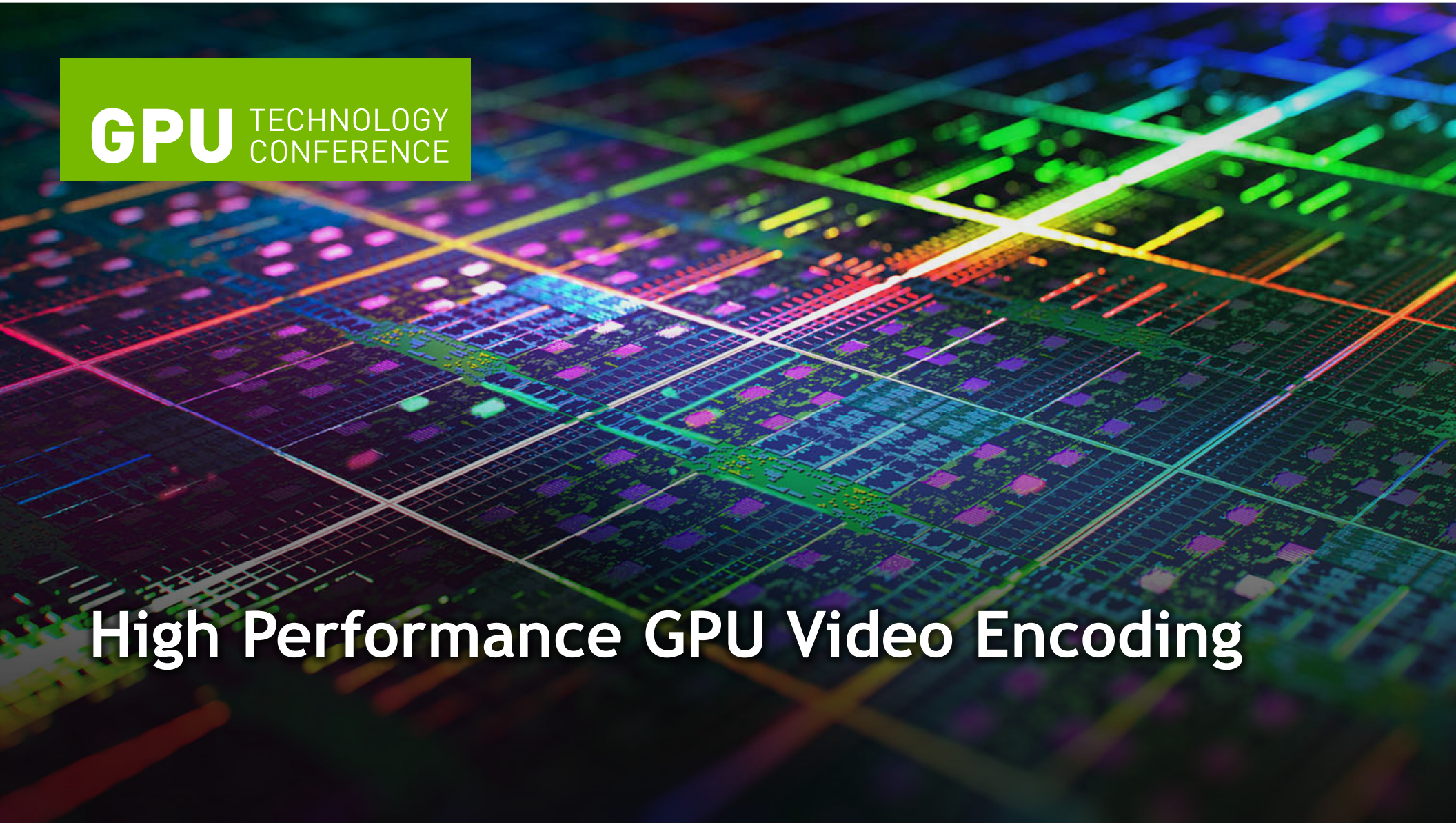


High Performance GPU Video Encoding

GPU TECHNOLOGY CONFERENCE

Agenda

- Why GPU Video Encoding
- NVIDIA H.264 Video Encoding Solutions
- Hardware Architecture
- Software Architecture
- Performance

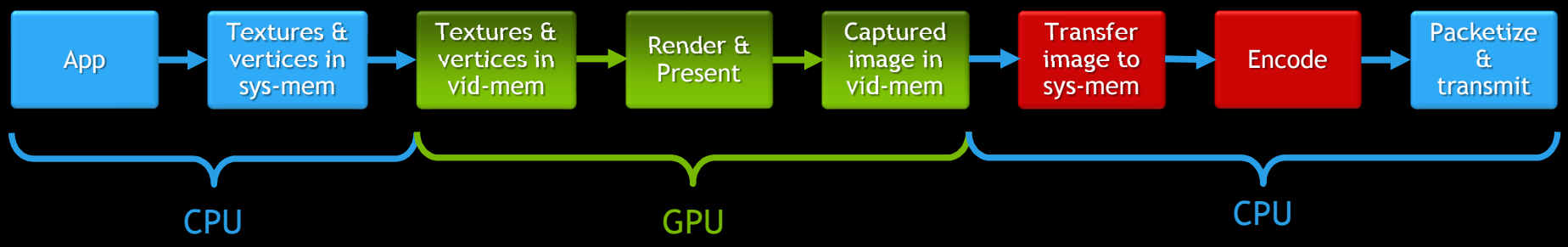
Why GPU Video Encoding?

