Solution Brief

Accelerated Computing Systems and Graphics
Intel® Data Center GPU Flex Series
Virtual Desktop Infrastructure

Intel® Data Center GPU Flex Series for Virtual Desktop Infrastructure

Outstanding user experience at scale, supporting open and key VDI infrastructure

For a visually rich graphical experience of virtual desktop infrastructure (VDI) select the Intel® Data Center GPU Flex Series. Delivering a highly performant, dependable, scalable future VDI solution for today's knowledge worker.

Deployment of virtual desktop infrastructure continues to accelerate, with the market segment forecast to grow at a CAGR of approximately 20.3% from 2023 to 2030, reaching about $50.5 billion by the end of that period.[1] Adding data-center-scale GPU resources to CPU virtualization solutions can dramatically improve user satisfaction and quality of service. Improved latency can deliver greater responsiveness, and higher framerates provide a smoother, more fluid visual experience, with reduced CPU core utilization. Relieving the CPU of the render + encode burden increases system headroom, enabling more sessions to be hosted per server, contributing to efficiencies in both capital and operating expense (CapEx and OpEx).

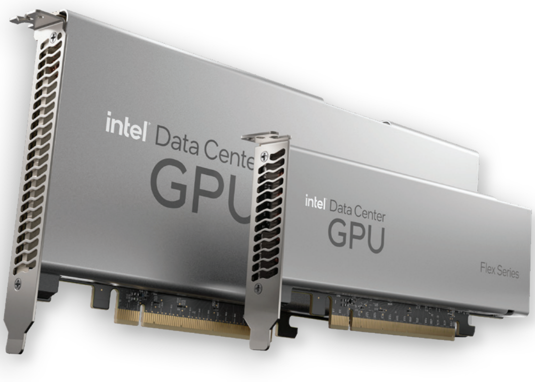
Intel® Data Center GPU Flex Series is easy to deploy, with support for the majority of popular VDI solutions. It delivers software licensing savings with no additional licensing costs for virtualization of GPUs and no management of licensing servers for VDI deployment. vGPUs deliver the GPU levels of performance, quality and capability users expect, with graphics, media and compute acceleration as well as CPU utilization offloads.

The Flex Series GPU helps enable remote access, to increase productivity wherever workers are and the need is. It offers an efficient, flexible solution to keep employees up and running with the resources, access and performance they need to be productive and innovative. The use of vGPUs also helps make businesses more future-ready, as the foundation for increasingly graphics-intensive office productivity applications used by knowledge workers and as AI becomes more integrated into mainstream workloads.

- **Intel Data Center GPU Flex 170** is a 150-watt PCIe card, optimized for performance. It supports up to 16 VMs and is designed for usages where especially high knowledge worker graphics performance is needed.
- **Intel Data Center GPU Flex 140** is a smaller, low-profile PCIe card at lower power (75 watts) that is ideal for edge and data center usages. Each PCIe card can support up to 12 typical knowledge-worker VDI sessions.

Open GPU architecture and programming model

Intel Flex Series GPU supports a unified open standards-based software stack, together with oneAPI cross-architecture programming and no ongoing license requirement.



Single Root I/O Virtualization (SR-IOV)

