams / Watt)*



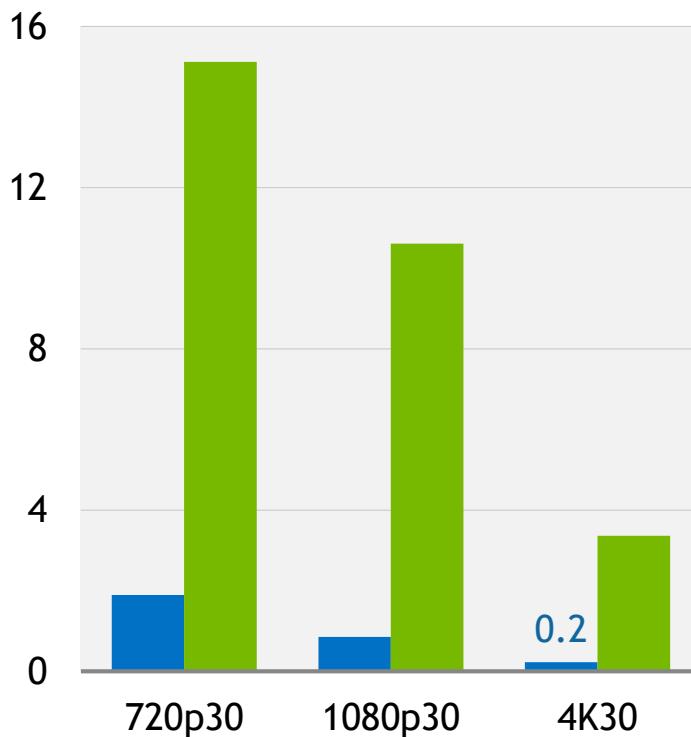
*H.264 hq Encode Quality
(PSNR YUV)*



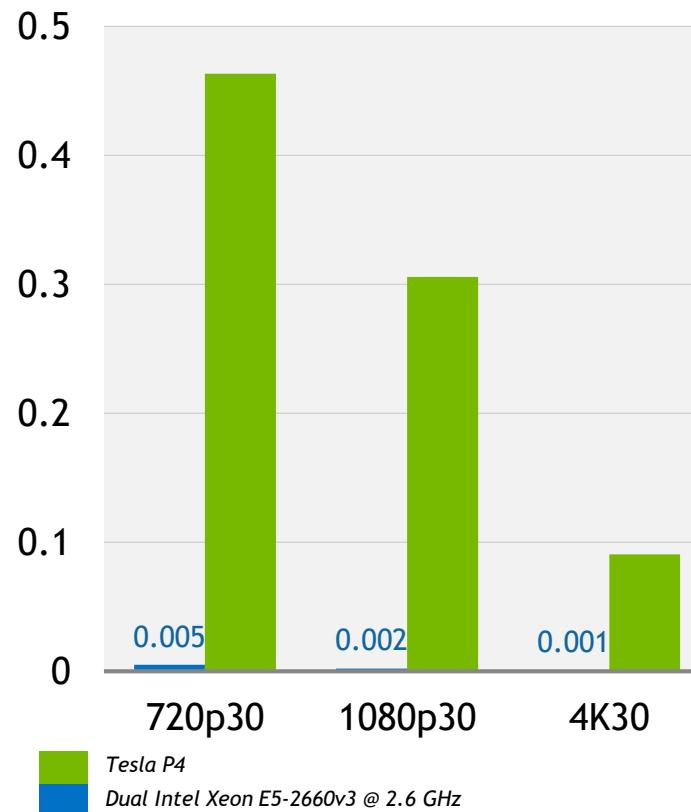
P4: REAL-TIME HEVC 4K60 ENCODE

Up to 15x more throughput, up to 30x better efficiency at ~ quality

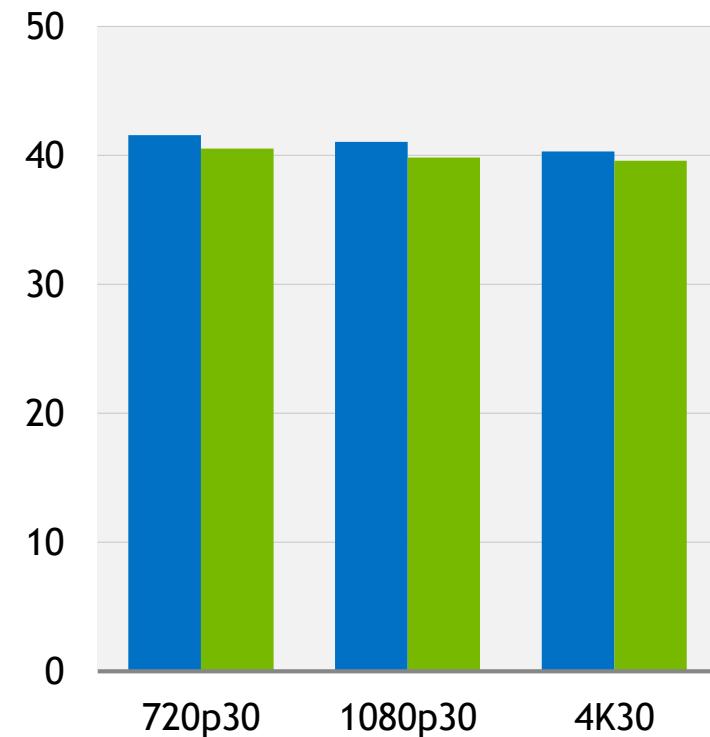
*H.265 hq Encode Throughput
(Streams)*



*H.265 hq Encode Efficiency
(Streams / Watt)*

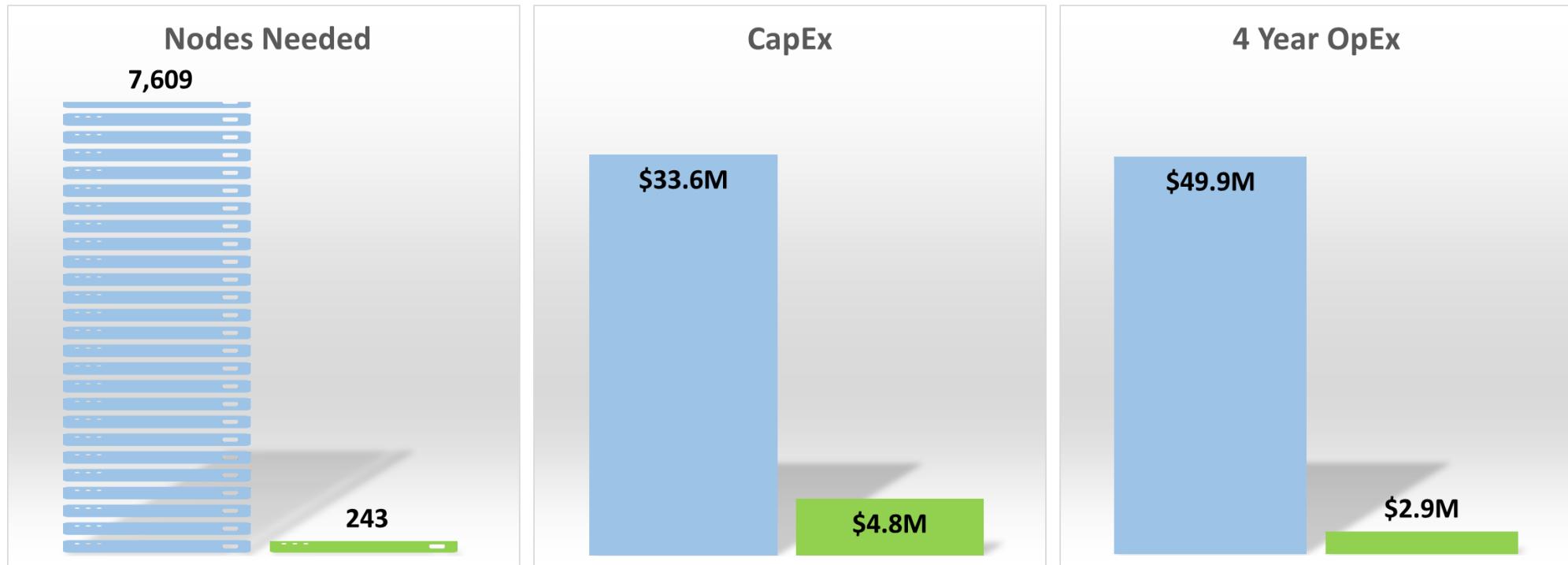


*H.265 hq Encode Quality
(PSNR YUV)*



GPU ENCODE REDUCES CAPEX 7X, OPEX 17X

Transcoding 20,000 720p30 Streams + 20,000 1080p30 H.264 Streams, hqslow



CPU Nodes 2xE5-2660v3, 128GB DDR4, 512GB SSD, 25 GE. Node price including core network \$4500

GPU Nodes 2xE5-2660v3, 8xP4 PCIe, 128GB DDR4, 512GB SSD, 25 GE

ROADMAP

ROADMAP

Video Codec SDK 8.2

- Q2 2018
 - Decode + inference optimizations
 - Reconfigure decoder without reinitialization
 - No init time, reuse context, lowers memory fragmentation
 - Report decoder errors
 - Inference can continue up to error slice
 - HEVC I-frame only decoding (H.264 already supported) - Q3 2018
 - Lower memory, IVA use-case

RESOURCES

Video Codec SDK: <https://developer.nvidia.com/nvidia-video-codec-sdk>

FFmpeg GIT: <https://git.ffmpeg.org/ffmpeg.git>

FFmpeg builds with hardware acceleration: <http://ffmpeg.zeranoe.com/builds/>

Video SDK support: video-devtech-support@nvidia.com

Video SDK forums: <https://devtalk.nvidia.com/default/board/175/video-technologies/>

Connect with experts (CE8107): Today, 26th March at 3:00 pm

