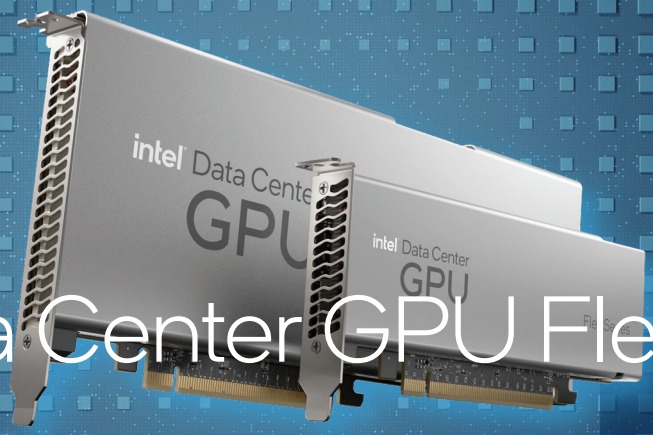
Product Brief

Accelerated Computing Systems and Graphics



Intel® Data Center GPU Flex Series

Intel Data Center GPU Flex Series is flexible, robust and the industry's most open GPU solution for the intelligent visual cloud.

Media processing and delivery, AI visual inference, cloud gaming and desktop virtualization are proliferating in data centers. That rapid growth comes to an industry largely beset by dependence on proprietary, licensed coding models such as CUDA for GPU programming. The use of software based on CUDA is also limited to proprietary GPUs, without portability to other accelerator architectures or CPUs. The resulting upward pressure on total cost of ownership makes proprietary GPU programming untenable at scale.

Intel® Data Center GPU Flex Series overcomes these limitations while delivering outstanding compute density and energy efficiency for visual cloud workloads. With acceleration for visual processing and AI built into the silicon, the GPU is based on Intel Xe HPG (high-performance graphics) microarchitecture. It provides capabilities and benefits that include:

- **Support for an open, flexible, standards-based software stack together with oneAPI unified programming,** which comprises open source components and libraries, tools and frameworks to build high-performance, cross-architecture media applications and solutions. This open approach helps the ecosystem break free of the technical and economic burdens of proprietary programming models.
- **Industry-first hardware-based open source AV1 encoder in a GPU,** improving bandwidth at the same quality by 30%, to save $23 million per 100,000 viewers per year or improve streaming quality over the same bandwidth.[1]

SUPPORTING STATS

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

Intel Flex Series 140 GPU compared to NVIDIA A10

HEVC 1080p60[1]

2X Decode throughput at half the power of the competition

Intel Flex Series 140 GPU compared to NVIDIA A10

across HEVC, AV1, AVC, VP9[1]

UP TO **68** 720p30 on select game streams

Single Intel Flex Series 170 GPU[2]

UP TO **46** 720p30 on select game streams

Single Intel Flex Series 140 GPU[1]

Hardware Specifications

The GPU will be available in two SKUs: the Intel Data Center GPU Flex Series 170 for maximum peak performance and the Intel Data Center GPU Flex Series 140 for maximum density. The graphics processor 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. Taking advantage of the Intel® oneVPL Deep Link Hyper Encode feature, the Flex Series 140 with its two GPUs can meet the industry's one-second delay requirement while providing 8K 60 real-time transcode. This capability is available for AV1 and HEVC HDR formats.

