Solution Brief

Accelerated Computing Systems and Graphics
Media Processing and Delivery

Intel® Data Center GPU Flex Series for Media Processing and Delivery

Media processing and delivery providers optimize both compute density and bandwidth consumption with Intel® Data Center GPU Flex Series, helping make their networks cost-efficient and future-ready.

Cisco reports that globally, business and consumer video accounts for 80 percent of all internet traffic.[1] Indeed, video assumes enormous social importance as user-created rich content becomes a primary means of communication and expression.

The demand for more sophisticated content produces upward pressure on costs to providers for both processing infrastructure and delivery bandwidth. Subscribers expect broadcast quality from internet delivery, including successively higher resolutions on a growing variety of end-user devices. At the same time, the buildout of 5G networks is massively increasing the throughput available to mobile subscribers.

In the face of flat average revenue per user (ARPU), meeting subscriber expectations for high quality creates substantial efficiency challenges for providers. High processing performance plays a pivotal role in meeting cost requirements by driving up the density of streams per server and providing optimal support for advanced streaming codecs, including AV1. That transition is critical for providers in order to deliver advanced video content with higher definition and frame rates.

Intel® Xeon® Scalable Processors are the gold standard for processing and delivering media today. They are compatible with common open source tools highly desired by media providers to overcome the constraints of siloed and proprietary environments. At the same time however, heterogeneous hardware and software environments persist for unique workloads, creating the need for a flexible solution that can meet emerging requirements while providing efficiency in the data center for high density amid workload complexity.

Intel Data Center GPU Flex Series is a general-purpose data center graphics processor optimized for media stream density and quality, with server-class reliability, availability and scalability. It can be used separately or in combination with Intel Xeon processors providing a flexible, open media solution to meet today's changing and challenging media delivery needs.

SUPPORTING STATS

5X **Media transcode throughput at half the power of the competition**

Intel Flex Series 140 GPU compared to NVIDIA A10

HEVC 1080p60[3]

2X **Decode throughput at half the power of the competition**

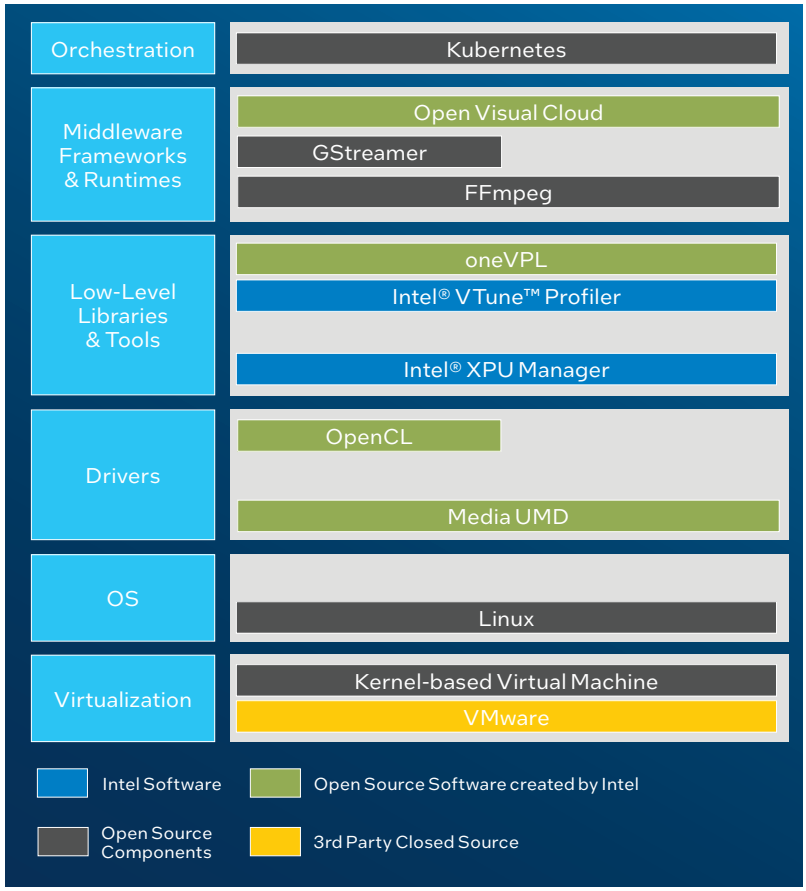
Intel Flex Series 140 GPU compared to NVIDIA A10

across HEVC, AV1, AVC, VP9[3]