Benefits of Encoding on GPU

- Low power
  - Fixed function hardware
  - Reduced memory transfers
- Low latency
- High performance
- Higher density
  - 2x channel density @ ~50% power consumption
- Scalability

GPU TECHNOLOGY CONFERENCE

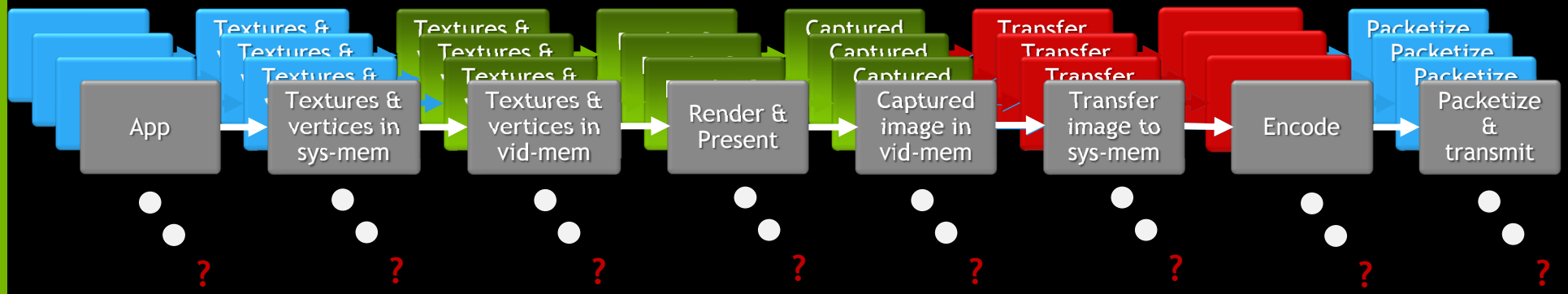
Cloud Streaming - Encode on CPU



Large memory transfers	Power, Latency
CPU-intensive tasks	Power
CPU-intensive tasks	Cost/seat ↑ , channel density ↓

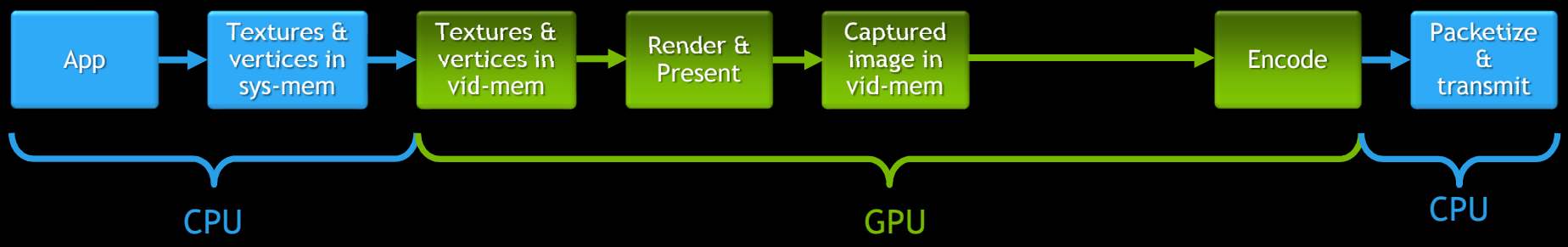
GPU TECHNOLOGY CONFERENCE

Cloud Streaming - Encode on CPU



Large memory transfers	Power, Latency
CPU-intensive tasks	Power
CPU-intensive tasks	Cost/seat ↑ , channel density ↓
Fixed number of CPUs	Limited scalability

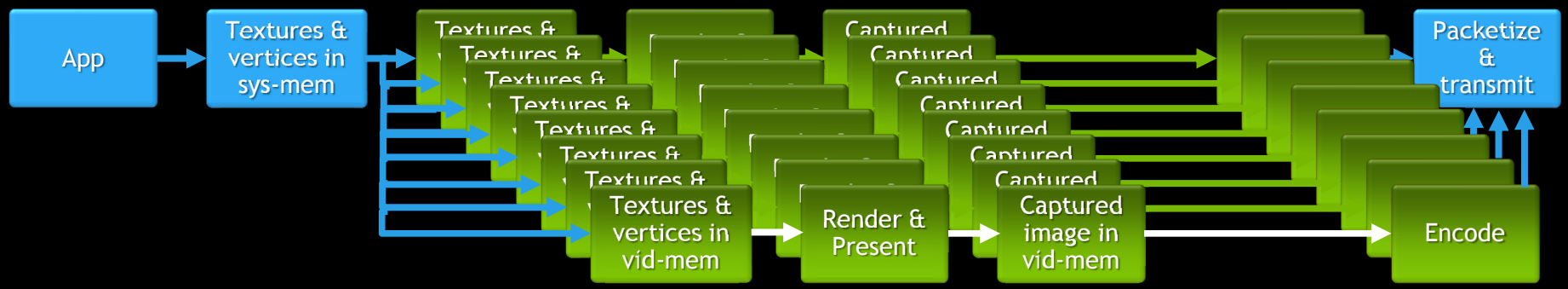
Cloud Streaming - Encode on GPU



No large memory transfers	Low-power, low-latency
Use CPU only where needed	Low power
Use CPU only where needed	Cost/seat ↓, channel density ↑

GPU TECHNOLOGY CONFERENCE

Cloud Streaming - Encode on GPU



No large memory transfers	Low-power, low-latency
Use CPU only where needed	Low power
Use CPU only where needed	Cost/seat ↓, channel density ↑
Add GPUs as needed	Excellent scalability

GPU TECHNOLOGY CONFERENCE

NVIDIA H.264 Video Encoding Solutions

GPU TECHNOLOGY CONFERENCE

NVIDIA H.264 Video Encoding Solutions

NVIDIA CUDA

- Hybrid processing (CPU + CUDA)
  - ME, intra-prediction, mode decision in CUDA
  - VLE on CPU
- Performance scales with CUDA cores
- Works on all GPUs (Tesla, Fermi, Kepler, …)

NVIDIA NVENC

- Fully hardware accelerated
  - ME, intra-prediction, mode decision, VLE
- High performance, low power
- Kepler+ GPUs

GPU TECHNOLOGY CONFERENCE

NVIDIA H.264 Video Encoding Solutions

NVIDIA CUDA

- Distributed with CUDA SDK libraries
- No low-latency streaming
- All Platforms – GeForce, Quadro, Tesla, GRID
- Windows only

NVIDIA NVENC

- Proprietary software API
- Optimized for low-latency streaming
- Better visual quality
- Quadro, GRID and Tesla
- Windows & Linux

