Solution Brief

Sports, TV Broadcasters, Entertainment, and Gaming
CTOs, CEOs, CIOs, and Broadcasting Companies

Gcore Boosts Live Stream Density with Intel® Data Center GPU Flex 140

Gcore Video Streaming integrates Intel Data Center GPU Flex 140 for low-latency live streaming and high broadcast video quality

Live streaming has opened new doors for the mass distribution of content, including by enabling the creator economy and expanding live sports streaming by broadcasters and over-the-top (OTT) providers.

These use cases are driving the rapid growth of live-streaming audiences. According to research from GameSquare, a record 2.76 billion hours of live content has been consumed on streaming platforms in 2023 – a 75% increase since 2019.[1]

To keep up with the growing industry needs, streamers have two options for building out their streaming infrastructure – either endlessly investing in their own new infrastructure and undertaking costly trials to evaluate new technology, or teaming up with a global, cloud-based streaming services provider that fully covers infrastructure construction, maintenance, and upgrades.

As a leading streaming services platform provider, Gcore is at the cutting edge of creating and adopting the latest technologies that cut down on latency and eliminate interruptions to video streaming.

Gcore Delivers Complete Streaming Platform on Demand

Gcore Video Streaming is a cloud-based video streaming platform that supports content delivery network (CDN)-based delivery and embedded players and playback. The high-capacity streaming service has delivered low-latency, scalable, live broadcasts to audiences of more than 1 million simultaneous viewers.

The service supports adaptive bitrate (ABR) encoding at no additional cost, which works with any viewer and enables smooth viewing even over unstable internet connections. The platform supports a variety of formats, codecs, and multiscreen options to meet user demands for videos on all platforms - from 360p (5" phone) to 4K@60fps (75" smart TV) with codecs and resolutions transparently adjusted for maximum quality and minimum latency.

This computationally heavy workflow used to require that the service provider purchase additional software/hardware to boost throughput and improve latency. These acceleration technologies are standard with Gcore Video Streaming.

Infrastructure for Streaming Video

Gcore Video Streaming provides the following features on Gcore's infrastructure:

- Low-latency streaming with mass-market delivery latency of 2-4 seconds
- Simple dashboard software development kit (SDK) and application programming interface (API)
- Over 180 global points of presence
- ABR encoding at zero additional cost
- Support for 4K ABR live streaming at broadcast video quality

