WHITE PAPER

FPGA

Finding Acceleration in a Video-Centric World

Intel® FPGAs and H.265 intellectual property enable edge computing at the lowest TCO.

Author

Neal Forse
Video and Vision Strategic Marketing Manager
Intel Programmable Solutions Group

Table of Contents

Introduction

By 2019, 80% of global Internet usage will be for video content, and annual global IP traffic is projected to reach 2.3 zettabytes (ZB) per year by 2020[1] leading to a projected 270% growth in video. Storing, encoding, queuing, transmitting, and decoding this huge volume of video requires new systems and technologies to keep pace with the demand. Today, providers rely on custom hardware and consolidated data centers to manage network content. Users expect on-demand content to arrive when they want it (immediately)—which is forcing providers to shift their distribution model.

The following elements are key to the success of this new model:

- Complex video algorithms require acceleration for timely processing. FPGAs are the most server footprint efficient programmable (non-hardened) accelerators available for media workloads.
- As the industry progresses towards the Future-X network,[2] media-centric servers do not scale for capital expenditures (CapEx) or operating expenses (OpEx).
- Workload agnostic servers, enabled with FPGAs, can enable a non-hardware-segmented data center infrastructure, lowering the total cost of ownership.

Accelerating media workloads

Video is recorded, compressed, and distributed using a variety of video formats and algorithms. Intel® Xeon® processors (such as the Xeon-SP) are computationally powerful enough[3] to host media workloads for legacy video formats without hardened accelerators for video processing. For example, today's media infrastructure vendors successfully use the Xeon-SP for:

- MPEG-2 and advanced video coding (AVC) codecs (H.264)
- Color space conversion
- Resolution scaling
- Frame rate conversion
- Noise filtering

However, as video algorithms become more complex—and the computational demands to perform these algorithms increase—accelerating the x86 hardware is critical. For example, high efficiency video coding (HEVC) or H.265, a next-generation codec, can be an order of magnitude more computationally complex than AVC.[4] Therefore, implementing H.265 algorithms in a pure x86 domain is

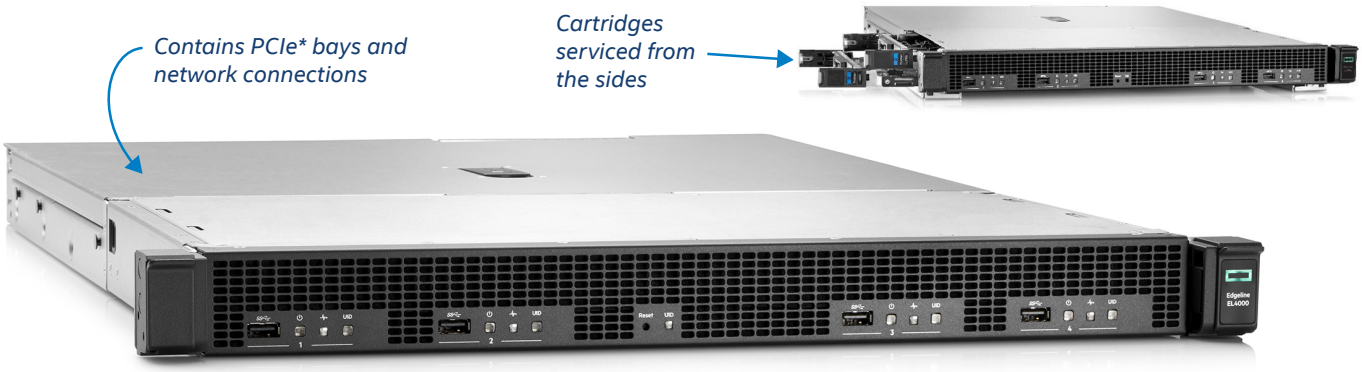

Figure 1. DCP Solution: HPE Edgeline EL4000

unrealistic. Additionally, it is challenging to execute a single workload across multi-socket hardware platforms.[5]

To support next-generation video requirements, Intel has developed video acceleration algorithms targeted for Intel FPGAs. FPGAs are functionally agnostic, allowing the developer to configure them with customized fuctions. Intel has partnered with eBrisk to provide a solution that combines an Intel Xeon processor with FPGA acceleration on an Intel Arria® 10 FPGA; the resulting HEVC encoding solution has a rich feature set with high video quality.[6] HEVC is computationally intense, requiring a heavy video processing workload. Implementing the HEVC encoder on an FPGA doubles throughput while increasing power consumption by less than 20%.