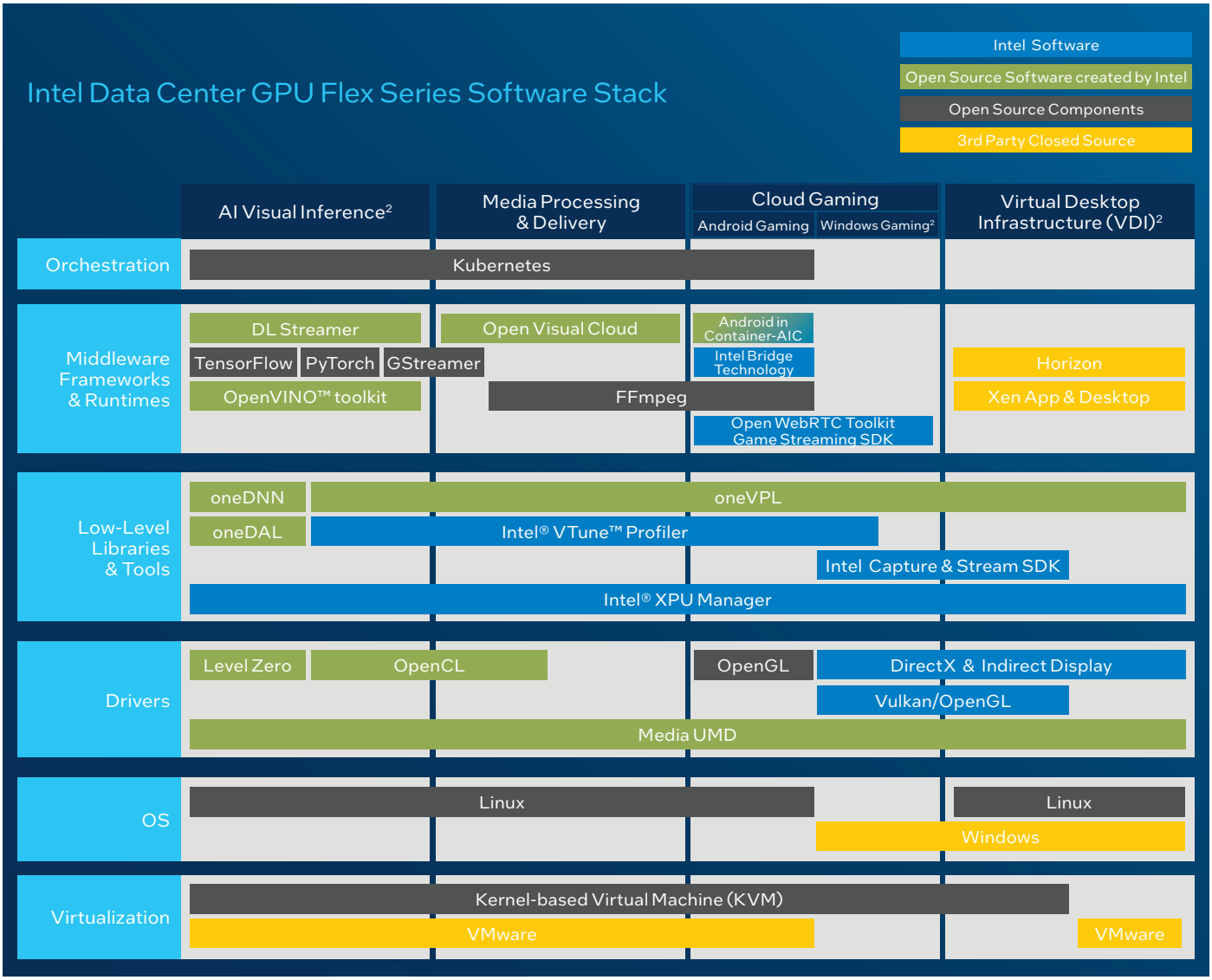
	Intel® Data Center GPU Flex Series 140	Intel Data Center GPU Flex Series 170
Target Workloads	Media processing and delivery, Windows and Android cloud gaming, virtualized desktop infrastructure, AI visual inference[2]	
Card Form Factor	Half height, half length, single wide, passive cooling	Full height, three-quarter length, single wide, passive cooling
Card TDP	75 watts	150 watts
GPUs per Card	2	1
GPU Microarchitecture	Xᵉ HPG	
Xᵉ Cores	16 (8 per GPU)	32
Fixed Function Media	4 (2 per GPU)	2
Ray Tracing	Yes	
Peak Compute (Systolic)	8 TFLOPS (FP32) / 105 TOPS (INT8)	16 TFLOPS (FP32) / 250 TOPS (INT8)
Memory Type	GDDR6	
Memory Capacity	12 GB (6 per GPU)	16 GB
Virtualization (Instances)[3]	SR-IOV (62)	SR-IOV (31)
Operating Systems	Linux (Ubuntu, CentOS, Debian), Windows Server 2019/2022, Windows Client 10, Red Hat® Enterprise Linux	
Host Bus	PCIe Gen 4	
Host CPU Support	3rd Generation Intel Xeon® Scalable Processors	

Software Stack, by Use Case

The Flex Series GPU supports an open, flexible, standards-based software stack with oneAPI cross-architecture programming. The stack includes open source components and libraries, tools and frameworks so developers can create high-performance, cross-architecture media applications and solutions to meet a wide range of use cases. This open approach removes the barriers to proprietary models where code portability and the ability to adopt new architectures across multiple vendors is limited.

The common set of software capabilities integrates into popular middleware and frameworks, and the stack is delivered in validated productized containers or reference stacks. The containers can be orchestrated with Kubernetes on bare metal or in VMs using SR-IOV virtualization with tools to assign and manage workloads. The toolset is designed to speed time-to-market and enable flexible deployment of multiple workloads on the same GPU.

Intel enables the software ecosystem through industry collaborations, initiatives and standards bodies. It also provides ongoing leadership, investment and technical contributions to the open source community.



Note: oneDNN is the oneAPI Deep Neural Network Library. oneDAL is oneAPI Data Analytics Library. oneVPL is the oneAPI Video Processing Library. oneVPL, oneDNN, oneDAL, and Intel VTune Profiler are in the Intel® oneAPI Base Toolkit (individual tools can be downloaded separately). Intel-optimized TensorFlow & PyTorch are in Intel® AI Analytics Toolkit.

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

intel®

[1] Performance varies by use, configuration and other factors. Learn more on the Performance Index site.

[2] Reflects capabilities of Intel Data Center GPU Flex Series that will be available when product is fully mature.

[3] VMs will vary by use case.

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/349353-001US