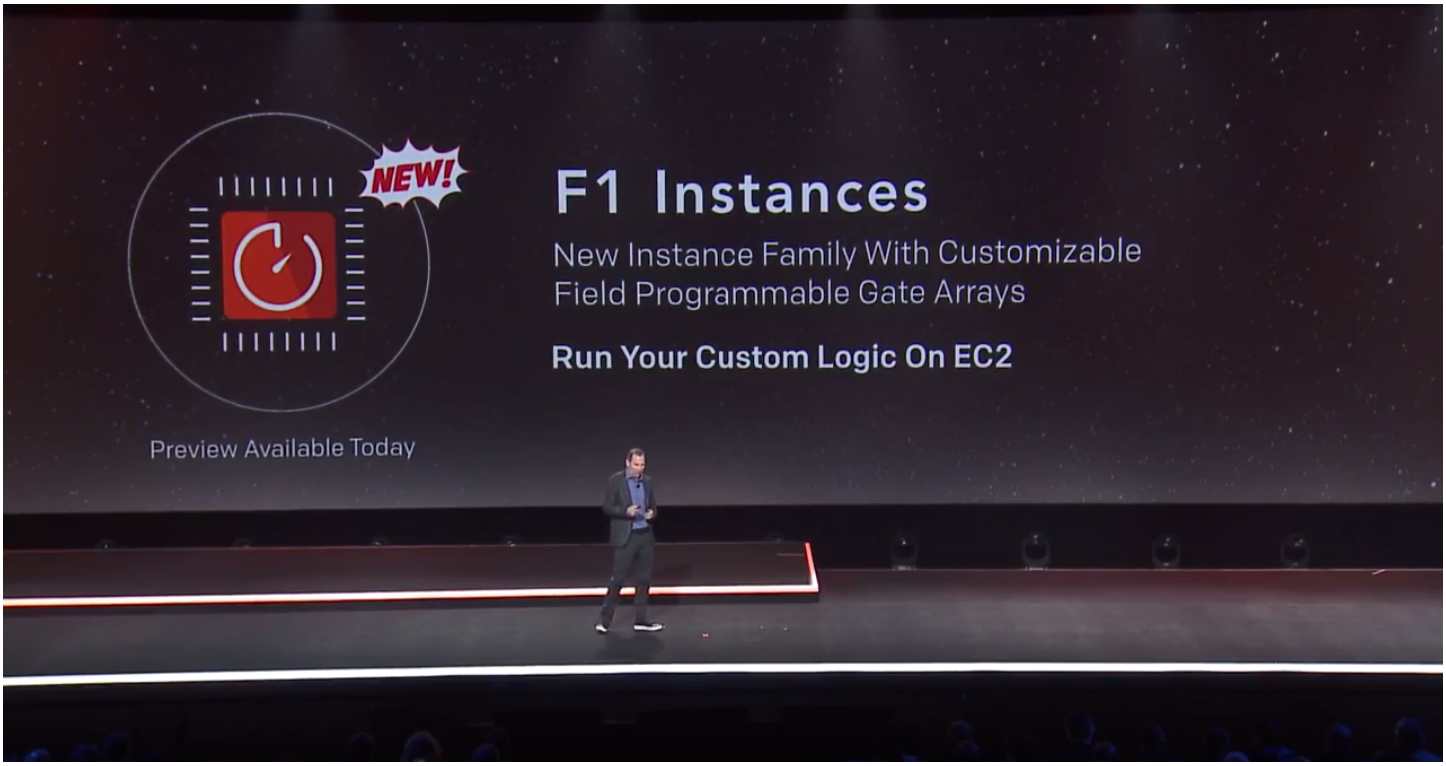
Live Video Encoding Using New AWS F1 Acceleration

