Video Acceleration in the Cloud Using FPGAs

Presented By

Sean Gardner
Marketing Mgr., Cloud Video

Johan Janssen
Chief Video Architect
The market is recognizing us and our technology...

Dan Rayburn is key industry analyst & veteran

Xilinx enabling all live applications
Live streaming seeing explosive growth

Live Streaming Market Size ($B)

2016

2021

21% CAGR

Markets & Markets report
Live video demands a new approach

- Video will be 82% of all internet traffic by 2021 - Cisco VNI report
- Video will be 73% of all wireless traffic by 2023 - Ericsson Mobility report

Graph showing increase in pixels from SD (480) to 8k with 96x more pixels.
New codecs help one hand…

Bandwidth

Codec Complexity

MPEG2
H.264
HEVC / VP9
AV1

1080p60 (mbps)

Codec Complexity

8 x less bandwidth

125 x more complex

8 x less bandwidth

1080p60 (mbps)
CPUs are too slow for live video...

Frames per second using Intel E5 Xeon

- Very Slow
- Slow
- Medium
- Fast
- Very Fast
- Live Video

What they live with for real-time applications

Compression efficiency customers would like

Measured on AWS using dual socket Xeon E5-2666 v3 @ 270W
**Intel performance scaling not keeping up**

**Estimated Performance Gain Over 10 Generations**

*Historical 2-Socket Integer Throughput Performance*

<table>
<thead>
<tr>
<th>Description</th>
<th>Growth in Compute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live video growth</td>
<td>2 x</td>
</tr>
<tr>
<td>Resolution growth</td>
<td>96 x</td>
</tr>
<tr>
<td>Growth in codec complexity</td>
<td>125 x</td>
</tr>
<tr>
<td><strong>Total growth in compute</strong></td>
<td><strong>223 x</strong></td>
</tr>
</tbody>
</table>

Slide taken from Intel XDF presentation.
The “Pareto Principle” of live video distribution

20% of video streams generate 80% of streaming traffic

Less popular streams (80% of streams but 20% of bandwidth)
Together Xilinx & NGCodec can save on streaming costs
The difference compression performance makes

<table>
<thead>
<tr>
<th>Encoded Bitrate</th>
<th>Data per Mth. (TB) Per Stream</th>
<th>Cost per Mth ($0.05 per GB)</th>
<th>Monthly Cost @ 100k Streams</th>
<th>Annual Cost (100k Streams)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4Mbps</td>
<td>1.21</td>
<td>$60.48</td>
<td>$6,048,000</td>
<td>$72,576,000</td>
</tr>
<tr>
<td>2.8Mbps</td>
<td>0.85</td>
<td>$42.34</td>
<td>$4,234,000</td>
<td>$50,808,000</td>
</tr>
<tr>
<td></td>
<td>Annual Savings</td>
<td></td>
<td></td>
<td>$21,768,000</td>
</tr>
</tbody>
</table>
Hardened solutions come at a cost

Encoder saves significant OPEX costs for bandwidth on egress traffic

- 35% lower bitrate for same quality

Measured June 2018 SDK

Bitrate (kbps) vs. PSNR for competitor and NGcodec HEVC PSNR.
Overview of FPGA Video Acceleration

> Johan Janssen
> Chief Video Architect
## FPGA HEVC encode vs. x265 encoding configurations

<table>
<thead>
<tr>
<th>HEVC encode comparison</th>
<th>1080p fps @ x265-slow preset</th>
<th>Device Power</th>
<th>performance/W</th>
<th>performance/W Improvement</th>
<th>Available</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual Socket E5-2680 v3 2.5GHz</td>
<td>10 fps</td>
<td>240 W</td>
<td>0.042 fps/W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single socket VU9p</td>
<td>120 fps</td>
<td>&lt;40 W</td>
<td>3.0 fps/W</td>
<td>72x</td>
<td>Dec 2018</td>
</tr>
<tr>
<td>Dual socket VU9p</td>
<td>240 fps</td>
<td>&lt;80 W</td>
<td>3.0 fps/W</td>
<td>72x</td>
<td>Dec 2019</td>
</tr>
</tbody>
</table>