GPU TECHNOLOGY CONFERENCE

Power vs. Performance

GPU TECHNOLOGY CONFERENCE

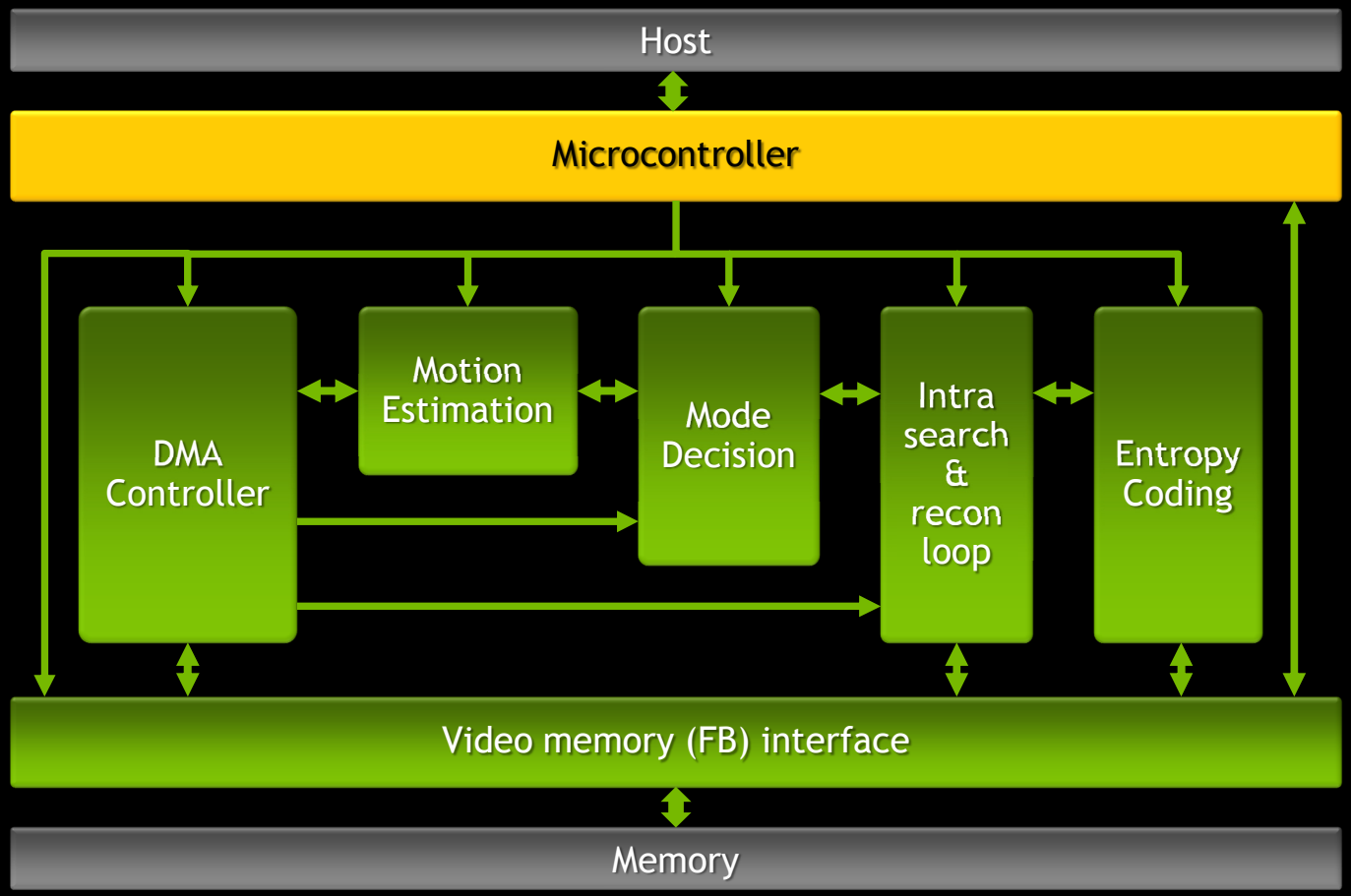
NVENC Features

Feature	What it enables
H.264 base, main, high profiles	Wide range of use-cases
Up to 8x HD encode (1080p @ 240 fps)	Faster than real-time encoding
Flexible ME, QP maps	Customizable quality, region of interest encoding
YUV 4:2:0 and planar 4:4:4 support	High quality encoding without chroma subsampling
MVC	Full resolution stereo encode
Up to 4096 × 4096 in HW	High resolution encode
API	NVENC SDK (Flexible API, Win/Linux, x86) GRID SDK (Capture+ Encode, Win - now, Linux - future)
NVENC and CUDA parallelism	Simultaneous and parallel HW and CUDA encoding for increased performance