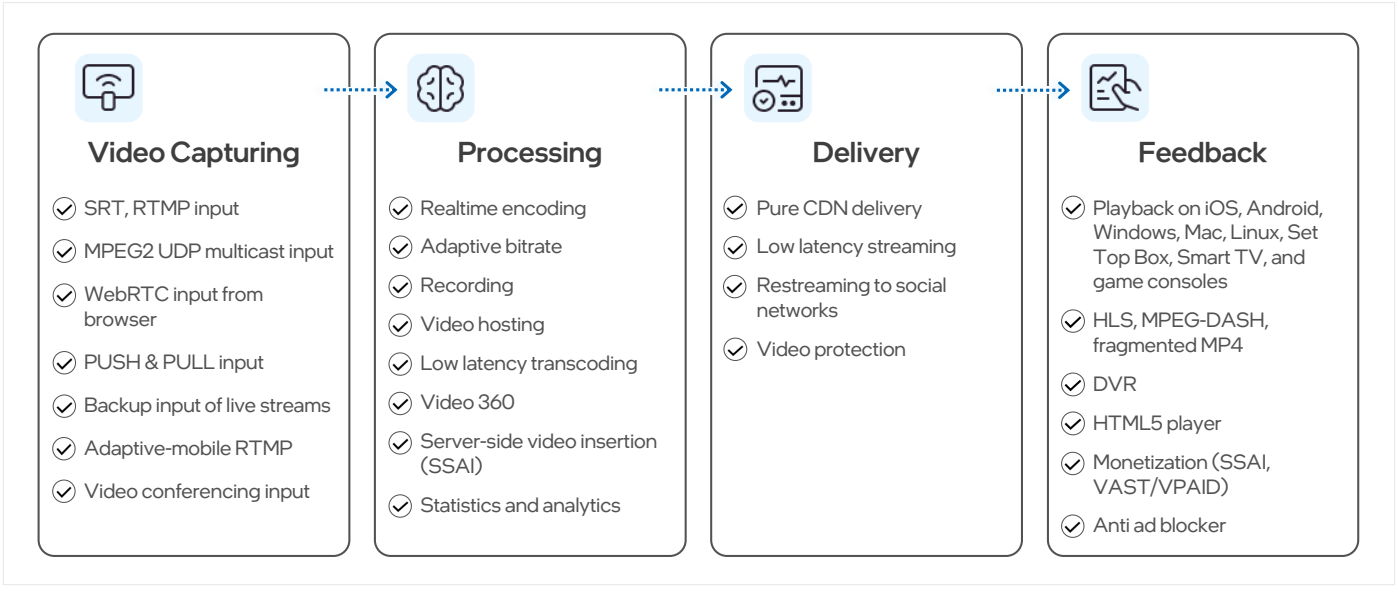


Figure 1. Gcore Video Streaming benefits.

Geography of Streaming Services

Gcore has built out a CDN with more than 180 points of presence (PoPs) worldwide with a total capacity of 100 million viewers. This robust, global network allows customer streams to be delivered around the globe with lowest possible latency.

Gcore Transcoding Performance is Enhanced by Intel Data Center GPU Flex 140

Before the video processor transmits the original video signal it needs to transcode the stream.

In the transcoding stage, the video is converted to a new format by adjusting parameters including encoding, bitrate, and resolution. Transcoding is a compute-heavy process that usually limits the number and quality of simultaneously processed streams.

Gcore stands out for its use of the Intel Data Center GPU Flex 140 to provide live streaming transcoding for H.264, HEVC, AV1, and Full HD / 4K /8K. With its integration with FFmpeg, the Intel Data Center GPU Flex 140 can encode high-density AVC, HEVC, and AV1 streams while still achieving ultra-low latency.

Figure 2. Gcore CDN worldwide locations.

The GPU has built-in AV1 encode functionality that creates a royalty-free AV1 codec. For Gcore's customers, the use of Intel Data Center GPU Flex 140 for Gcore Video Streaming offers the simplicity of simply picking the required number of streams, resolution, and other settings. Gcore then takes care of the backend technical aspects with support for even the highest density 4K/8K formats using H264/HEVC/AV1 codecs.

Each Intel Data Center GPU Flex 140 card (see Figure 3) features two GPU devices. Because of its low 75W power consumption, Gcore servers can support up to six cards per server. That provides 12 logical GPUs per system for a total throughput of ~200 1080p@60 streams per server. This allows Gcore to offer services at a low cost per stream but with very high quality and built-in codec advantage.

Media Performance and Density

Intel Data Center GPU Flex 140 cards provide simple preset configurations to optimize the encoder for density and quality. This lets users select a density and quality point that best meets the needs of their service.

Intel Data Center GPU Flex Series 140 achieves the highest performance levels for modern-generation video standards like HEVC and AV1 while still supporting ultra-high density and high-quality transcode. When using high-level APIs like FFmpeg, the GPUs provide three convenient operating presets that offer different tradeoffs between speed and quality.