Commonly used X265 preset for encoder benchmarking

Dec 2018

Dual socket VU9p 240 fps <80 W 3.0 fps/W 72x Dec 2019

© Copyright 2018 Xilinx
FFmpeg is the leading multimedia framework, able to decode, encode, transcode, mux, demux, stream, filter and play pretty much anything that humans and machines have created

- It supports the most obscure ancient formats up to the cutting edge
- Highly portable: FFmpeg compiles, runs, across Linux, Mac OS X, Microsoft Windows, under a wide variety of build environments, machine architectures, and configurations
Integration into FFmpeg framework: building the ecosystem

Customer Application

FFmpeg
(video codecs, Scalars, Compositing etc.)

XMA
(XRT, Partial Reconfiguration, Video IP)

SDAccel Target Board

FPGA

Customer Application

FPGA h.264 encode plugin
FPGA HEVC encode plugin
FPGA VP9 encode plugin
Xilinx h.264 decode plugin
Xilinx ABR Scalar plugin
Xilinx Yolo plugin

x86 Server

FPGA

FPGA HEVC

FPGA VP9 encode

Xilinx h.264 decode plugi

Xilinx ABR Scalar plugi

Xilinx Yolo plugin

FPGA h.264 encode plugin

FPGA h.264 encode plugin

FPGA VP9 encode

Xilinx h.264 decode plugi

Xilinx ABR Scalar plugi

Xilinx Yolo plugin

FPGA VP9 encode
Video Transcoding ABR Example (Single VU9p)

RTSP/RTMP Live Stream

RTSP Client → Demux Aud/Vid → Decode → ABR Scaler → Encoder (low frame rate) → Mux Aud/Vid

Optional

RSTP/RTMP Stream to Inference

RTSP/RTMP to Media Server

Xilinx Alveo PCIe Card

Decode → ABR Scaler → Encoder

2 x HEVC or VP9 1080p60 Encoder
Video Transcoding ABR Example (Single VU9p)

RTSP/RTMP Live Stream → RTSP Client → Demux Aud/Vid → Decode → ABR Scaler → FFmpeg

Optionally:
- Encoder (low frame rate) → Mux Aud/Vid

Optional:
- RSTP/RTMP Stream to Inference
- RTSP/RTMP to Media Server

Xilinx Alveo PCIe Card:
- Decode
- ABR Scaler
- 10 x H.264 1080p60 Encoder
“Edge-to-Cloud Video Analytics”

Video Inference

- Database (mysql/postgresql)
- HTTP Server

Streaming Server

- File
- RTSP/RTMP Live Stream
- Demux Aud/Vid
- h.264 Decode
- Inference
- Post-Process
- Overlay
- Encode

Optional

- RTSP/RTMP to Media Server

Xilinx Alveo PCIe card

- h.264
- xDNN

h.264 or HEVC or VP9

Metadata

RTSP/RTMP Client

File

Database

HTTP Server

cloud
Integrating accelerators into FFmpeg framework

```bash
ffmpeg \
-f rawvideo -pix_fmt yuv420p -s:v 1920x1080 -r 30 -an -i \
/home/ffmpeg/VU9P/TestSequences/Kimono1_1920x1080_24.yuv \ 
-frames 240 -c:v libx264 -preset medium -profile:v high -crf 23 -bf 4 -refs 3 -g 30 -b:v 4000k -maxrate 4000k -bufsize 8000k -f h264 -r 30 -y ./sw_outdir/x264_medium_out0_br4000k.h264
```

```bash
$ ffmpeg \
-f rawvideo -pix_fmt yuv420p -s:v 1920x1080 -r 30 -an -i \
/home/ffmpeg/VU9P/TestSequences/Kimono1_1920x1080_24.yuv \ 
-frames 240 -b:v 4000k -g 30 -c:v xlnx_h264_enc-hq -f h264 -y ./hw_outdir/out0_br4000k.h264
```

```bash
$ ffmpeg \
-f rawvideo -pix_fmt yuv420p -s:v 1920x1080 -r 30 -an -i \
/home/ffmpeg/VU9P/TestSequences/Kimono1_1920x1080_24.yuv \ 
-frames 240 -b:v 4000k -g 30 -c:v xlnx_HEVC_enc -f h265 -y ./hw_outdir/out1_br4000k.h264
```