GPU TECHNOLOGY CONFERENCE

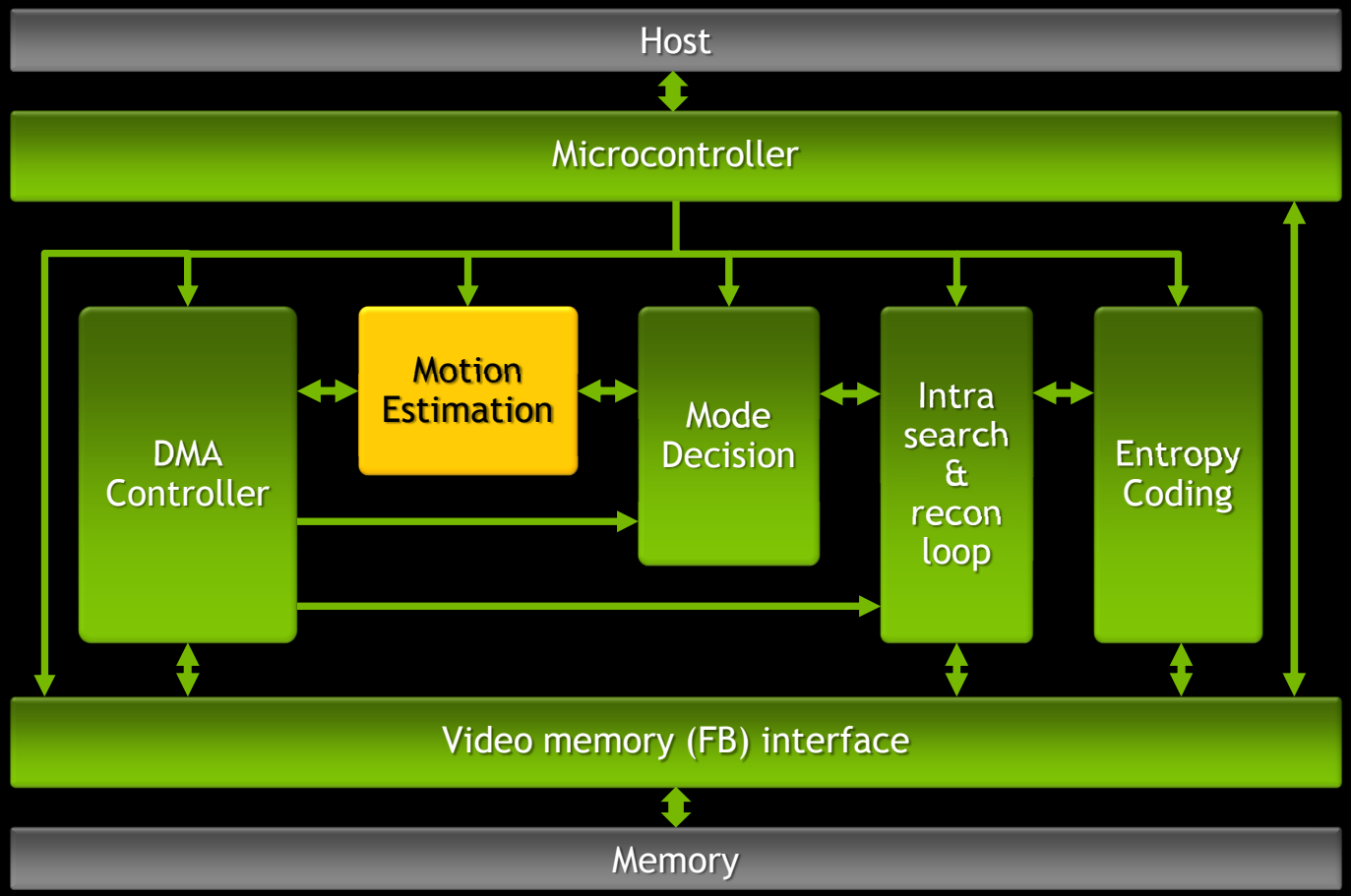
NVENC Hardware Architecture

NVENC Arch: Microcontroller



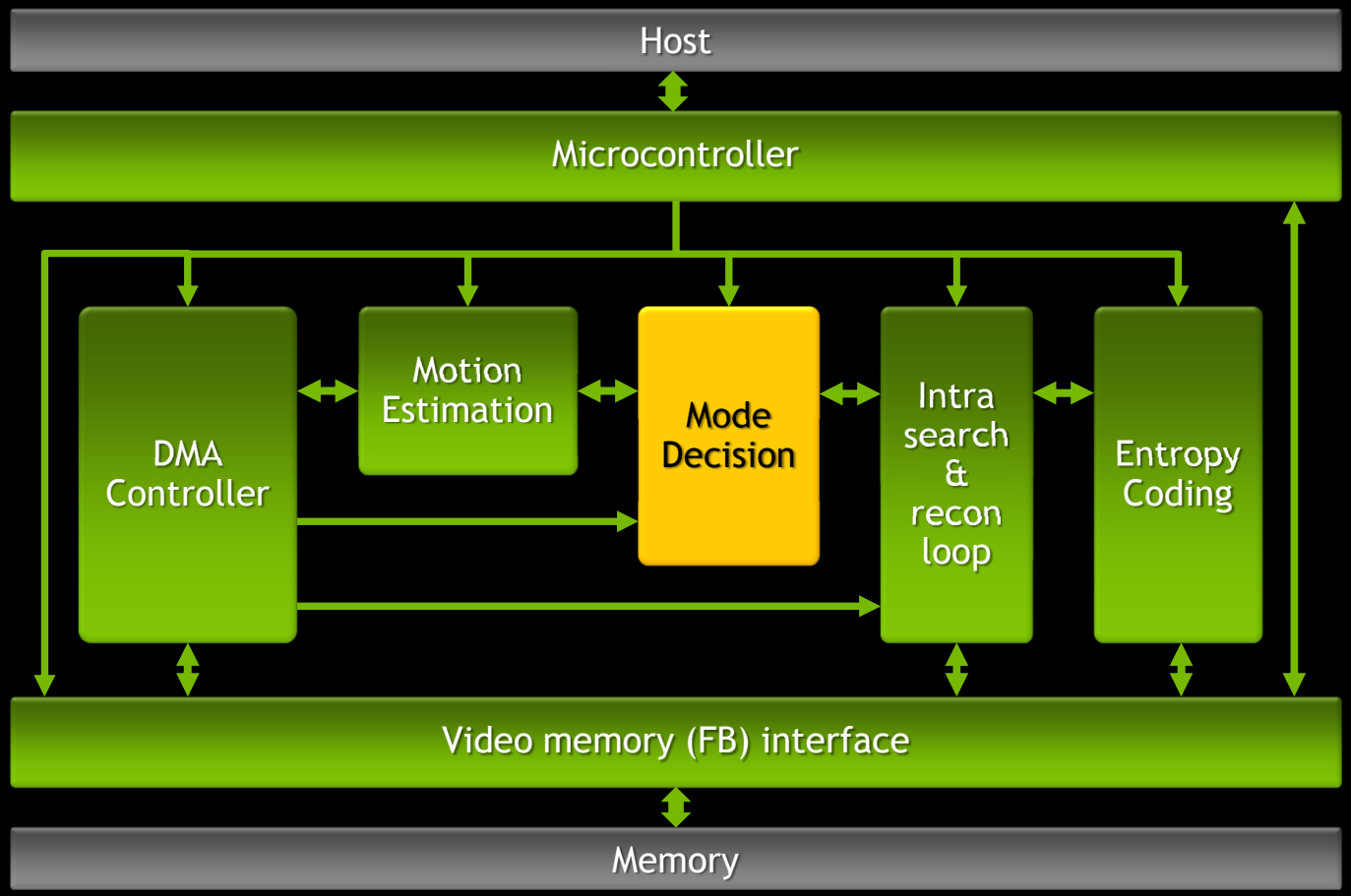
- NVIDIA proprietary microcontroller
- Runs firmware
- Programs encoder blocks
- Rate control