Intel's Discrete Configurable Platform (DCP) supports flexible, configurable cloud-based applications. For example, the HPE Edgeline EL4000 Converged Edge System—featuring a Xeon + Intel Arria 10 DCP—brings data management, video analytics, and security to the cloud's edge.[7] Additionally, the EL4000 supports the eBrisk eBLive A-5000 HEVC encoder, enabling high-quality broadcast video distribution.[6]

To provide an even higher level of flexibility in the cloud, Intel is combining the Xeon processor's power with flexible Intel Arria 10 or Intel Stratix® 10 FPGAs in a Multi-Chip Package (MCP). This MCP can fit multiple instances of a 4Kp60 10 bit HEVC encoder design—the industry's highest-order codec at the highest frame rate, bit depth and resolution (4K)—in a single Xeon motherboard socket, which is the industry's smallest footprint server socket. Figure 2 provides a video support roadmap.

Migrating to a new infrastructure model

Globally, IP video traffic will be 82% of all consumer Internet traffic by 2020, up from 70% in 2015.[1] This anticipated jump in video demand will require a new distribution paradigm. Today, video processing is centralized; existing cloud architectures experience more than 100 ms latency between the end consumer and the central media processing hub. This lag time between the hub and end user is unacceptable for a variety of applications. For example, the human visual cortex maps to human spatial awareness within 7 ms.[2], [8] Therefore, in virtual reality (VR) cloud gaming[9] use cases, the latency must be around 10 ms or less between video frames that contain a delta in the user's point of reference (and are subsequently presented to the human eye). To achieve this low latency, the lag time must be reduced by a factor of 10.

What is left to scale?

In *The Future-X Network: A Bell Labs Perspective*, Marcus Weldon explains that transmission systems consist of spectrum (Hz), spectral efficiency (bits/second/Hz), and spatial dimensions (bits/second/Hz/m²). Figure 3 shows how these elements work together in a "Trangle of Truth." As Weldon states,

> "[W]e have nearly reached the fundamental limit of capacity." [2]

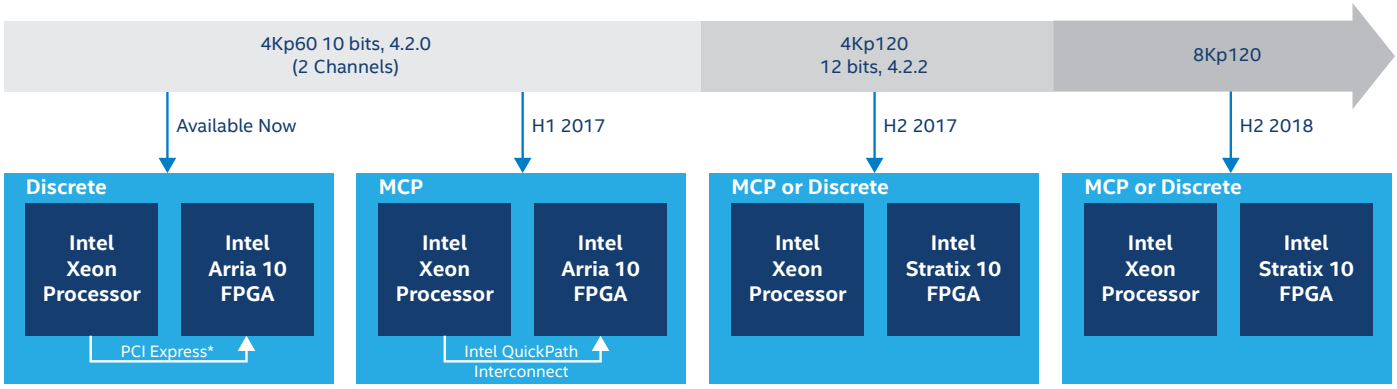

Figure 2. Video Support Roadmap

Projected roadmap—subject to change

- Operators cannot scale the Hz dimension more than 2X over what is used today due to a simple physical constraint: the RF power is too prohibitive to propagate these higher frequencies over the distances required.
- Due to advances in redundancy encoding, the spectral efficiency or bits/sec/Hz is already very near the Shannon limit, which is the theoretical maximum efficiency for a given transmission medium.
- Space, the remaining dimension, is only limited by the area of the Earth. By geographically dispersing the computing (cloud) infrastructure, operators can scale bandwidth beyond 2X moving forward. As part of this dispersion, video processing hardware must be located at the network edge.

Pushing to the network edge

The speed of light has 4.5 ms latency for every 1,000 km.[2] Signal transmissions contain additional latency based on the number of hops the signal takes to get from its source to destination as well as the time it takes to encode and decode the signal. Figure 4 compares Ethernet and Internet latency to the speed of light.

To reduce the overall transmission latency significantly, video processing must move from the centralized cloud to the network's edge. Additionally, to achieve 10 ms latency, this new video edge must be around 100 km from the end user. Dispersed video processing supports applications such as:

- Just in time transcoding (JITT) for video on demand (VoD)
- Live event streaming, for example, sports events

As illustrated by the Triangle of Truth, the bits/second/Hz/m² or space dimension is the only way to increase user bandwidth.[10] Pushing video computing to a non-centralized, geographically distributed network is inevitable.