Open Standards Architecture

Code developed for GPUs under proprietary programming models such as CUDA lacks portability to other hardware, creating a siloed development practice that locks organizations into a closed ecosystem. By contrast, the Flex Series GPU supports an open, standards-based software stack together with oneAPI cross-architecture programming so developers can build high-performance media applications and solutions that run seamlessly across Intel CPUs and GPUs.

Open standards code development based on oneAPI benefits from a large open ecosystem that includes open source tools, APIs, and drivers. That flexibility helps organizations reduce the complexity, cost, and time requirements to bring new services and solutions to market, streamlining adoption of new architectures and enabling engineers and programmers to innovate instead of maintaining code.



Industry Ecosystem

Extending the benefits of its standards-based open architecture and optimized software stack, the Flex Series GPU draws on a broad ecosystem of service providers, independent software vendors (ISVs), original equipment manufacturers (OEMs), and others to support a wide range of media use cases.

These companies draw on the published oneAPI oneVPL specification to integrate the GPU with their technologies for media processing and delivery. The openness and transparency of the programming model also encourages uptake by the open source community, creating a virtuous cycle to further enhance the software stack.

To effectively realize the underlying hardware's capabilities at delivering these media streams, Intel is enabling the software ecosystem to take advantage of them. This work helps ensure that software standards, frameworks, and open source technologies such as FFmpeg, GStreamer, and Handbrake—as well as the customers that use them—can attain performance on GPUs that has typically been possible only on CPUs. To reach that goal, Intel invested substantially to enable programmability for media processing and delivery across CPU and GPU architectures.

oneAPI Video Processing Library

The Intel® oneAPI Video Processing Library (oneVPL) offers optimized media transcode performance across integrated and discrete GPUs. oneVPL provides a video-focused API for video decoding, encoding, and processing in applications spanning media processing and delivery, broadcasting, streaming, video on-demand (VoD), cloud gaming, and remote desktop solutions.

Low-level encoder and rate controls provided by oneVPL enable developers to fine-tune encoder configurations for unique density/performance balance by use case beyond standard frameworks such as FFmpeg and GStreamer. They can also implement their own rate controls to marry customer domain expertise with Intel hardware innovation. oneVPL is backward-compatible with Intel® Media SDK core API.[3]

oneVPL can be downloaded individually for free. It is also included in the Intel® oneAPI Base Toolkit, which is a core set of tools and libraries for developing high-performance, data-centric applications on Intel CPUs and GPUs.

Analyze Application Performance with Intel® VTune™ Profiler

Accelerate application compute-intensive tasks by identifying the most time-consuming parts of GPU code and optimizing GPU offload schema and data transfers for SYCL, OpenCL code, Microsoft DirectX or OpenMP offload code. Analyze GPU-bound code for performance bottlenecks caused by microarchitectural constraints or inefficient kernel algorithms.
To learn more about supported operating systems, see system requirements.

High-Efficiency Codecs

Even as large-scale data storage has become progressively cheaper, CDN and other delivery costs remain high. Improved compression helps media processing and delivery providers reduce operating costs.

The Alliance for Open Media—a cross-industry consortium founded by Amazon, Cisco, Google, Intel, Microsoft, Mozilla, and Netflix—introduced the open source AV1 codec in 2018. This next generation codec built into the GPU brings the highest quality real-time video, scalable to any modern device at any bandwidth. It enables delivery of commercial or non-commercial user-generated content with low computational footprint, optimized for internet streaming. It does all this at 30% better compression with no degradation in streaming quality, reducing the cost per stream.