NVENC Arch: Motion Estimation



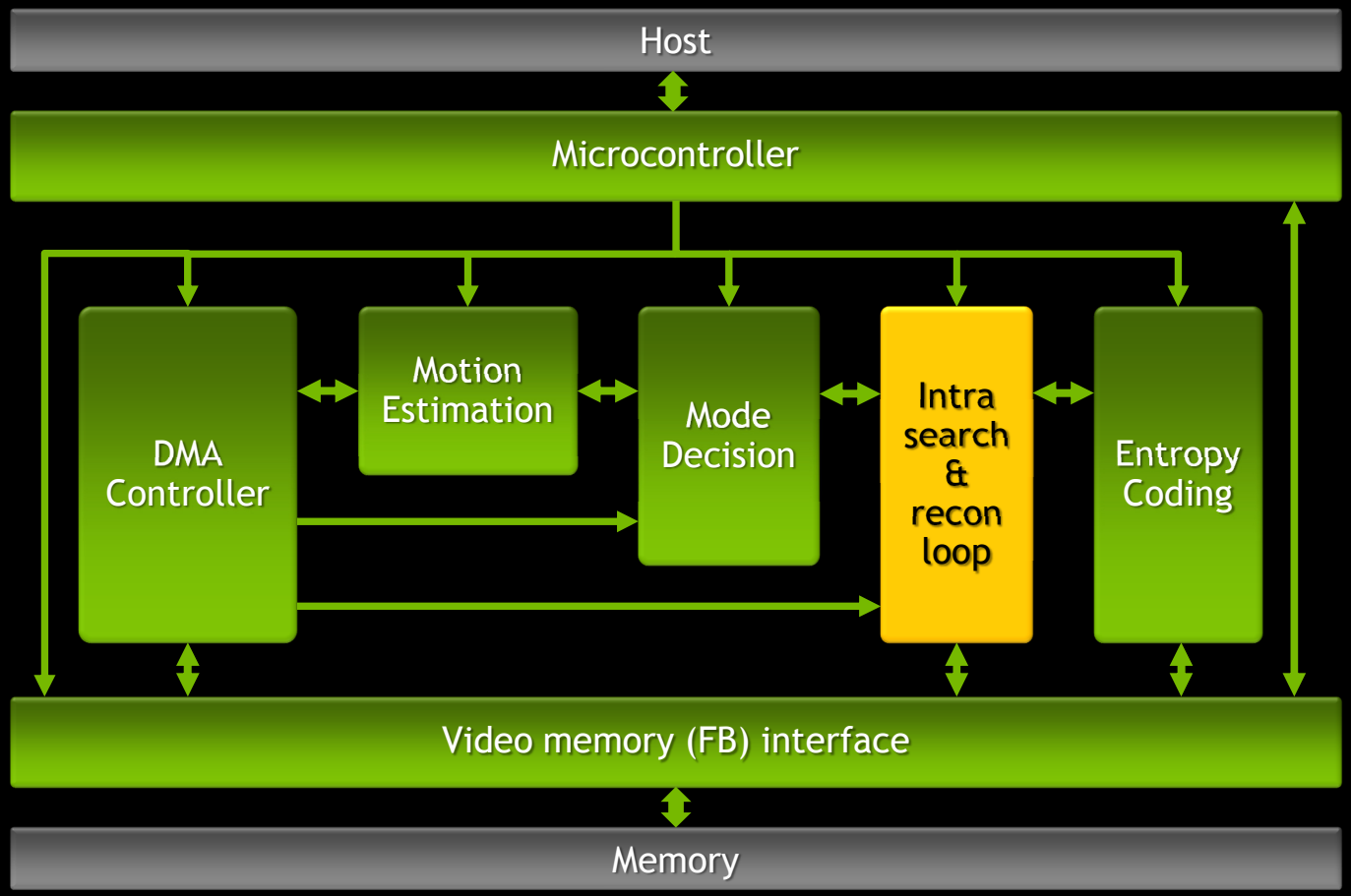
- Exhaustive full-pel search (L0, L1, Bi)
  - Temporal
  - Spatial
  - Coloc
  - Constant
  - External
- Half-pel and quarter-pel refinement
- Motion compensation

NVENC Arch: Mode Decision



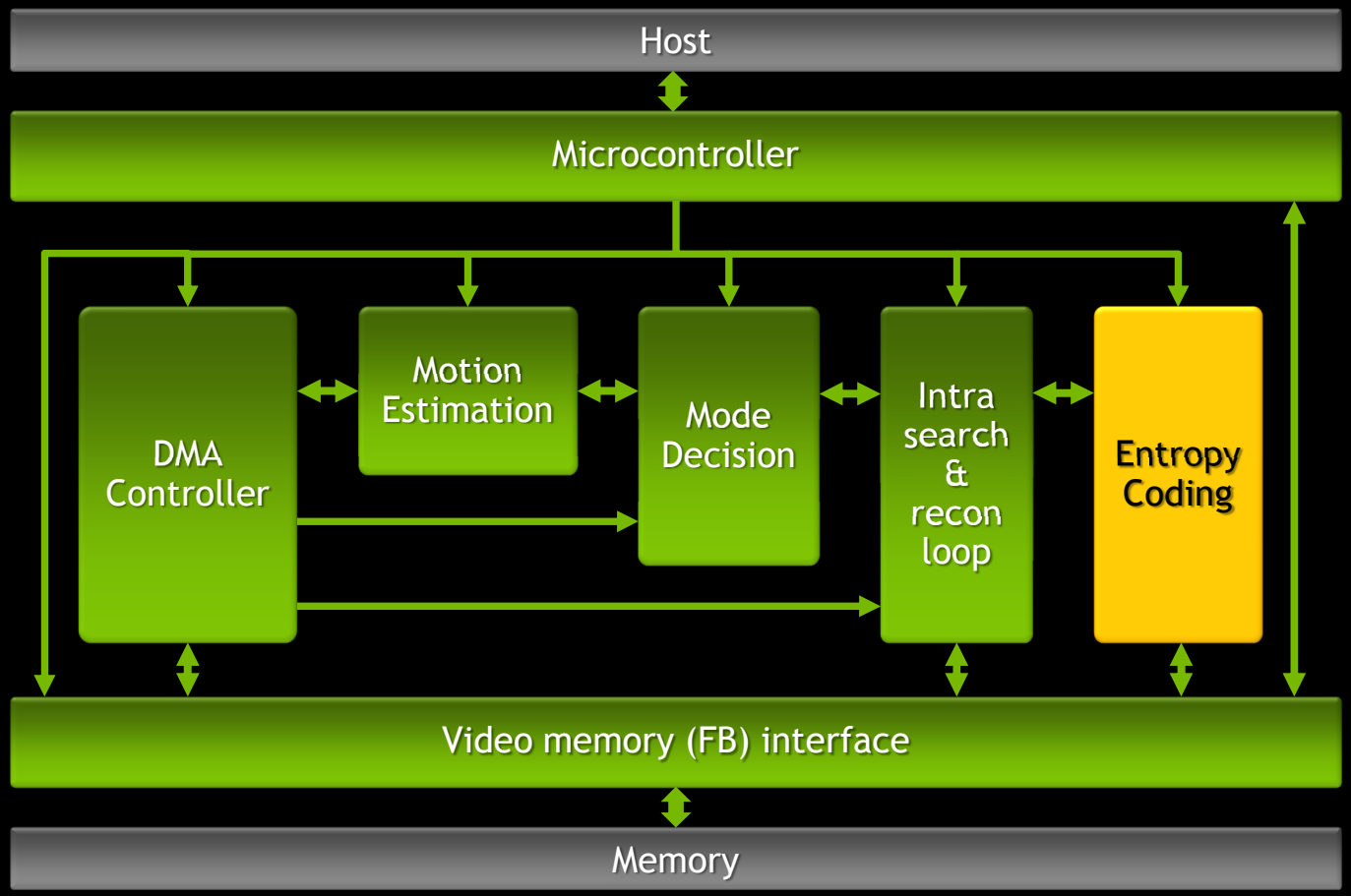
- Calculates inter-MB cost
- Compares to intra-MB cost and decides final winner