Cloud services are beginning to move towards the edge, blurring the lines between the cloud and connected devices. For example, Amazon Web Services (AWS) Greengrass program "extends AWS cloud capabilities to local devices, thereby enabling them to collect and analyze data closer to the source of information, while also allowing devices to securely communicate with each other on local networks."[11]

Lowering total cost of ownership

Today, content distributors use custom hardware for each video processing step, such as middleware, and encryption, ad insertion, as well as JITT VoD, live event streaming, and virtual reality. Each function is mapped to a specific workload server, and video encoding and decoding is the most computationally intensive part of the flow.

Commercially distributed content, enabled by media workloads in the data center or cloud, typically requires 99.999% uptime, also called five-nines.[12] Additionally,

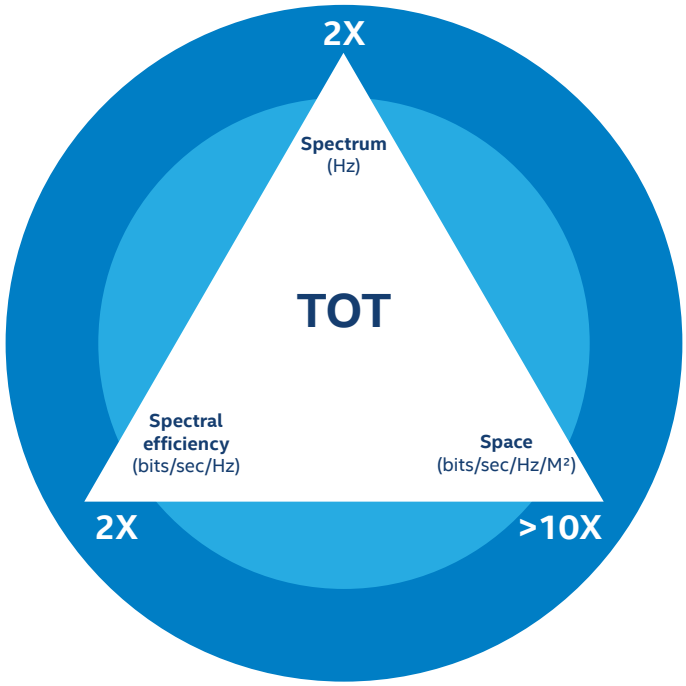

Figure 3. Weldon's Triangle of Truth[2]

Reprinted, by permission, from Marcus K. Weldon, The Future X Network: A Bell Labs Perspective. (c) 2016 by Taylor & Francis, LLC.

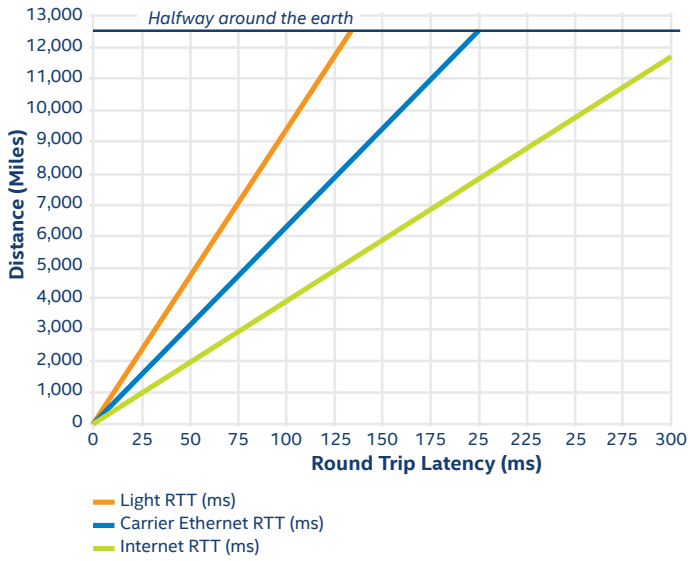

Figure 4. Comparing Round-Trip Latency

providers are usually bound by strict service level agreements with content distributors that include heavy penalties for downtime. To achieve this up time, operators typically employ 1+1 redundancy (known as hot standby) so that they can immediately switch to a dedicated standby unit

"The [cloud integrated network] CIN is comprised of two related and essential architectural elements: the edge cloud that is embedded in the network; and a re-partitioned access and aggregation network that delivers ultra-high-bandwidth and ultra-low-latency to and from this edge cloud." —*The Future-X Network: A Bell Labs Perspective*

when they detect a unit failure. Therefore, each function-specific server has a dedicated redundant server.

To support video at the network's edge, operators need more hardware at the edge. It is clear that the traditional architecture of custom hardware and 1+1 redundancy cannot scale—the CapEx is too prohibitive.