The Benefits of Xilinx FPGA for Live Video Encoding



VERSION 1.0

Author: Oliver Gunasekara

NGCodec Inc.
440 North Wolfe Road
Sunnyvale, CA 94085–3869
USA

https://ngcodec.com
info@ngcodec.com
+1 408 766 4382

ABSTRACT

In today's mobile world, where live video is rapidly gaining ubiquity in everyday life, NGCodec is leading the charge to overcome the difficulties and sacrifices in quality associated with video encoding using traditional software methods. This white paper discusses the benefits of hardware encoding using Xilinx® FPGA in the new Amazon Web Services (AWS) F1 instances. We open with a background on video encoding and an overview of the encoding process. This is further contextualized with a discussion of the applications of cloud video transcoding and an exploration of the differences between file-based and live video encoding. Following on from this, we explore the limitations of traditional software encoding methods for live video encoding. Having established the drawbacks of relying on CPUs and GPUs, we discuss the superior results that can be obtained through hardware encoding with AWS FPGA F1 instances. Our paper goes on to delve into the methodology behind NGCodec's FPGA F1 design using the Xilinx Vivado® HLS tool suite and to summarize how we ported our RealityCodec™ H.265/HEVC video encoder to AWS Elastic Compute Cloud (EC2) F1 instances in only three weeks. Finally, the paper outlines our roadmap for a new, twofold business model to make hardware encoding with FPGA F1 instances available to customers of all sizes and closes with an opportunity for readers to try out NGCodec's video encoding capabilities for themselves.

TABLE OF CONTENTS

On the cover: Amazon Web Services CEO Andrew Jassy announces the new EC2 F1 Instances at AWS Re:Invent 2016 in Las Vegas, Nevada.

I. VIDEO ENCODING IN TODAY'S MOBILE WORLD

At its most basic, video encoding involves converting a raw source video into a much smaller bit rate that can be streamed to an internet player, mobile device, digital set-top box, or similar, as shown below in Fig. 1. Video transcoding converts one format of compressed video to another by decoding the source video and subsequently re-encoding. Video encoding uses modern compression techniques to reduce redundancy in video data by leveraging both spatial (inside the frame) image compression and temporal (across multiple frames) motion compensation.

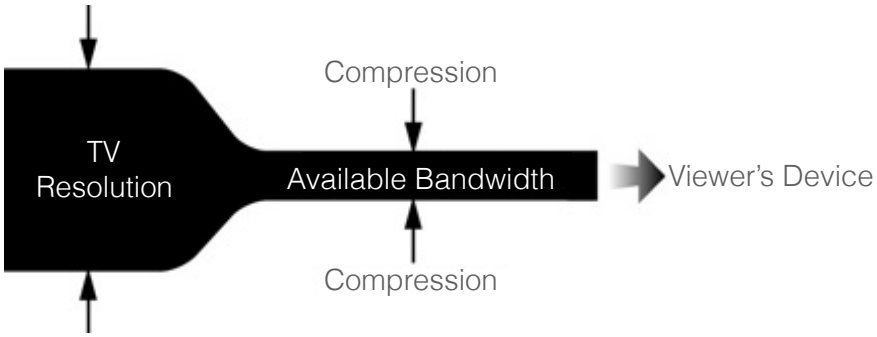


FIG. 1: VIDEO ENCODING PROCESS OVERVIEW

Different international video standards exist to define how to decode video for playback. The latest standard, H.265/HEVC, is the basis for NGCodec's RealityCodec video encoder technology. Our implementation delivers a video signal that is a mere 0.13 percent the size of the original raw video (compression ratio of 750:1). It is important to understand that each implementation of the H.265/HEVC encoder is proprietary and makes trade offs between computing requirements, bit rate, and video quality.

There are two major applications for cloud video encoding today. The first, file-based video encoding, is currently the most prevalent. Any video you watch online was likely encoded in a data center and stored, ready to stream to consumers via a service such as YouTube®. The second, live video transcoding, is growing rapidly, driven primarily by mobile technology and social media. As consumers grow increasingly accustomed to being able to consume content on their mobile devices while on the go and stay connected to family and friends wherever they may be, the ability to incorporate live video into these interactions becomes ever more desirable. A prime example of live video transcoding in use today is Facebook® Live, which allows a user to engage in a one-to-many broadcast of a live video to his/her Facebook friends, any of whom can view the live feed and interact by commenting or leaving a reaction, such as a "Like," in real time. Another example is watching a live sports match, concert, or news broadcast, which is streamed to many users simultaneously over the internet via a service such as YouTube TV or the Comcast® XFINITY® Stream App.

