Product Brief

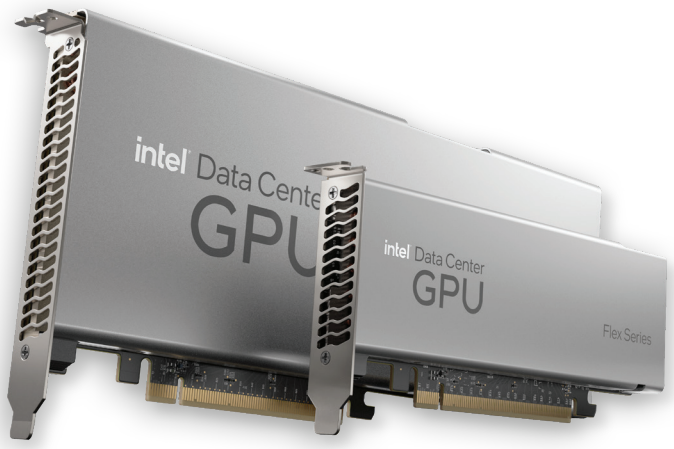
Intel® Data Center GPU Flex Series

Intel® Data Center GPU Flex Series for Cisco Servers

Intel® Data Center GPU Flex Series delivers flexibility and performance for a wide range of workloads on Cisco servers. It is the industry's most open GPU solution for the intelligent visual cloud.

The visual computing market is growing dramatically as media processing and delivery, AI visual inference, cloud gaming and desktop virtualization proliferate in data centers. With that growth, providers feel competitive pressure to reduce the cost per stream and improve stream quality and density over the same bandwidth. Multiple types of discrete chips are typically employed for media transcode, cloud gaming, real time media analytics and virtualization, creating complexity and siloed workloads.

Intel Data Center GPU Flex Series addresses these challenges with a flexible, open solution to meet today's changing and challenging media delivery needs, with future-readiness for emerging requirements. The Intel Flex Series GPU accelerates visual computing performance for media stream density and quality, with server-class reliability, availability and scalability, offloading and accelerating work from the Intel Xeon processor.



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]

Breakthrough Compute Density and Latency

The Intel Flex Series GPUs offer a seamless hardware-software media solution that meets both compute density and latency requirements with one chip.

Industry-First Built-in AV1 codec

Services built on the royalty-free open-source AV1 codec mean lowering operational expenses while providing higher video quality. Advanced video coding (AVC), High Efficiency Video Coding (HEVC) and VP9 support also comes standard with the Intel Data Center GPU.

The Industry's Only Open-Source Media Solution Stack

Deploy easily across CPUs and GPUs with Intel oneAPI, and remove dependence on proprietary, licensed coding models, such as CUDA for GPU programming, which limit software portability. The Intel Flex Series GPU supports an open, flexible, standards-based software stack supported by Intel oneAPI that enables developers to build high-performance, cross-architecture applications and solutions. This helps organizations reduce the complexity, cost and time requirements to bring new solutions to market, enabling engineers and programmers to innovate instead of maintaining code. When existing code investment is significant, developers can use the Intel oneAPI Compatibility Tool to migrate to Intel DPC++.

Licensing Cost Savings for VDI

There are no additional virtualization costs for GPU software licensing in Virtual Desktop Infrastructure (VDI) usages, and no management of licensing servers for VDI deployment.

Hardware Specifications

Select the Intel Data Center GPU Flex 170 for maximum peak performance and the Intel Data Center GPU Flex 140 for maximum density and low profile. 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 8K60 real-time transcode.[1] This capability is available for AV1 and HEVC HDR formats.

	Intel® Data Center GPU Flex 140	Intel Data Center GPU Flex 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[4]	
Host Bus	PCIe Gen 4	
Host CPU Support	4th Gen Intel® Xeon® Scalable Processors	

Software Stack oneAPI Tools

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.

- Intel oneAPI Video Processing Library (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.
- Intel oneAPI Deep Neural Network Library (oneDNN) is an open-source cross-platform performance library of basic building blocks for deep learning applications.

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.

