White Paper

Version 1.0

Virtual Desktop Infrastructure (VDI) Performance on Intel® Data Center GPU Flex Series

Author

Narasimha C V,
Accelerated Graphics VDI Solutions,
Intel Corporation

Table of Contents

Introduction

The evolution of modern-day Virtual Desktop Infrastructure (VDI) usage models, applications and workflows require VDI Server solutions to integrate dedicated hardware accelerators to deliver,

- VDI sessions at high and consistent framerates to the client with uncompromised image quality whilst supporting modern display topologies
- Low latency graphics rendering and encode performance
- Efficient use of CPU and GPU resources per Virtual Machine (VM) with sufficient resource headroom to future proof VM configurations to support the latest workloads
- Improved density of users per VDI server thus lowering Total Cost of Ownership (TCO)

This whitepaper is intended to highlight the performance of the Intel® Data Center GPU Flex Series for VDI usage models. Various aspects such as end-user experience, GPU utilization, hardware encode performance, scalability using Single Root I/O Virtualization (SR-IOV), CPU-GPU offload, VM density and TCO will be discussed in the sections that follow.

Intel® Data Center GPU Flex 140 is a 75-watt low profile PCIe Gen4 GPU card for accelerating cloud gaming, media processing and delivery, VDI, and AI Visual Inference applications in data center servers. Each PCIe card has 2 GPUs, each with eight Xe cores, two media engines and 6GB of GDDR6 memory attached. Considering a graphics local memory provisioning granularity of ~1GB per virtual GPU, each Flex Series 140 GPU can support up to 12 VDI sessions for a typical knowledge worker persona with low-to-moderate graphics performance requirement per display pixel, making it a compelling solution for high density VDI deployment.

To meet workloads with higher graphics performance requirement per display pixel, a Flex Series 170 GPU can be considered. It is a 150-watt, full-size Gen4 PCIe card that has a single GPU node with 32 Xe cores, two media engines, and 16GB of GDDR6 memory. This whitepaper primarily focuses on performance studies executed on the Flex Series 140 GPU for VDI use cases specifically targeting the knowledge worker profile.

VDI and GPU Virtualization

Virtual Desktop Infrastructure (VDI) has a >10-year history as a means of centralized hosting of interactive Windows desktop environments and applications that are delivered from server infrastructure to client endpoints via standard remote streaming protocols. VDI encompasses a wide range of hosting configurations, user profiles, use cases and applications. Everything from a software-only rasterizer with no GPU of any kind, to a dedicated GPU per user may be employed to deliver the required user experience for each deployment.

From a graphics perspective, the level of performance/quality/capability that the user experiences in each connection will depend in large part on the level of GPU acceleration that is available to the target OS. If the target is a VM, the acceleration level will depend on what the VM Manager (VMM) makes available and how the administrator has configured the exposure of GPU resources for a given session.

Fundamentally, there are four means of providing a VM with access to graphics, media, and compute acceleration:

1. CPU Rasterization (no GPU): This is the traditional VDI configuration, running on data center servers that have no graphics resources.
2. Paravirtualization: This uses API-forwarding techniques, specialized hooks to allow VMs to request access to GPU resources and the Hypervisor then routes these requests to the Host OS Kernel mode graphics driver for GPU acceleration.
3. Device Passthrough: The PCI-e GPU device is assigned to the VM in full.
4. Virtual GPU (vGPU): Exposing GPU capability using Hardware assisted Virtualization (SRIOV, Hard Partitioned Graphics).

Growth of VDI and Graphics

In the past decade, and especially the past few years, large enterprises have seen the growing popularity of remote work, a steep rise in the need for remote desktops, workstations, and app streaming. Convenience of remote working aside, VDI brings great benefits to enterprises such as:

- Centralized management of user applications and profiles
- Improved security – data never has to leave the datacenter
- Bring Your Own Device (BYOD) – deliver applications to any device, anywhere

VDI users can be broadly classified based on their target use cases and applications. Figure 1 shows the relative requirement of GPU resources across four user personas/profiles – task worker, knowledge worker, power user and designer.