The encoding process for both file-based and live video is essentially the same: the source video must be encoded multiple times at varying resolutions, bit rates, and (sometimes) different compression standards. The video is typically broken into two-second segments, such that the end client has the capability to dynamically evaluate, at corresponding intervals, the available connection speed and bandwidth and select from the various quality levels accordingly. This is known as adaptive bit rate, or ABR. Apple HLS (HTTP Live Streaming) and MPEG-DASH (Dynamic Adaptive Streaming over HTTP) are popular examples. The viewer may be able to perceive changes in quality throughout the playback experience, but the hope is that the video will merely improve or deteriorate in quality, rather than freeze altogether. This has rendered obsolete store-and-buffer playback styles, in which playback interruptions can lead to high levels of video abandonment. Video abandonment is of particular detriment to video streaming services whose revenue is advertisement-based: when the viewer stops watching, he/she also misses the placed ads from sponsors of the site.

II. LIVE VIDEO BEFORE F1: A STREAM OF SACRIFICES

In a live video broadcast over the internet, a single video stream is sent from the source to the cloud. It is then transcoded—decoded in the cloud and re-encoded into multiple bit rates for ABR—before being sent on to the end viewer. Today, this is achieved purely through software, typically open source encoding projects such as x264 or x265, using many central processing units (CPUs). The difficulty with this approach for live video is that there is a limit to the amount of parallelism that can be exploited to make the video smaller; this capability is defined by the number of cores within the server in question. Because the frames per second (FPS) must be maintained to avoid jerky playback, the computing requirement must not exceed this FPS at any time. As such, the highest quality settings in the software encoder cannot be used. For our purposes, we will look at the x265 open source software video encoder as an example.

Encoding software like x265 contains a great many presets, allowing the user to customize settings and trade overall computing requirements for the end size of the video. For file-based videos, this technology can produce very high-quality results with the x265 'veryslow' preset: the video can be encoded many times longer than real-time constraints allow, yielding the best compression, but with considerable cost of encoding resources.

For live video, by contrast, software encoding is simply unable to achieve the maximum quality offered by the encoder technology. Fig. 2 compares 1080p50 source video encoded with different x265 presets (for video quality) and the resulting frames that can be encoded per second on the AWS c4.8xlarge instance type. The necessary tradeoffs to satisfy computing requirements mean significant reductions in quality. Instead of running the encoder at a slow setting, which will produce the best end quality, sacrifices are necessary to achieve the target frame rate. The fundamental problem with software-based encoding for live videos is that the best compression—that is, the highest quality video for the lowest bit rate—and the finest end result in video quality are unattainable with the available compute level. By comparison, NGCodec's encoder can achieve 80 FPS and surpass the quality of even the x265 'veryslow' preset.

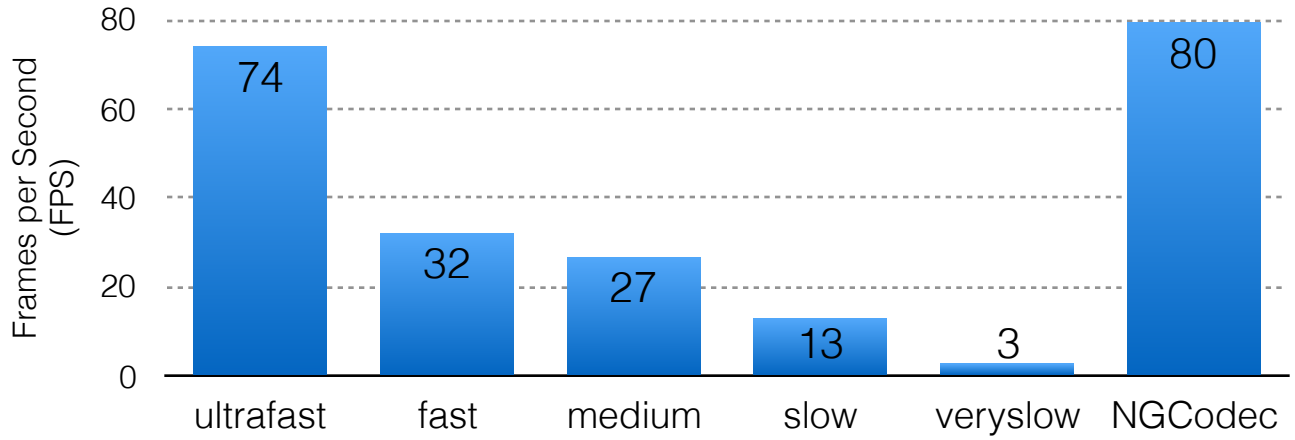


FIG. 2: x265 SOFTWARE ENCODER PRESETS & CORRESPONDING FPS

III. FPGA ACCELERATION

With the slowing of Moore's Law, cloud infrastructure providers have realized that a new type of accelerator is required. Initially, graphics processing units (GPUs) were tried. For some applications, like the training of neural networks, these provide more performance than CPUs, but they are more challenging to program. Unfortunately, many new applications require complex decision-making as well as number crunching; these have experienced limited gains with GPUs.