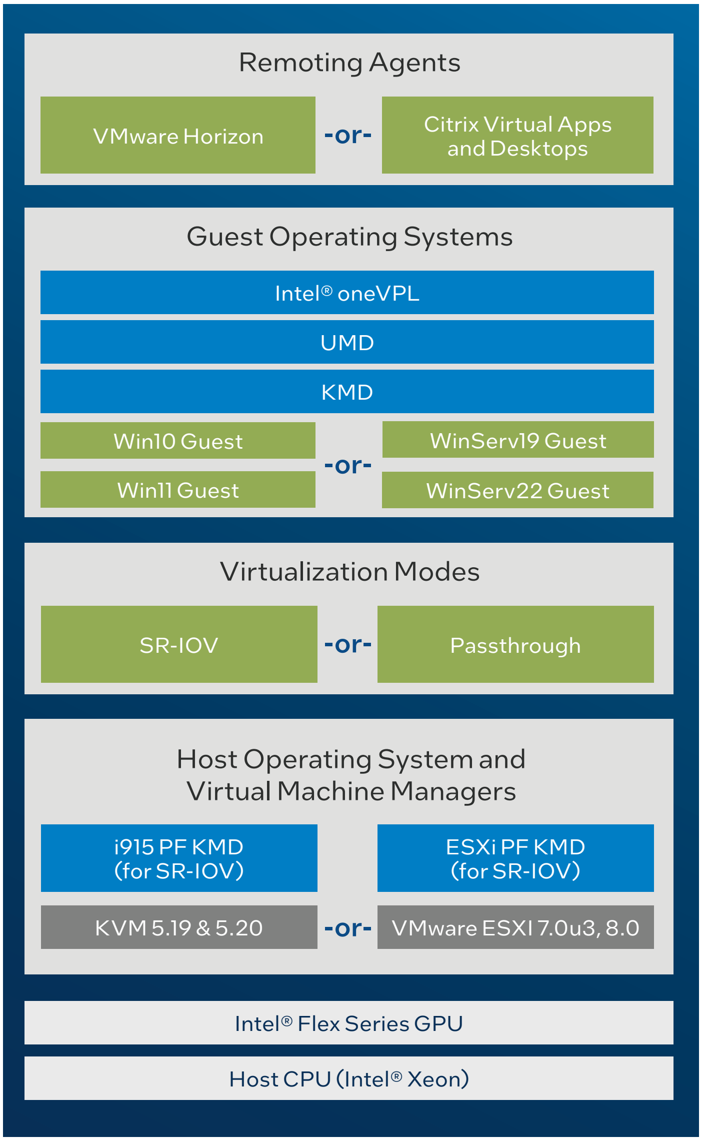
Intel Data Center GPU Flex Series supports hardware-assisted GPU virtualization using Single Root I/O Virtualization (SR-IOV) — an open, royalty-free PCIe standard. GPU resources are fractionalized and assigned to virtual functions (VFs) during the vGPU provisioning phase of the setup process, which can subsequently be independently assigned to each VM. This flexibility in vGPU administration improves scalability of the VDI server deployment. Depending on the service level agreement of each VDI session, a larger or smaller vGPU profile can be created and assigned to each VM.

High-efficiency codecs

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 Intel Flex Series GPU brings the highest quality real-time video scalable to any modern device at any bandwidth. It enables delivery of virtual desktops with a low computational footprint, optimized for internet streaming. In addition to AV1, the GPU also supports existing HVEC, AVC and VP9 workloads.

Cost-effective delivery of excellent user experiences

To validate performance of Intel Data Center GPU Flex Series in VDI sessions, Intel performed testing across the key performance indicators (KPIs) described in the remainder of this section. The testing is based on a typical mainstream knowledge worker persona, with tasks based on office productivity applications, various web-browsing usages and media playback. Details of the testing and discussion of performance results are given in the white paper, "Virtual Desktop Infrastructure (VDI) Performance on Intel Data Center GPU Flex Series".



Take the next steps

Explore the following resources to assess the opportunity for desktop virtualization solutions in your organization based on the Intel Data Center GPU Flex Series.

- **Get details about the GPU platform.** Product page: Intel Data Center GPU Flex Series
- **Dive deeper into solution architecture and performance.** White paper: "Virtual Desktop Infrastructure (VDI) Performance on Intel Data Center GPU Flex Series"
- **See smooth VDI experiences running on the GPU using VMware Horizon.** Demo: "Accelerate Remote Desktops with Intel Data Center GPU Flex Series"
- **Try out the full hardware and software stack, hands-on.** Intel Developer Cloud
- **Explore support for software providers.** Intel VDI Ecosystem

	Key Feature	Production Validated
Remoting Stack	VMware Horizon	Supported V8.2206
	Citrix Virtual Apps and Desktop	Supported V7.2206
Guest OS	Windows 10 Enterprise	Supported
	Windows 2019 Server	Supported
	Windows 2022 Server	Supported
	Windows 11 Enterprise	Supported
API	oneVPL support	Supported
CPU	Platform support	IceLake Sapphire Rapids
VMware virtualization modes	vDGA, vGPU (using SR-IOV), vSGA	Supported
Performance data	Intel® Data Center GPU Flex Series - Overview ›	

CPU-GPU offload while delivering high quality graphics performance[2]

Initial results indicate that virtual machines with Flex Series 140 GPU deliver high quality graphics performance, consistent frame rate of remoted frames and latencies in addition to significant CPU resource offload (~80% in 8vCPU VM) depending on the workloads. Ongoing testing shows substantially improved end-user experience of each VDI session and higher density of VDI sessions per server (using lower vCPU configurations per VM), thereby delivering better TCO for the VDI server deployment. This excess headroom also provides a degree of future proofing in supporting new workloads.

Minimal virtualization overhead at scale[2]

The tested performance of Intel Flex Series GPU SR-IOV virtualization scales linearly with an increasing number of VFs for typical knowledge worker VDI use cases. The overhead of vGPU time-division multiplexing is minimal; in extended validation of VDI use cases, it accounts for <1% of the overall time under test.