NVENC Arch: Microcontroller



- H.264 intra search
- Forward DCT & quantization
- Recon-loop (IDCT, IQT, deblocking)

NVENC Arch: Microcontroller



- CAVLC & CABAC

GPU TECHNOLOGY CONFERENCE

NVENC Software Architecture

Using NVENC

Direct Encode

NVENC SDK

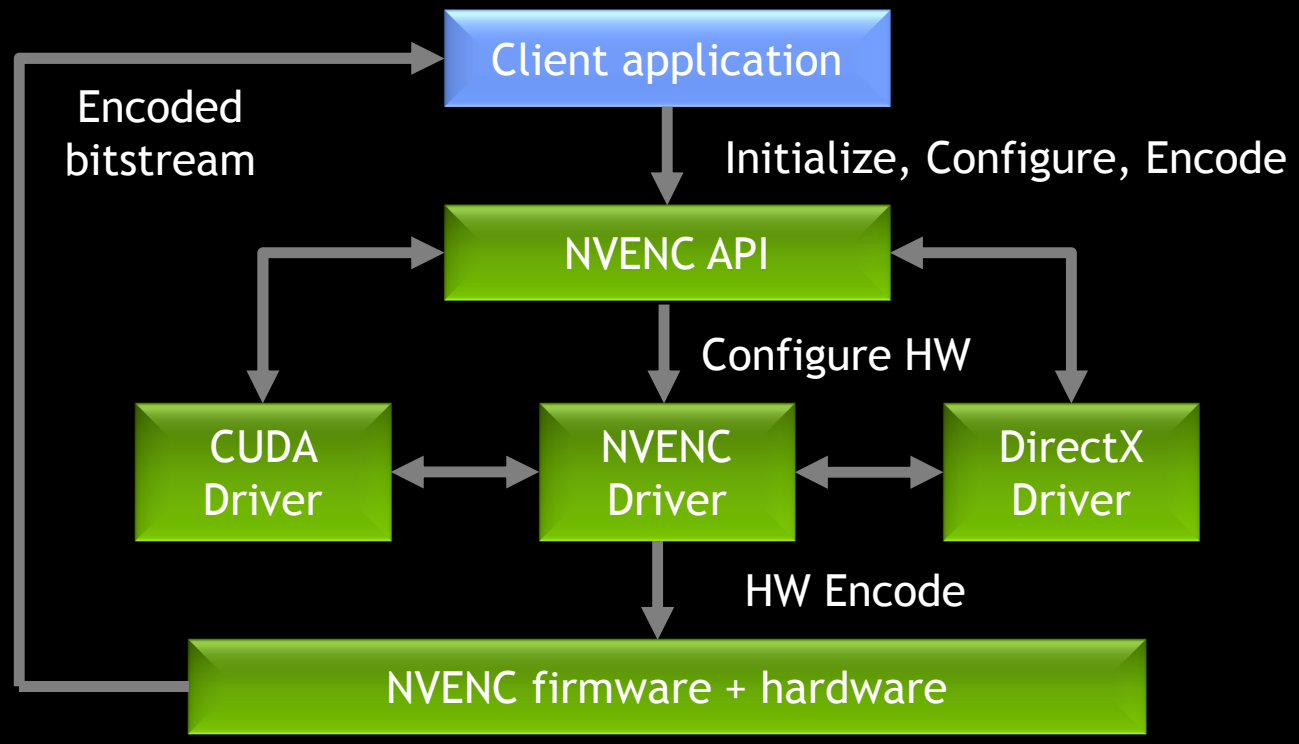
- No capture
- Transcoding
- Archiving
- Video editing
- CUDA + encoding
- D3D, CUDA interop
- Exhaustive encoder settings