The FPGA (field-programmable gate array) is a special type of chip that can be programmed by rewriting its circularity in the field. This is identical to how a traditional ASIC (application-specific integrated circuit) is designed, but unlike FPGAs, ASICs can not be changed in the field and typically take years to develop from the initial design to deployment. Fig. 3 shows a comparison of different accelerators used in data centers.

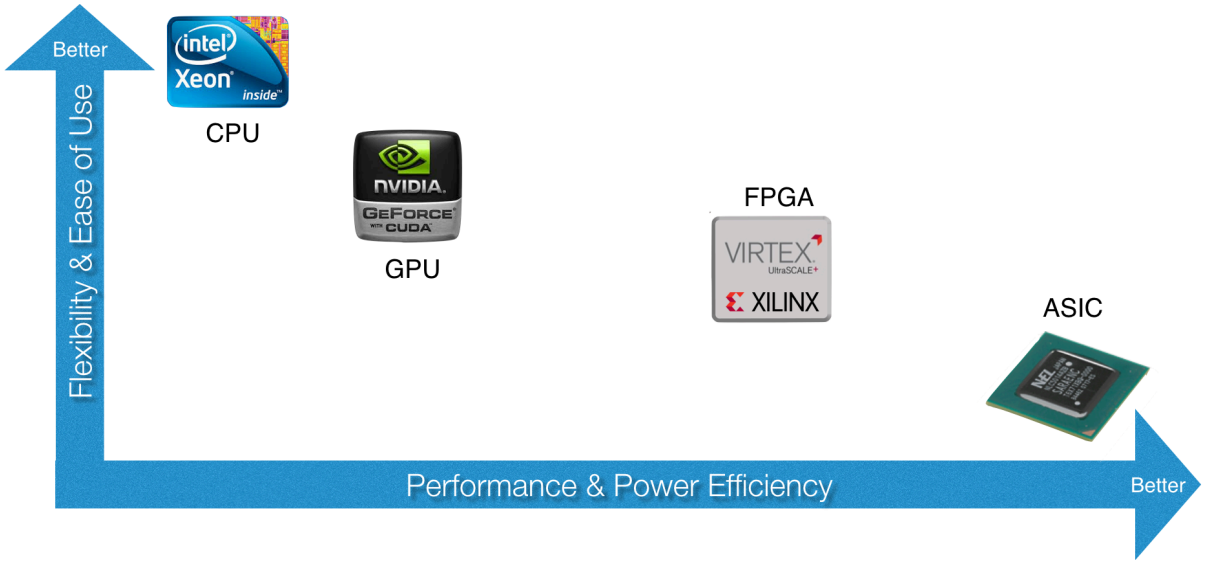


FIG. 3: COMPARISON OF CHIP TYPES & CAPABILITIES

Due to how FPGAs are programmed versus running software on CPUs and GPUs, FPGAs offer twenty times more performance and are around twenty times lower power than CPUs. The drawback is that they are more difficult to program and use.

IV. METHODOLOGY BEHIND OUR FPGA F1 DESIGN

The H.264 and H.265 standards define only the *decoding* of a video, not the encoding; this presents a significant opportunity for companies such as NGCodec to differentiate ourselves from our competition and offer desirable benefits and capabilities to customers through our proprietary video encoder algorithms. The output from our RealityCodec encoder is fully standards-compliant, meaning that it can be decoded and played on any device that supports H.265/HEVC. We have been granted one US patent and have seven others pending for our encoding technology.

Developing video encoders involves a group effort from a variety of teams within NGCodec. Initially, our algorithm team develops a bit accurate software model in C/C++, in which they invent our encoding algorithms. This is called 'N265'. This command line utility runs on a PC, taking in raw video and outputting a compressed H.265/HEVC video file. It runs very slowly—about one frame per minute—as it models the video running through the entire hardware pipeline.