In addition to AV1, the GPU also supports existing HEVC, AVC, and VP9 codecs. Providers can maximize quality for the channels with the greatest viewership and highest-profile content using the SVT-AV1 software encoder developed by Intel in cooperation with the Alliance for Open Media. Drawing on the open source Scalable Video Technology (SVT) project for core libraries, the encoder is highly-optimized for Intel Xeon Scalable processors, providing outstanding performance and power efficiency on the same servers that host Intel Data Center GPUs.

These codecs can be accessed using standard frameworks, such as FFMEG or Gstreamer, or with oneVPL, which provides additional access to more controls and parameters. Both the hardware and software encoders provide a range of performance/quality presets so providers can make TCO-oriented adjustments according to the requirements of specific use cases.

SUPPORTING STAT

30% Better Compression

Without degradation in streaming quality[2]

Higher Performance with Lower Total Cost of Ownership (TCO)

Media processing and delivery providers have a strategic imperative to optimize TCO while meeting subscriber demands for more sophisticated content. The Flex Series GPU supports that objective by increasing the density of streams that can be supported per server without compromising quality, so service providers can handle a given subscriber base with fewer servers. It supports as many as eight simultaneous 4Kp60 streams or 30+ 1080p60 streams per card.[2] By supporting large numbers of streams per server, the GPU enables providers to address growing subscriber bases with smaller data center footprints, helping reduce CapEx associated with equipment and facilities costs. High performance per watt helps drive TCO down further by reducing OpEx at the same time that it supports corporate green initiatives by reducing the system's carbon footprint.

The GPU can significantly reduce the cost of media processing and delivery using AV1 versus x264 Medium. These savings extend across the total cost of service, including both preparation and distribution costs.

Service providers can scale graphics processing capacity from one to multiple GPUs per server to vary the mix of CPU and GPU resources as needed. The flexibility of this architecture is well suited to meeting the evolving needs of service providers with fast, high-quality real-time video decoding, encoding, processing, and media format conversion.

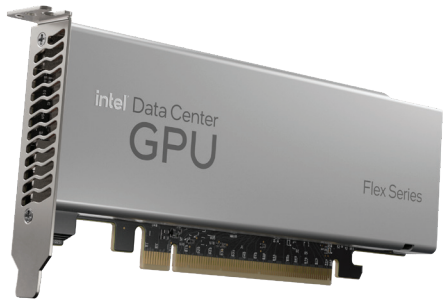
Intel® Xᵉ Architecture

Built on the Intel Xᵉ architecture, the GPU has up to 32 Intel Xᵉ cores and ray tracing units, up to four Intel Xᵉ Media Engines, AI acceleration with Intel Xᵉ Matrix Extensions (XMX), and support for hardware-based SR-IOV virtualization.

SUPPORTING STATS

8 Simultaneous 4Kp60 Streams - OR - **30+** Simultaneous 1080p60 Streams

per PCIe card[2]



The Future for Media Processing and Delivery

Media content providers are in a constant state of technology adaptation in the effort to deliver outstanding quality and customer experiences while keeping their eyes squarely on the bottom line. The industry transition replacing proprietary, specialized technologies with open, flexible, standards-based ones is a key contributor to this balance, along with innovation moving forward. The Flex Series GPU contributes to this transition with a seamless hardware and software media processing and delivery solution, untethering graphics programming from restrictive proprietary environments.

Learn more about the Intel® Data Center GPU Flex Series at www.intel.com/FlexSeriesGPU

[1] Cisco Systems, "VNI Complete Forecast Highlights." https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-forecast-highlights/pdf/Global_2021_Forecast_Highlights.pdf.

[2] Performance varies by use, configuration, and other factors. Learn more at https://www.intel.com/PerformanceIndex.

[3] Minor exceptions apply. See Intel, "Upgrading from Intel® Media SDK to Intel® oneAPI Video Processing Library: Transition Guide." https://www.intel.com/content/www/us/en/develop/documentation/upgrading-from-msdk-to-onevpl/top/developer-details/removed-features-details-and-mitigations.html.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See configuration disclosure for configuration details. No product or component can be absolutely secure.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Your costs and results may vary.

Intel technologies may require enabled hardware, software, or service activation.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a nonexclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

0822/MH/MESH/349354-001US