Low latency graphics from rendering and encode performance[2]

Flex Series 140 GPUs would enable a VDI server administrator to configure the server for high VM density (such as 2/4vCPU per VM) and deliver VDI sessions to knowledge workers with excellent user experience offering larger and multiple virtual displays at consistently higher frame rates with low framebuffer encode latencies.

Lower total cost of ownership with virtual GPU density[2]

Flex Series 140 GPUs will offer considerable ease and flexibility in setting up vGPUs without any licensing costs and license servers. All vGPU profiles will support VDI sessions with concurrent applications at various resolutions of single and multiple display configurations at a minimum of 30 fps as determined by the VDI remoting agent.

Business Benefits

The Intel® Data Center GPU Flex Series:

- **Easy to deploy:** Supports the majority of popular VDI solutions
- **Software licensing savings:** No additional licensing costs for virtualization of GPUs
- **Excellent graphics experience:** Virtual GPU and CPU utilization offloads
- **Enable remote access:** Increase productivity where the workers are and the need is
- **Future-ready:** Provide headroom for increasingly graphics-demanding office and web applications

Efficient GPU utilization[2]

Flex Series 140 GPUs will offer flexible GPU resource assignment to VFs via SR-IOV provisioning to ensure that adequate GPU resources can be assigned to each VF while keeping per-VM and overall GPU utilization to an optimum level with sufficient resource headroom. GPU telemetry data can be collected using Intel tools, such as XPU Manager. Current testing shows there is adequate GPU resource headroom (both engine capacity and local memory) for typical knowledge worker VDI use cases, and the Flex Series 140 GPU can deliver 12x 1080p sessions per card at just 30-35% GPU utilization and 60% local memory utilization. Experiments reveal that there is sufficient GPU resource headroom to accommodate more complex, concurrent GPU-focused applications at higher display resolutions and multiple displays.

The future of desktop virtualization

The Intel Flex Series GPU offers solution providers a cost-effective platform for delivering virtual desktop experiences with customer-winning quality. Drawing on the open-standards DNA of Intel architecture, the GPU is code-compatible with Intel CPUs. The open programming environment supports flexible development The Intel Data Center GPU Flex Series — through both hardware and software — drives high-density, high-quality desktop virtualization instances. It is built to run with Intel® Xeon® processors, complementing each other in the same system to handle diverse, complex workloads as efficiently as possible.

	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[3]	
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	X^e HPG	
X^e Cores	16 (8 per GPU)	32
Media Engine	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 GB per GPU)	16 GB
Virtualization (Instances)[4]	SR-IOV (12 per GPU)	SR-IOV (16 per GPU)
Operating Systems	Linux (Ubuntu, CentOS, Debian), Windows Server 2019/2022, Windows Client 10, Red Hat® Enterprise Linux	
Host Bus	PCIe Gen 4	

[1] Zion Market Research, March 14, 2023. “Global Demand for Virtual Desktop Infrastructure (VDI) Market Size Will Surpass $50.5 Bn by 2030 at 20.3% CAGR.” https://www.globenewswire.com/en/news-release/2023/03/14/2626677/0/en/Global-Demand-for-Virtual-Desktop-Infrastructure-VDI-Market-Size-Will-Surpass-50-5-Bn-by-2030-at-20-3-CAGR-Zion-Market-Research.html.

[2] Testing by Intel, 06/14/2023. Intel Server Board M50CYP, Intel Data Center GPU Flex 140, 2x Intel® Xeon® Gold 6336Y processors (36 MB Cache, 2.40 GHz, QXRV), 8x 16GB 3200MHz PC4-25600 ECC Registered 1.2 Volts DDR4 RDIMM, 1x 960GB Intel® SSD D3-S4610 Series SATA 6GB/s 2.5” SSD TLC, Intel® Ethernet Network Adapter X710-T2L for OCP 10Gbps Dual-Port Modular LOM, ESXi v8.0, Windows 10 Enterprise (10.0.19044), Horizon Agent v2212 -8.10.0-62933987, Horizon Client v2209, display resolutions 1080p/1440p/2160p, vGPU Profiles V1/V3/V6, AMC V6.6.0.0, BMC 2.88.097ec61c, IFWI ES029, PC Mark10 2.1.2525.64, VM config: 8 vCPUs/8 GB system memory, Secure Boot disabled.

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

[4] VMs will vary by use case.

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

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.

0923/JAW/MESH/353947-001US