In the second stage of development, our FPGA implementation team rewrites the algorithmic model into high-level synthesis (HLS) to achieve something that can be compiled into register-transfer level (RTL). This involves a significant investment of time and effort, yielding a result that ultimately is turned into Verilog RTL, which is then place-and-routed to form a bitstream for the FPGA. We use the Xilinx Vivado® HLS tool suite for compiling our C++ into RTL and have consistently achieved sound results with it. The benefits of using HLS include accelerating our design time and having the ability to iterate on that design much more quickly, with a high level of verification.

Last, our firmware team configures the encoder correctly and adds in the rate control, which is a critical part of any encoder design. This ensures, for example, that if 1 Mbps is requested, then 1 Mbps is ultimately delivered, even in videos featuring high levels of motion or scene changes. An outline of NGCodec's development flow is depicted in Fig. 4.

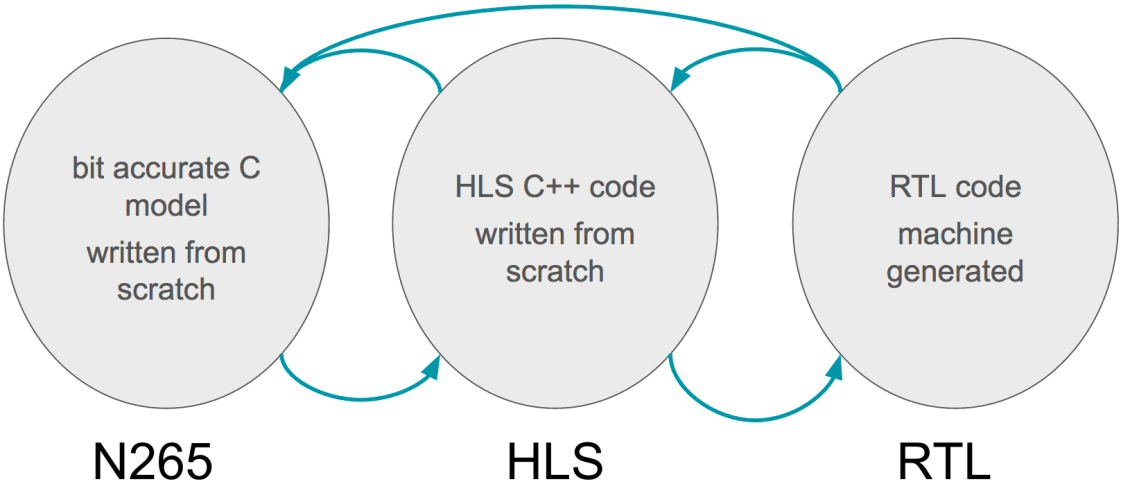


FIG. 4: NGCODEC DEVELOPMENT FLOW

V. PORTING TO F1 FOR AMAZON WEB SERVICES

NGCodec's team of fifteen spent two years developing our RealityCodec H.265/HEVC video encoder. Capitalizing on this substantial investment of time and engineering effort, we were able to port our technology to the AWS Elastic Compute Cloud (EC2) F1 Instances on a highly accelerated schedule. In just three weeks, we completed porting and met the deadline for a live demo at the AWS re:Invent 2016 conference in Las Vegas, Nevada, where we showcased our ability to live-transcode H.264 to H.265 and halve the bit rate while maintaining the same video quality.

On stage at re:Invent, we streamed HD video in H.264 from an iPhone to the AWS EC2 F1 Instance in the 'us-east-1' data center. In the AWS data center, we decoded the H.264 video and sent it to our RealityCodec running on the Xilinx FPGA, where it was re-encoded in H.265. We then sent the H.265 video back to a laptop on stage in Las Vegas, as shown in Fig. 5. The laptop on stage used a software decoder to display the video stream in real time.