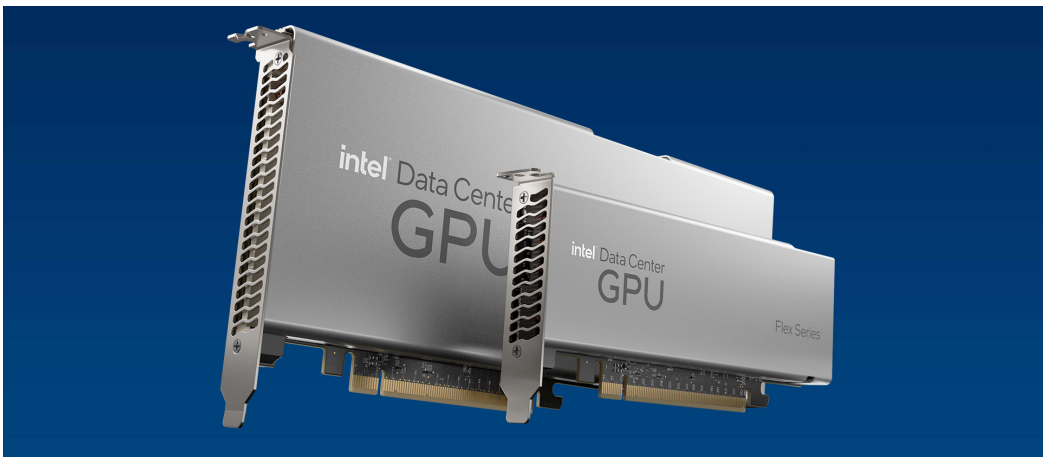


Figure 3. Two versions of the Intel Flex Series Intel Data Center GPU Flex 140 cards.