Change 20 characters to get acceleration
XMA Architecture

Key features

- Video domain specific interfaces with seamless integration with FFmpeg
- Low-level plugin can be reused with any media framework
- Supports multiple processes sharing different kernels on the same device
- Supports multiple channels on a single kernel
- Ensures a kernel resource is reserved for the lifetime of a video session
## Video IP Offering (each IP has FFmpeg plugin)

<table>
<thead>
<tr>
<th>Codec</th>
<th>Partner</th>
<th>Description</th>
<th># of 1080p (VU9P)</th>
<th>Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>H.264 HDE</td>
<td>Alma</td>
<td>High Density Encoder</td>
<td>12 x 1080p60</td>
<td>Today</td>
</tr>
<tr>
<td>H.264 HQE</td>
<td>IDT</td>
<td>High Quality Encoder</td>
<td>1 x 1080p60 (3 in Mar)</td>
<td>Today</td>
</tr>
<tr>
<td>H.264 HDD</td>
<td>VYU Synch</td>
<td>High Density Decoder</td>
<td>12 x 1080p60</td>
<td>Today</td>
</tr>
<tr>
<td>HEVC-HDE</td>
<td>NGCodec</td>
<td>High Density Encoder</td>
<td>2 x 1080p60</td>
<td>Today</td>
</tr>
<tr>
<td>HEVC-HQE</td>
<td>NGCodec</td>
<td>High Quality Encoder</td>
<td>1 x 1080p60 (roadmap 2)</td>
<td>Today</td>
</tr>
<tr>
<td>HEVC-HDD</td>
<td>Path Partner</td>
<td>High Density Decoder (Up to 8 12bits 4:4:4)</td>
<td>6 x 4kp30</td>
<td>Today</td>
</tr>
<tr>
<td>HEVC-HEIFD</td>
<td>Path Partner</td>
<td>HEIF Decoder (Up to 60Mbps per core)</td>
<td>10 x 4kp15</td>
<td>Today</td>
</tr>
<tr>
<td>VP9-HQE</td>
<td>NGCodec</td>
<td>High Quality Encoder</td>
<td>1 x 1080p60 (roadmap 2)</td>
<td>Q2 CY2018</td>
</tr>
<tr>
<td>AV1</td>
<td>NGCodec/Xilinx</td>
<td>Future</td>
<td></td>
<td>2019</td>
</tr>
<tr>
<td>Perseus+</td>
<td>V-Nova</td>
<td>4x density / bitrate improvement</td>
<td>Various</td>
<td>Q2 CY2018</td>
</tr>
<tr>
<td>WebP Enc.</td>
<td>Xilinx</td>
<td>High Density Encoder</td>
<td>Resolution dep.</td>
<td>Today</td>
</tr>
<tr>
<td>ABR Scalar</td>
<td>Xilinx</td>
<td>High Density Scalar</td>
<td></td>
<td>Today</td>
</tr>
</tbody>
</table>
Xilinx Alveo ABR video transcoding solution

Xilinx has a solution to address all workload needs
## Video Optimized platforms

<table>
<thead>
<tr>
<th>Form Factor</th>
<th>Details</th>
<th>Sampling Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Profile</td>
<td>Power 75W</td>
<td>Aug 2018</td>
</tr>
<tr>
<td></td>
<td>Suitable for any server form factor</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Low power with passive cooling</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lowest cost option</td>
<td></td>
</tr>
<tr>
<td></td>
<td>24 x 1080p30 H.264 encodes in real-time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>24 x 1080p30 H.264 decodes in real-time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 x 1080p60 HEVC encodes in real-time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10 x 1080p60 HEVC decodes in real-time</td>
<td></td>
</tr>
<tr>
<td>Form Factor</td>
<td>Details</td>
<td>Oct 2018</td>
</tr>
<tr>
<td>FH/ Length</td>
<td>Power 225W</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Highest density per slot</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ability to mix workloads Video + ML / AI</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 x performance of single VU9P</td>
<td></td>
</tr>
</tbody>
</table>