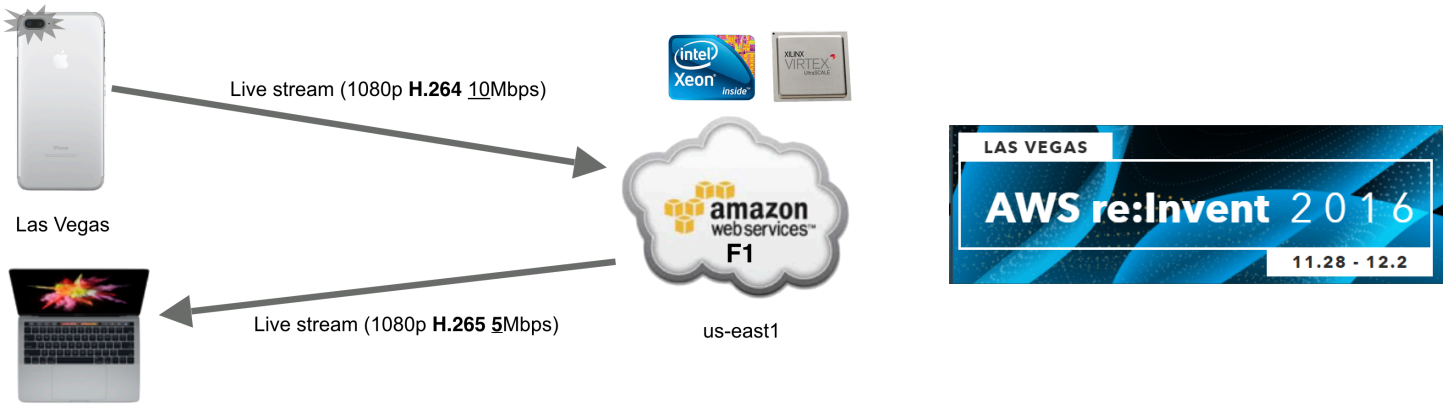


FIG. 5: NGCODEC DEMO AT AWS RE:INVENT 2016

Delivering the AWS re:Invent demo was an amazing achievement by three different teams working together:

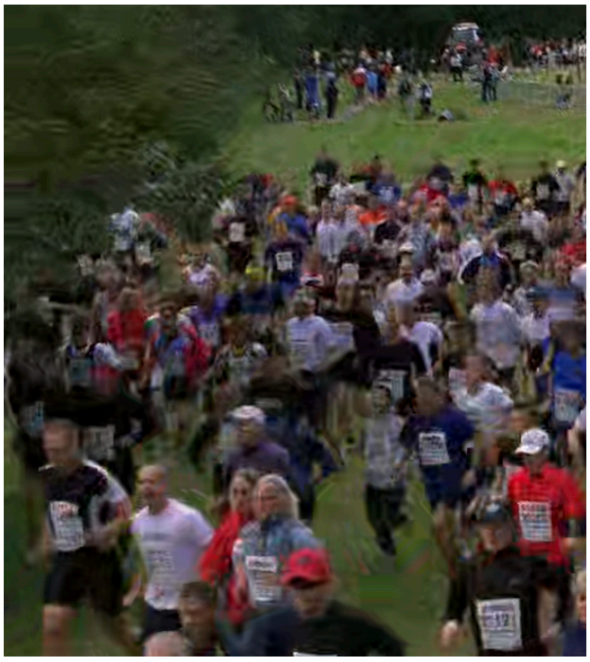
- **AWS F1 Infrastructure Team:** Primarily located in Austin, Texas, they provided the hardware development kit (HDK) and supported the NGCodec team throughout.
- **NGCodec Team:** We ported our H.265/HEVC RealityCodec to the AWS F1 Instance and created a plugin for FFmpeg.
- **Cogniance Professional Services Team:** Hired by NGCodec to build the supporting cloud software infrastructure for the live transcoding demo, they handled initial prototyping with x265 software encoding.

VI. BENEFITS OF HARDWARE ENCODING WITH F1 INSTANCES

For live video, the primary benefit to video encoding with FPGA F1 instances is that we can achieve a higher quality video at the same bit rate, and do it at a desirable 60 frames per second. A second benefit, relevant only in certain cases, is lower latency and reduced lag time between live stream source and end viewing. Third, the cost of encoding is significantly reduced. AWS's pricing for the FPGA F1 instance is forthcoming; we hope to present a more economical option to users as a result of the significantly lower power consumption required for FPGAs versus CPUs. Finally, we can support up to 32 independent encoded streams of video on a single F1 instance.