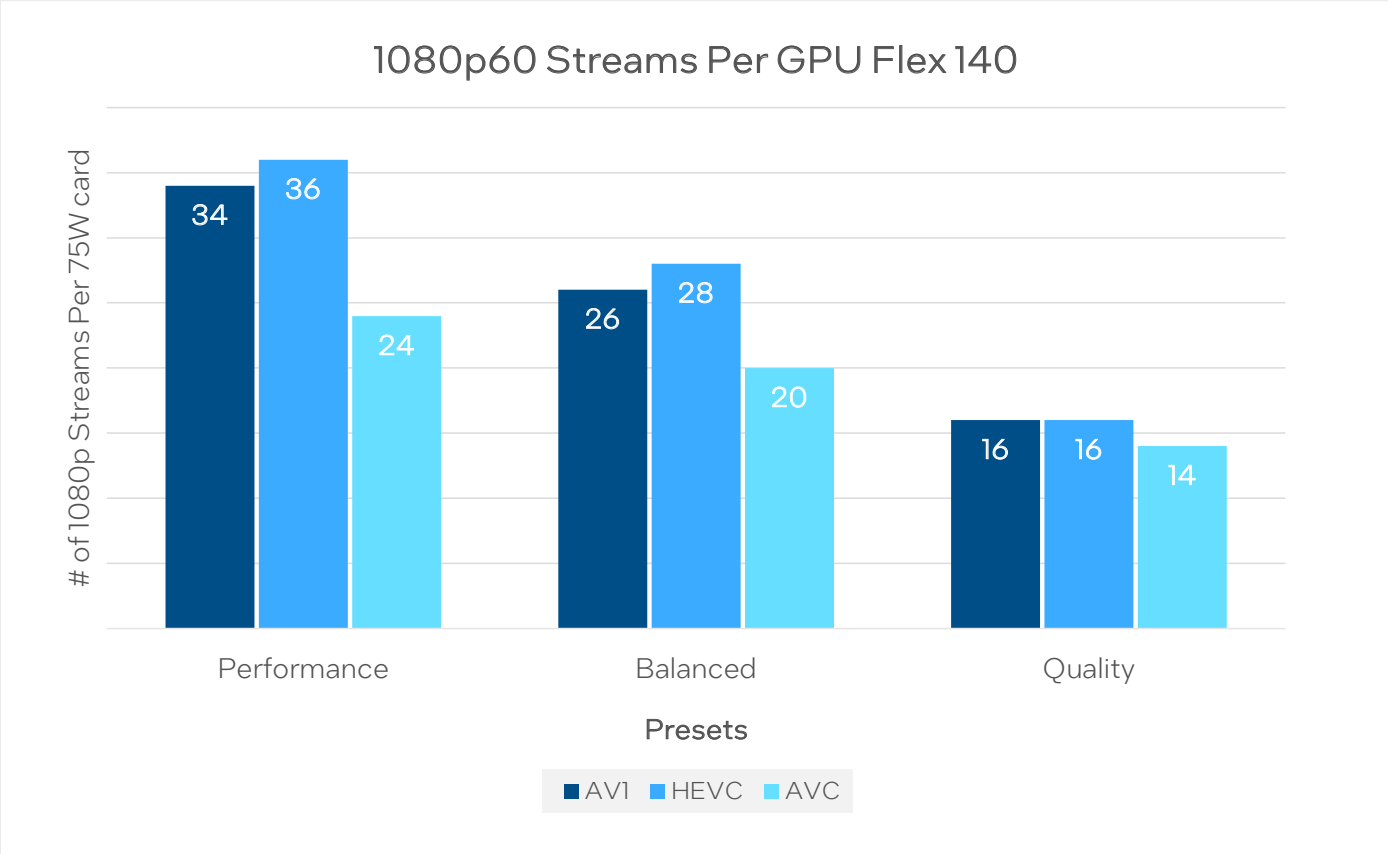


Figure 4. Performance for AV1, HEVC, and AVC transcoding across three presets (higher is better).

Intel Data Center GPU Flex Series 140 Performance

Figure 4 shows the measured performance for AV1, HEVC, and AVC transcoding across three presets, "performance," "balanced," and "quality."[2] The Intel Data Center GPU Flex 140 delivers up to 36 streams of 1080p60 HEVC in real-time in its "performance" mode, while AV1 and AVC can achieve up to 34 and 24 streams, respectively. When using the "quality" preset, the Intel Data Center GPU Flex Series 140 can deliver 16 streams of 1080p60 HEVC and AV1 and 14 streams of 1080p60 AVC at video quality levels documented in the following section.

Intel Data Center GPU Flex 140 Series Video Quality

Intel's advanced software bitrate controller is designed to boost GPU video quality for AVC, HEVC, and AV1 using various compression efficiency technologies and content-adaptive quality optimization tools. Some of the other key features include:

- Adaptive long-term reference
- Scene-change detection
- Adaptive motion compensation temporal filter
- Adaptive pyramid quantization
- Adaptive GOP
- Adaptive reference frames
- Adaptive custom quantizer matrix
- Look ahead
- Persistence adaptive quantization

As shown in Figure 5, the Intel Data Center GPU Flex 140 Series HEVC Encoder can achieve up to 14% bitrate reduction (BD rate) when compared to x265 medium preset for "non-low delay" use cases and 17% BD-rate savings in the "low delay" use case. In tests performed by Intel, "non-low delay" testing allowed for the use of lookahead by the codecs and was optimized for quality, while the "low delay" testing did not use lookahead and was optimized for minimum latency.

Intel Data Center GPU Flex 140 Series AV1 and AVC encoder achieved up to 37% and 20% BD rate savings, respectively, when compared to x264 medium preset when optimized for low latency usages.

HEVC BD RATE savings VS X265 Medium Preset			
Video Quality Preset	**Performance**	**Balanced**	**Quality**
Non-Low Delay	4%	11%	14%
Low Delay	9%	13%	17%

AVC BD RATE savings VS X264 Medium Preset			
Video Quality Preset	**Performance**	**Balanced**	**Quality**
Non-Low Delay	-7%	-5%	0%
Low Delay	17%	18%	20%

AV1 BD RATE savings VS X264 Medium Preset			
Video Quality Preset	**Performance**	**Balanced**	**Quality**
Non-Low Delay	11%	14%	17%
Low Delay	34%	36%	37%

Figure 5. BD rate savings under various presets.