Capture + Encode

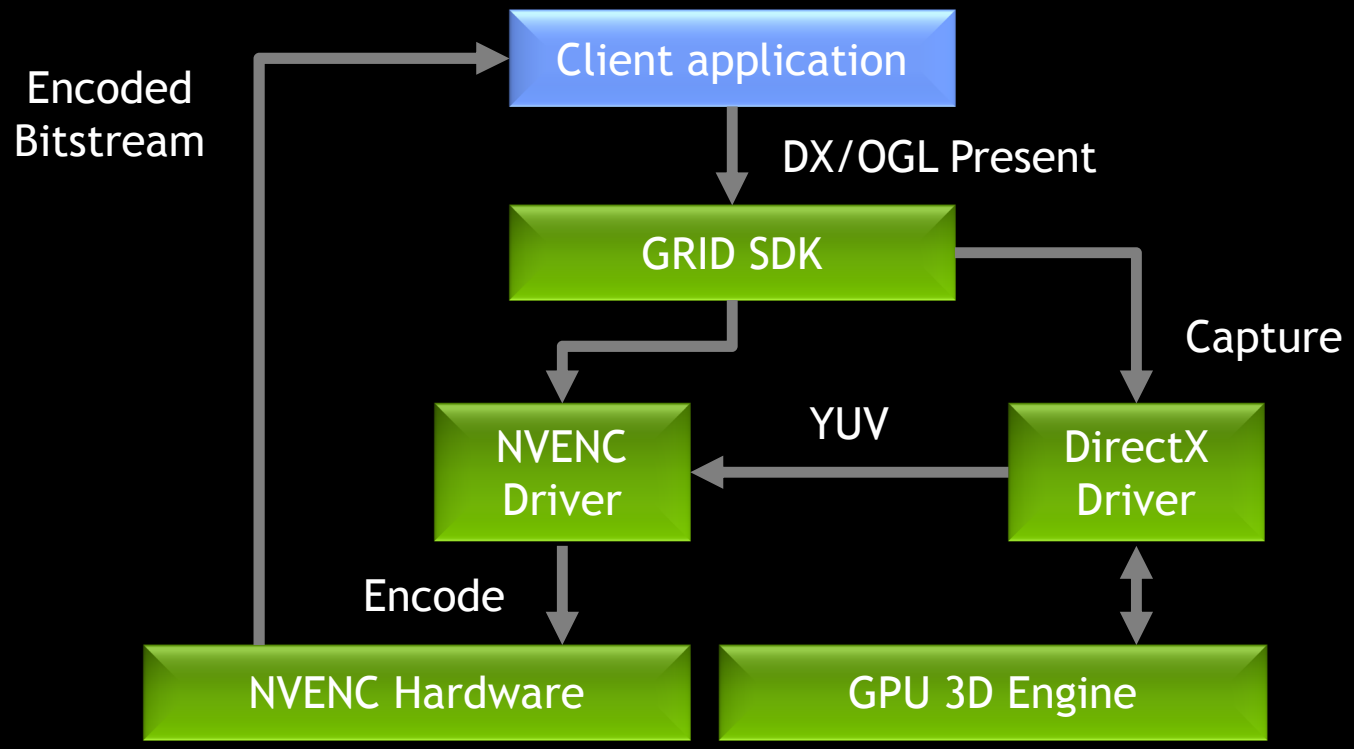
GRID SDK

- Capture + encode
- Optimized for low-latency apps
- Limited encoder settings

Direct NVENC Encode (NVENC SDK)



Capture and Encode (GRID SDK)



NVENC SDK

- Available on NVIDIA developer zone
  - https://developer.nvidia.com/nvidia-video-codec-sdk
- .DLL/.so, interface header, documentation, sample apps
- Unified API for Windows and Linux
- Works on x86/x64
- Various presets and API's for
  - Transcoding
  - Video conferencing
  - Remote graphics (Cloud gaming, remote desktop, capture & stream)
- Supports CBR, VBR rate control

NVENC SDK (Contd.)

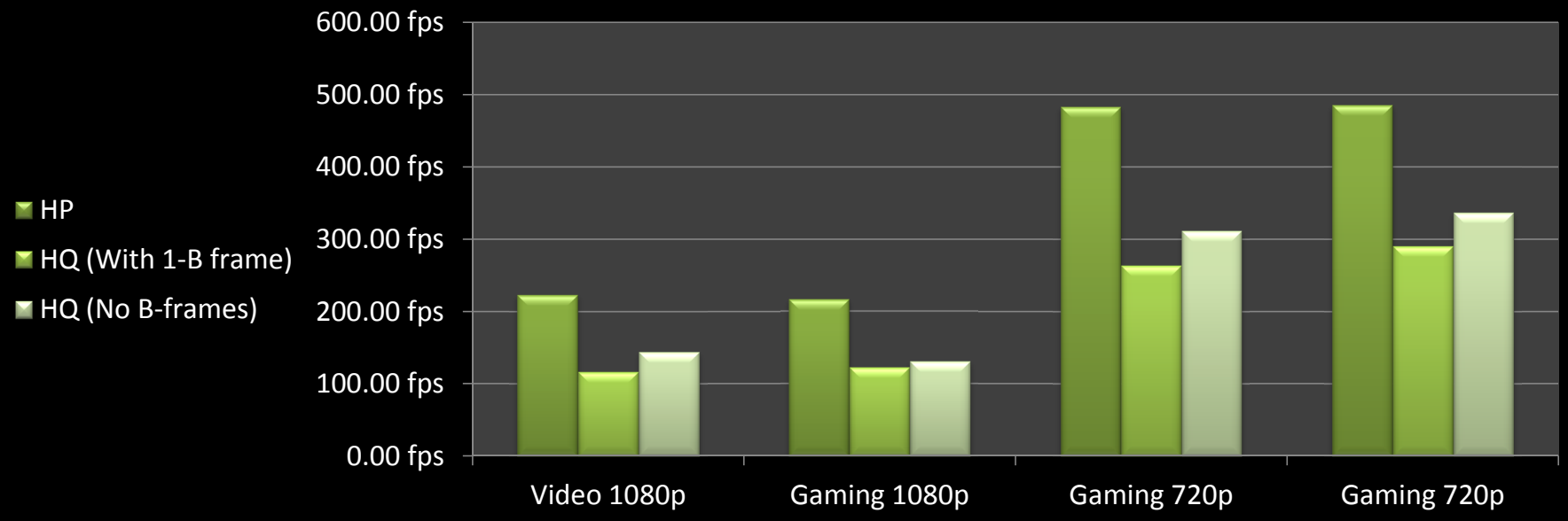
- Advanced features
  - Dynamic resolution change
  - Dynamic bitrate change
  - Reference picture invalidation
  - Temporal SVC
  - Intra-refresh
  - Two-pass rate control for constant quality

GRID SDK Encode

- Licensed from NVIDIA
- .DLL/.so, interface header, documentation, sample apps
- Windows (now) and Linux (future)
- Works on x86/x64
- Various presets and API's for
  - Remote graphics (Cloud gaming, remote desktop, capture & stream)
- Optimized for low latency and high quality