In a practical sense, the gain is ultimately that NGCodec can enable customers to achieve a higher-quality video by taking advantage of the greater compute capability of an FPGA. We are able to reduce source video to 0.13 percent of its original value with virtually no perceived loss of quality. Fig. 6 shows the difference in video quality between two x265 presets. Note the improved clarity of the runners' bib numbers and the tree on the top left.

x265 1080p50 2.5Mbps 'UltraFast'



x265 1080p50 2.5Mbps 'VerySlow'

FIG. 6: COMPARISON OF X265 PRESETS & RESULTING VIDEO QUALITIES

VII. BUILDING A BETTER BUSINESS MODEL

The strides that NGCodec has made in enabling FPGA F1 instances will allow us to explore new business models in areas such as cloud VR. Using the hardware acceleration capabilities of an FPGA, we can not only increase the quality of the video experience, but also decrease the latency significantly. This is not something that can be achieved through software encoding.

Our plan is twofold. First, we aim to make our technology available as part of the AWS Marketplace, where customers can buy F1 instances per use on an hourly basis and pair them with purchased code that runs in tandem. This will allow smaller customers to evaluate our technology without necessitating a potentially cost-prohibitive investment. Second, for larger customers, we hope to introduce a subscription model in which a monthly fee will allow the user to encode as much video as desired.

Across all our business models, NGCodec remains passionately focused on video encoding and visual quality. We hope to partner with companies specializing in other niche components of the overall video experience, such as audio and closed captioning, to enable a best-of-the-best video product and viewing experience.

VIII. LET US SHOW YOU

Hardware encoding using FPGA F1 instances offers many benefits to engineers encoding live video, including:

- exceptional compression—a video signal that is just 0.13 percent the size of the original raw video;
- lower bit rate while maintaining the same picture quality;
- reduced latency;
- significantly lower power consumption compared to CPUs;
- lower cost; and
- an improved viewing experience over traditional software encoding methods.

There is no need to take our word for it—we invite you to experience NGCodec-enabled video quality for yourself. Send us your raw video and we will encode it and return it to you so that you can see firsthand how video encoding using our encoder can be a game-changer for your company.

NGCodec's expert engineers are always available to answer your questions at +1 408 766 4382 or info@ngcodec.com. We look forward to partnering with you to make your next project a success. To learn more about NGCodec and our vision for next generation video compression, please visit our website at https://ngcodec.com.

REFERENCES & SUPPLEMENTAL INFORMATION

"Amazon EC2 F1 Instances (Preview): Run Custom FPGAs in the AWS Cloud." *Amazon Web Services.* Accessed March 27, 2017. https://aws.amazon.com/ec2/instance-types/f1/

Bossen, Frank, Benjamin Bross, Karsten Suhring, and David Flynn. "HEVC Complexity and Implementation Analysis." *IEEE Transactions on Circuits and Systems for Video Technology* 22, no. 12 (Dec. 2012): 1685–1696.

Gunasekara, Oliver. "Delivering a Moonshot: Being the First Company to Use the New Amazon F1 Instance." *NGCodec* (blog). December 2, 2016. https://ngcodec.com/news/2016/12/2/delivering-a-moonshot-being-the-first-company-to-use-the-new-amazon-f1-instance.

Gunasekara, Oliver. "NGCodec H.265/HEVC Video Compression." *Powered by Xilinx.* YouTube video, 3:45. Posted by "XilinxInc," January 18, 2017. https://www.youtube.com/watch?v=YIWXj8yhAzQ.

"H.265 : High efficiency video coding." *ITU: International Telecommunications Union.* Last modified December 22, 2016. http://www.itu.int/rec/T-REC-H.265.

"High Efficiency Video Coding." *Wikipedia.* Last modified March 21, 2017. https://en.wikipedia.org/wiki/High_Efficiency_Video_Coding.

Richardson, Iain. "Vcodex: Introduction to Video Coding." *Vcodex.* YouTube video, 11:47. Posted by "vcodexer," June 26, 2013. https://www.youtube.com/watch?v=gxefuXizO04.

Sole, Joel, Rajan Joshi, Nguyen Nguyen, Tianying Ji, Marta Karczewicz, Gordon Clare, Félix Henry, and Alberto Dueñas. "Transform Coefficient Coding in HEVC." *IEEE Transactions on Circuits and Systems for Video Technology* 22, no. 12 (Dec. 2012): 1765–1777.

ABOUT NGCODEC

NGCodec® has been in passionate pursuit of next generation video compression since 2012. With the support of investors including Xilinx and NSF, NGCodec's agile startup team has created RealityCodec™, a compressor-decompressor technology optimized for ultra-low latency, high-quality applications. NGCodec is headquartered in Sunnyvale, California. To learn more, visit https://ngcodec.com.