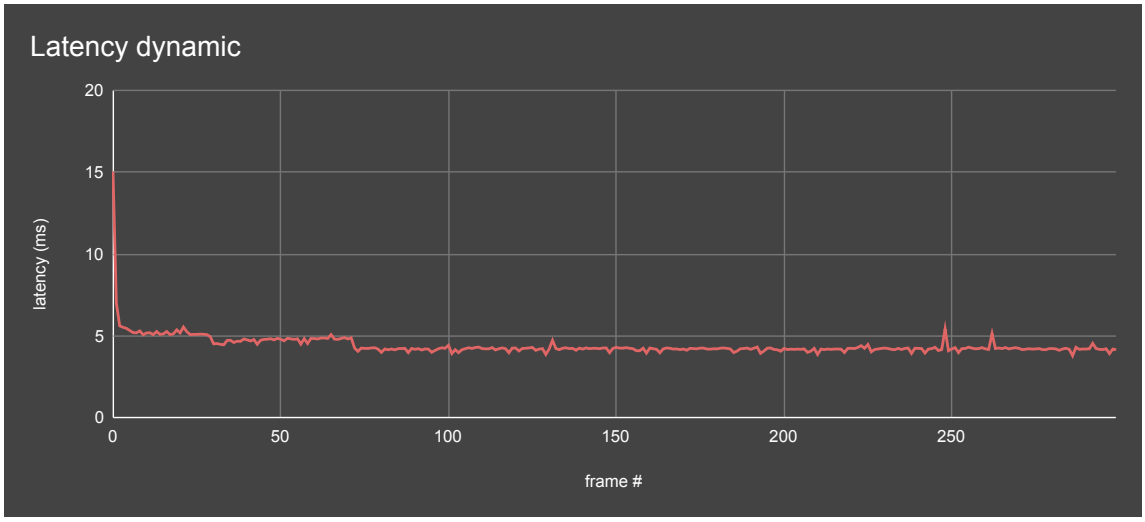
GPU Latency Benefits

The Intel Data Center GPU Flex 140 Series driver's low-latency session initialization and optimizations, coupled with dedicated hardware video processing blocks, ensure excellent low-latency performance at high density.

With an optimized video pipeline – including operating system support for interrupt handling, use of system memory latency tolerant buffers, and optimized driver memory allocation for video processing buffers – the solution was able to achieve excellent deterministic, low-latency, end- to-end per-frame processing.

Gcore's end-to-end latency testing showed the following results: single encoding process for non-B-frame content shows ~7% GPU load and 3-5ms of end-to-end encoding latency. The same process scaled up to x30 encoding instances resulting in a 97% GPU load with 30ms-50ms of latency.

Figure 6 shows the results from a single encoder:



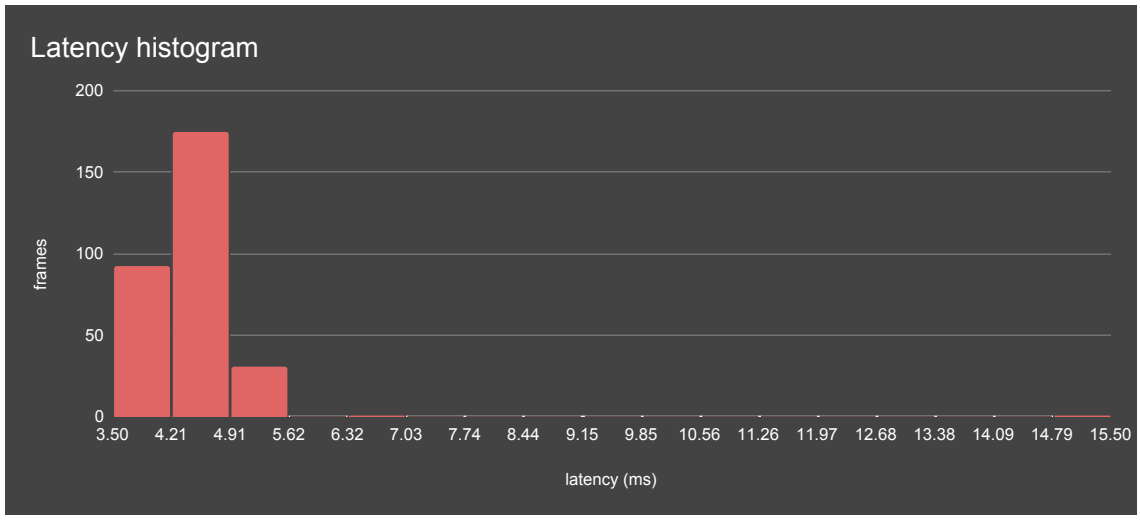


Figure 6. Latency results from single encoding process for non B-frame content.

Figure 7 shows the results from 30x multi-encoder:

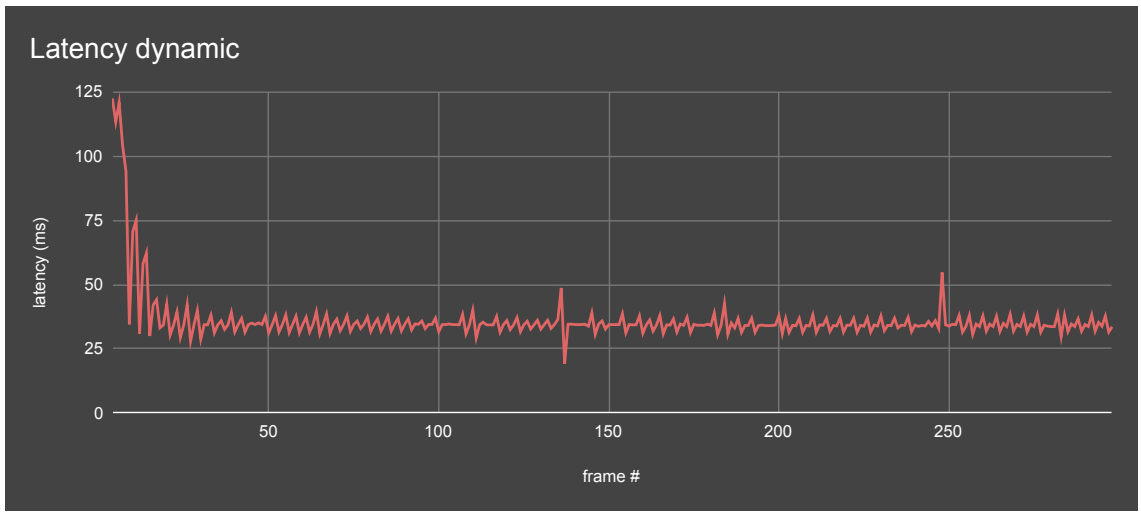


Figure 7. Latency results from 30x multi encoding process for non B-frame content.

Conclusion

Live streaming consumption is growing quickly. Gcore Video Streaming offers streaming content providers a cloud-based video transcoding solution that can be combined with Gcore CDN for seamless, global video distribution.

Gcore uses the Intel Data Center GPU Flex 140 to deliver performance and transcoding support for an intuitive and cost-effective experience. Benefits of the Gcore-Intel solution include:

- Widespread live streaming scale is available in a few clicks. Businesses can cut their Capex by using Gcore's manageable infrastructure.
- The Intel Data Center GPU Flex 140 provides a high-density deployment and low power consumption solution. It also allows Gcore to offer basic encoding for free (up to 1080p), and prices for HEVC/AV1 offer exceptional value even for low, per-minute volumes.

Gcore's solution, powered by Intel Data Center GPU Flex 140 innovations, provides the scalable, easy to use, video streaming platform that will drive the next wave of live streaming content.

Learn More

Gcore Live Streaming Platform

Intel® Data Center GPU Flex 140

Intel® Network Builders

[1] https://advanced-television.com/2024/01/11/top-gaming-live-streaming-trends-revealed/
[2] More information on the test set up and results can be found on GitHub.

0624/LV/H09/PDF ♻Please Recycle 360510-001US