GPU TECHNOLOGY CONFERENCE

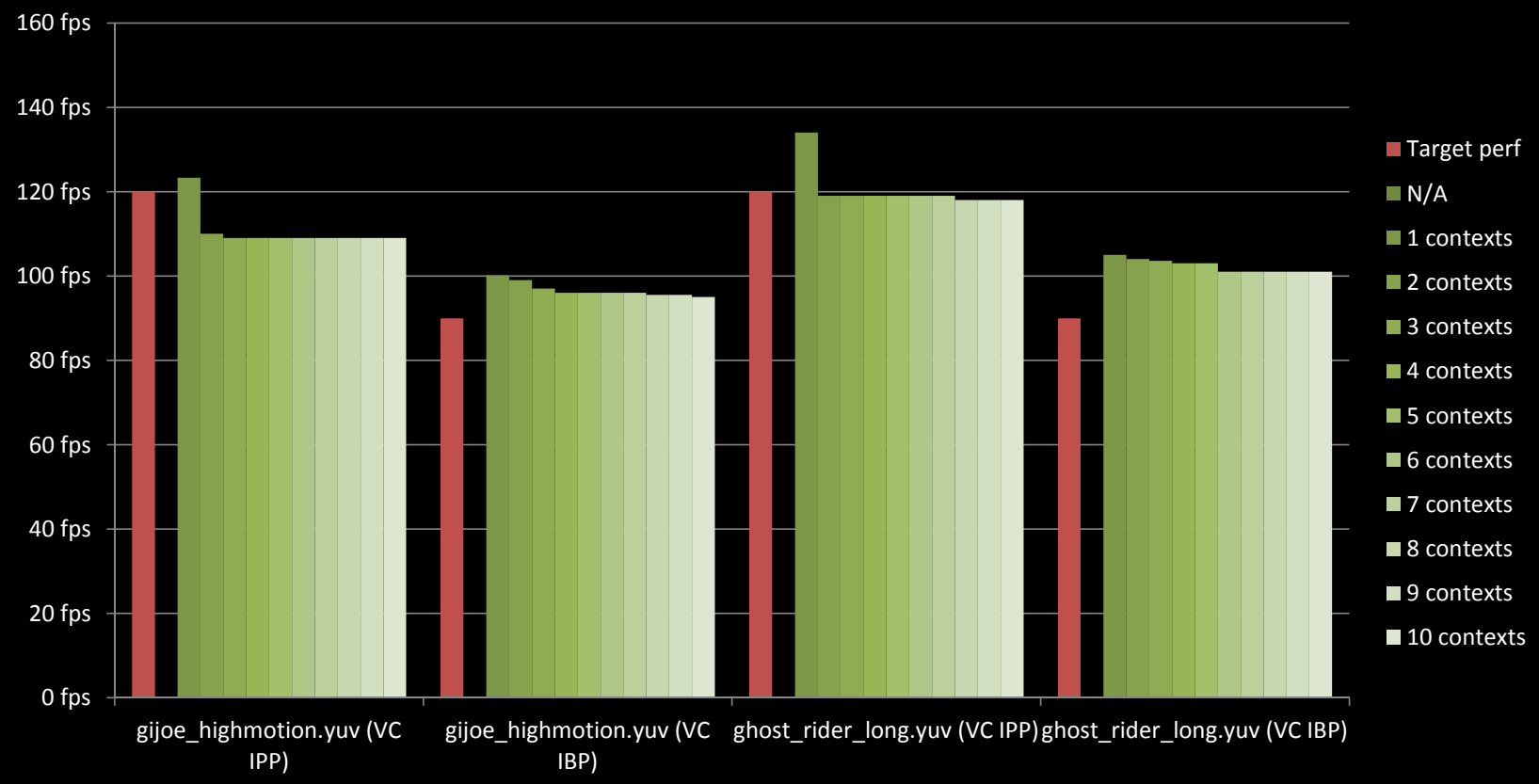
NVENC Performance



Preset	Video 1080p	Gaming 1080p	Gaming 720p	Gaming 720p
HP	223.21 fps	216.45 fps	483.09 fps	485.44 fps
HQ (With 1-B frame)	116.69 fps	122.55 fps	263.85 fps	290.70 fps
HQ (No B-frames)	144.30 fps	130.72 fps	311.53 fps	336.70 fps

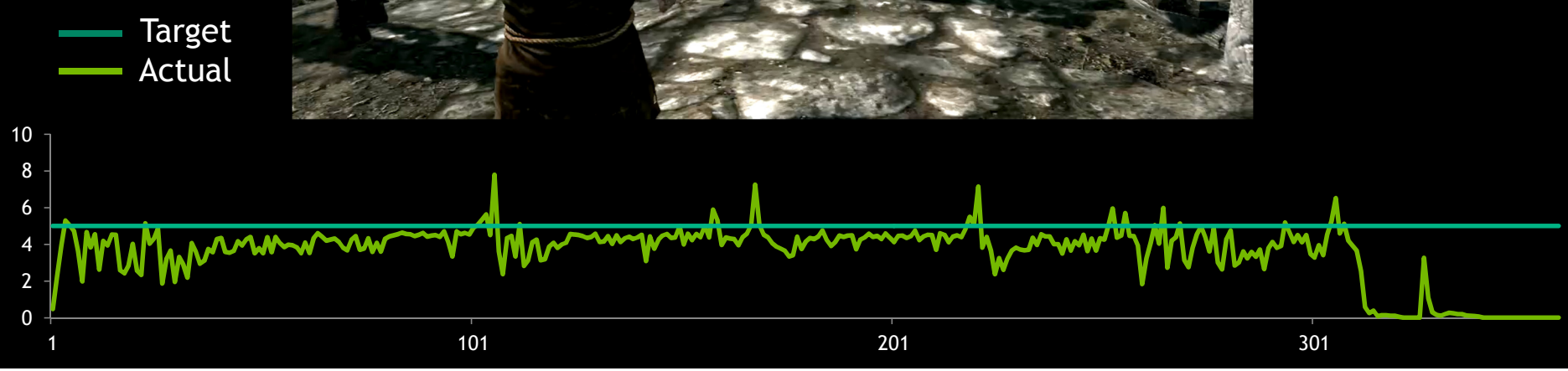
GPU TECHNOLOGY CONFERENCE

Performance (*n* simultaneous HD encodes)



- Encode bit rate = 30 Mbps
- Performance shown with more than 1 context is the sum of fps obtained from each context

Gaming Sequence (720p) - 5 Mbps



Gaming Sequence (720p) - 10 Mbps

GPU TECHNOLOGY CONFERENCE

Questions?