If, in contrast, each function is mapped to a computationally efficient, workload agnostic server, a single redundant server can take the place of any failed unit. In this scenario, operators can employ N+1 redundancy schemes: a warm standby. When a failure is detected, the operator's management system automatically configures the replacement unit and switches over to it. See Figure 5.

1+1 redundancy schemes scale acceptably in a centralized architecture. Separating the hardware by workload is not a significant cost burden, and the compute headroom per workload type, for the whole site, can be shared across a larger number of servers. However, the inevitable trend towards geographically dispersed computing infrastructure far from a centralized hub, dilutes the overall computing headroom. Moreover, because >80% of the workload is video, operators must build in redundancy.

With a workload agnostic hardware server processor, operators can instead deploy N+1 redundancy, reducing their CapEx investment per edge node by about 40% (i.e., 40% fewer servers would be needed at the edge) a massive CapEx and OpEx saving!

But what about the central cloud where the workload is not as latency sensitive, but 80% of central cloud traffic is still video? Using the same edge-agnostic FPGA accelerated processing hardware in the cloud that is deployed at the edge minimizes OpEx with automated workload orchestration and inventory management.

Conclusion

Supporting the anticipated leap in video bandwidth requires a fundamental shift in the distribution architecture. Dispersing central computing servers to a distributed model pushes video processing to the network edge. To support that new methodology cost effectively, 1+1 server redundancy must be replaced by N+1 redundancy with workload agnostic hardware. Intel's DCP and MCP, featuring the Xeon processor and Arria 10 or Stratix 10 FPGA, provide the flexibility needed to support agnostic hardware. Additionally, with Intel/eBrisk H.265 video encoding IP, this highly programmable workload agnostic hardware supports computationally intensive video processing.

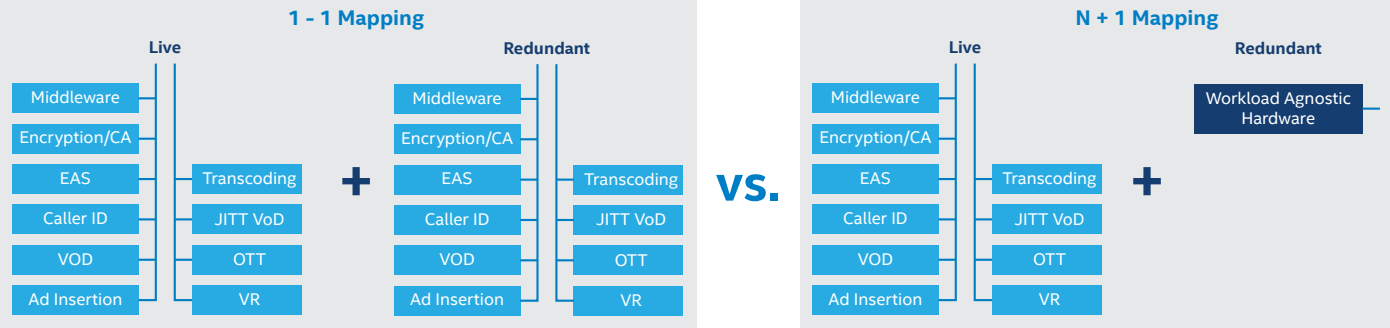

Figure 5. Comparing Redundancy Schemes

References

1 http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html
2 https://www.bell-labs.com/our-research/future-x-book/
3 http://ark.intel.com/products/family/88210/Intel-Xeon-Processor-E3-v5-Family
4 http://rd.springer.com/chapter/10.1007%2F978-3-319-25778-5_4
5 https://software.intel.com/en-us/blogs/2014/01/28/memory-latencies-on-intel-xeon-processor-e5-4600-and-e7-4800-product-families
6 http://www.prnewswire.com/news-releases/ebrisk-enables-4-hevc-4k60fps10bithdr-channels-in-1-rack-unit-610269645.html
7 https://www.hpe.com/us/en/product-catalog/servers/edgeline-systems/pip.hpe-edgeline-el4000-converged-iot-system.1008670180.html
8 http://arstechnica.com/gaming/2013/01/how-fast-does-virtual-reality-have-to-be-to-look-like-actual-reality/
9 https://www.superdataresearch.com/market-data/virtual-reality-industry-report/
10 http://www.nctatechnicalpapers.com/Paper/2015/2015-the-next-evolution-in-cable-converged-distributed-and-virtualized-access-network
11 https://aws.amazon.com/greengrass/
12 http://www.networkworld.com/article/2341775/lan-wan/five-nines--by-the-book.html

Where to get more information

For more information about Intel and Arria 10 FPGAs, visit **https://www.altera.com/products/fpga/arria-series/arria-10/overview.html**

For more information about Intel and Stratix 10 FPGAs, visit **https://www.altera.com/products/fpga/stratix-series/stratix-10/overview.html**

WP-01268-1.0