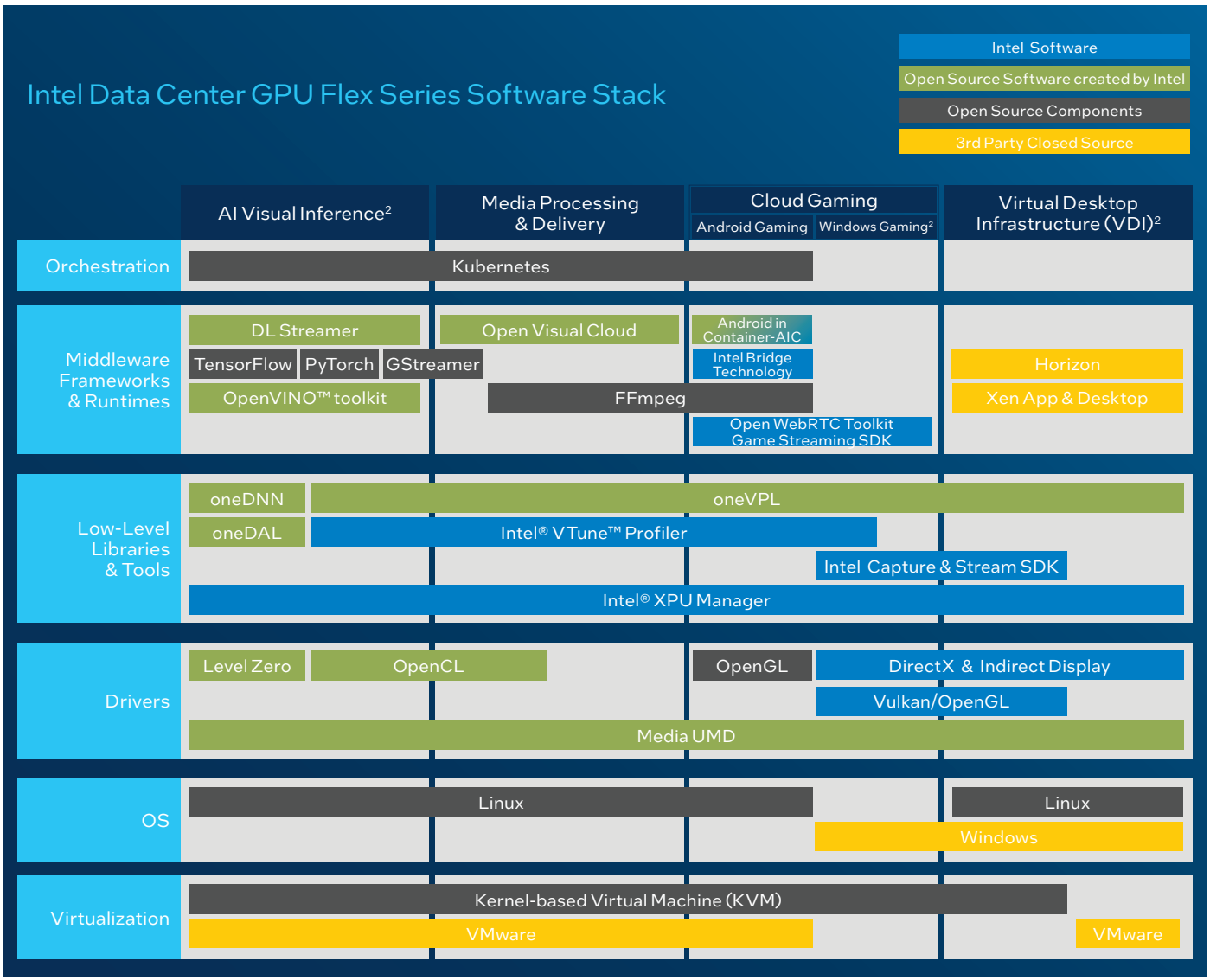
SUPPORTING STATS

8 Simultaneous 4Kp60 Streams - OR - **30+** Simultaneous 1080p60 Streams per PCIe Card[1]

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.

SUPPORTING STAT

Up to **150 TOPS INT8** (Tera Operations per Second) per PCIe card



Note: In the above figure, oneDNN is the oneAPI Deep Neural Network Library, oneDAL is oneAPI Data Analytics Library and 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 and PyTorch are in Intel AI Analytics Toolkit.

Product Availability

Intel Data Center GPU Flex Series is supported on the Cisco UCS C220 M7 and C240 M7 rack servers and the UCS X440p PCIe node for Cisco UCS X-Series Modular Systems. The Cisco product IDs for the GPUs are as follows:

- UCSC-GPU-FLEX140 (server-installed)
- UCSC-GPU-FLEX140= (spare adapter)
- UCSC-GPU-FLEX170 (server-installed)
- UCSC-GPU-FLEX170= (spare adapter)

Note: Servers supported as of the date of this publication. For up-to-date server compatibility, please check https://ucshcltool.cloudapps.cisco.com/public/. These versions of the Intel Flex GPUs are released directly by Cisco with a Cisco PID. The drivers and firmware must be updated with the Cisco software releases and HUU utility.

Learn more:

Intel® Data Center GPU Flex Series
Cisco UCS Servers

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

[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.

[4] Varies by workload; contact your Cisco representative for details.

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.

0323/MH/MESH/353910-001US