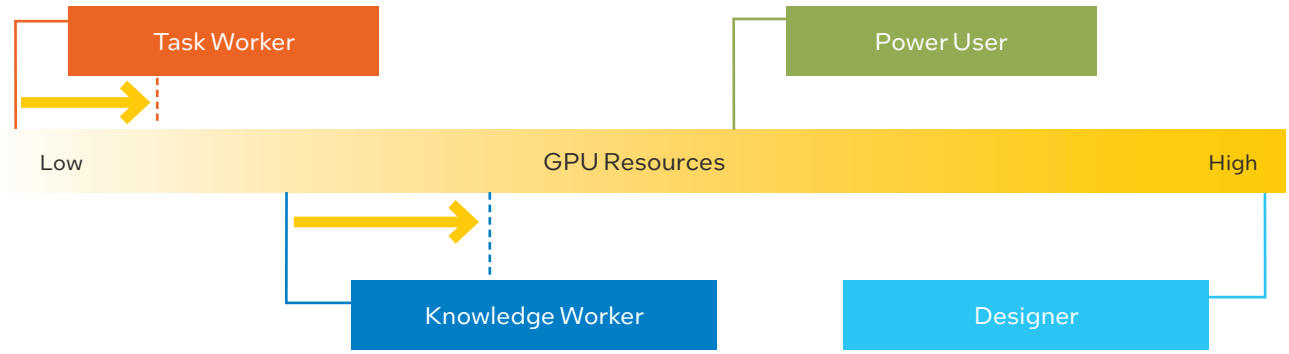


Figure 1. VDI user profiles mapped to GPU resource requirement

Power users and designers require high GPU performance to render complex images quickly at high quality. Fields of work that clearly mandate the need for a dedicated GPU include content creation and design, engineering, and architecture which may have specific benchmarking guidelines and certification requisites.

More interestingly, there is a rapidly growing need for GPU acceleration in the knowledge worker persona which can be attributed to:

- Adoption of higher resolution displays (QHD, 4K) and multiple monitor configurations
- Webpage and browser experience becoming more GPU intensive due to interactive content using HTML5, WebGL, embedded media
- Demand for higher fidelity VDI sessions – i.e., better image quality at typical resolutions such as 1080p
- Video Conferencing and collaboration applications becoming part of modern workflow
- Increase in use of graphics (via DX, OGL, DXVA APIs) in the out-of-box Desktop user experience and essential office productivity applications on the latest Windows client operating systems.

Benefits of Virtual GPUs in VDI deployments

The following are the key motivations for using Intel Flex Series vGPUs in VDI use cases across on premise, public and hybrid cloud deployments.

Higher VM density

In each VM, the vGPU serves all the 3D, media and compute workloads and encodes the framebuffers of the VDI session, thereby freeing up vCPU resources. This offload of CPU utilization towards GPU resources helps provide the resource headroom for 'futureproofing' the VM configuration to support the constantly evolving modern desktop applications that demand higher graphics performance. Freeing up CPU resources creates compaction possibilities and VM density improvement, i.e., more sharing of vCPUs across VMs, or reduction in vCPU allocation for each VM.

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 (VF) 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 (SLA) of each VDI session, a larger or smaller vGPU profile can be created and assigned to each VM.

Graphics User Experience

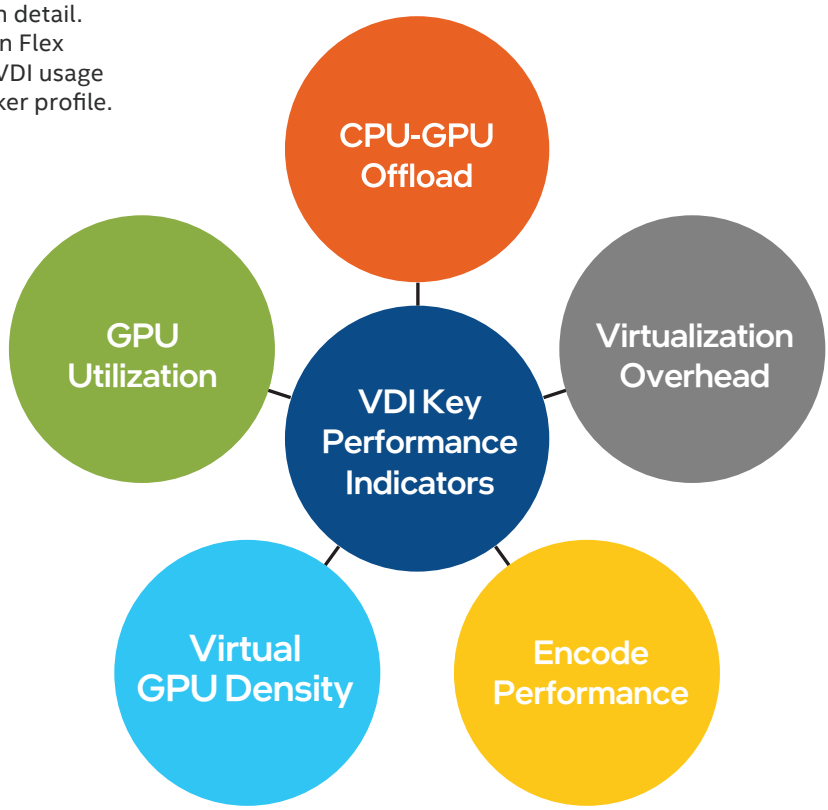
Presence of a vGPU in a VDI session results in overall better graphics end user experience. This can be perceived in the form of:

- Higher Frames Per Second (FPS) of remoted frames as observed on Client endpoint devices
- Consistency of FPS
- Improved end-to-end, click-to-photon latency
- Scalability to higher display resolutions (1440p, 4K, 5K and above) and multiple displays per VDI session.
- Faster rasterization and lower frame encode latency
- Highest image quality and support for the latest media encode formats
- Better user interface responsiveness and quicker app startup times owing to CPU-GPU offload

Key Performance Indicators

This section will cover each of the VDI Key Performance Indicators and discuss the performance studies in detail. Several technical studies have been performed on Flex Series 140 GPUs to quantify the performance of VDI usage scenarios primarily targeting the Knowledge worker profile.

Knowledge worker scenarios (see Figure 2) are typically office productivity applications (including video conferencing) with additional casual desktop usage such as web browsing and media playback. It is a mix of applications that exhibit both sustained GPU utilization and ones which demand GPU performance in spikes.



VDI Remoting Agent	Frame buffer Encode - AVC, HEVC, AV1
Video Conferencing	Microsoft Teams
	PCMark10 (Video Conferencing)
Office Productivity	Microsoft Office 365 Apps - Excel, Powerpoint, Word, Outlook
	PCMark10 (Essentials, Productivity)
Web browsing (Render)	Google Earth and Maps
	WebGL, HTML5 interactive content
Web browsing (Media)	Instagram, Spotify, Youtube Music
	Youtube Video, MSN Video
Web browsing (Casual)	Chrome browsing
	Edge browsing stress test
Media Playback	Media Playback - AVC, HEVC, AV1, VP9 (MTV, VLC, Browser) - 1080p, 4k

Figure 2. Knowledge Worker workload set used for VDI validation

In our testing methodology, usually two to three concurrent applications are instantiated per VM with varying complexity and temporal spacing. Concurrent VDI sessions with GPU hardware encoding are launched at various virtual display resolutions, and it is ensured that each workload set fits within the available GPU resource budget for each vGPU profile. Figure 3 calls out the system configuration details.

HW Config	GPU	Intel® Flex 140
	Motherboard	Intel® Server Board M50CYP
	Processors	2 x Intel® Xeon® Gold 6336Y (36M Cache, 2.40 GHz, QXRV)
	Memory	8 x 16GB 3200MHz PC4-25600 ECC Registered 1.2 Volts DDR4 RDIMM
	Storage	1 x 960GB Intel® SSD D3-S4610 Series SATA 6GB/s 2.5" SSD TLC
	Network	Intel® Ethernet Network Adapter X710-T2L for OCP 10Gbps Dual-Port Modular LOM
SW Config	ESXI OS Version	v8.0
	Windows Client OS	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	vCPU	8
	System Memory	8 GB
	Secure Boot	Disabled

Figure 3. System Configuration used for VDI performance validation as of 06/14/2023

CPU-GPU Offload

With the introduction of a virtual GPU to a VM, the end user not only observes an immediate improvement in overall graphics performance, but also a dramatic reduction in CPU utilization.

Figure 4 depicts the performance comparison for a typical benchmark such as PCMark10 (link), which covers a wide range of tasks usually performed in modern PC use cases.

PCMark10: V1 profile	2 vCPUs, 16GB RAM		
	With vGPU	CPU Only	GPU / CPU only Performance
App startup	8500	8689	0.98
Video Conferencing	6773	3425	1.98
Web Browsing	6474	3783	1.71
Spreadsheets	4129	4532	0.91
Writing	5211	5459	0.95
Photo Editing	5978	1233	4.85
Rendering and Visualization	3136	-	-
Video Editing	3528	1272	2.77

Figure 4. Performance of vGPU-accelerated vs CPU-only VMs

For Digital Content Creation and Essentials segments, the VM with a vGPU clearly outperforms the CPU-only VM and this is obvious in the end user experience as well. It can be reasoned that in the productivity case, the CPU-only VM performs marginally better. But this comes at the cost of higher CPU utilization and reduces the ability of a VM configuration to support newer and/or more concurrent applications. Note that in the CPU-only case (see Figure 5), the VM is operating at >60% CPU utilization for ~43% of the time under test. On the VM with a vGPU (refer Figure 6), this is as low as ~18% and nearly 60% of the time under test the VM is operating at a CPU utilization of <40%.

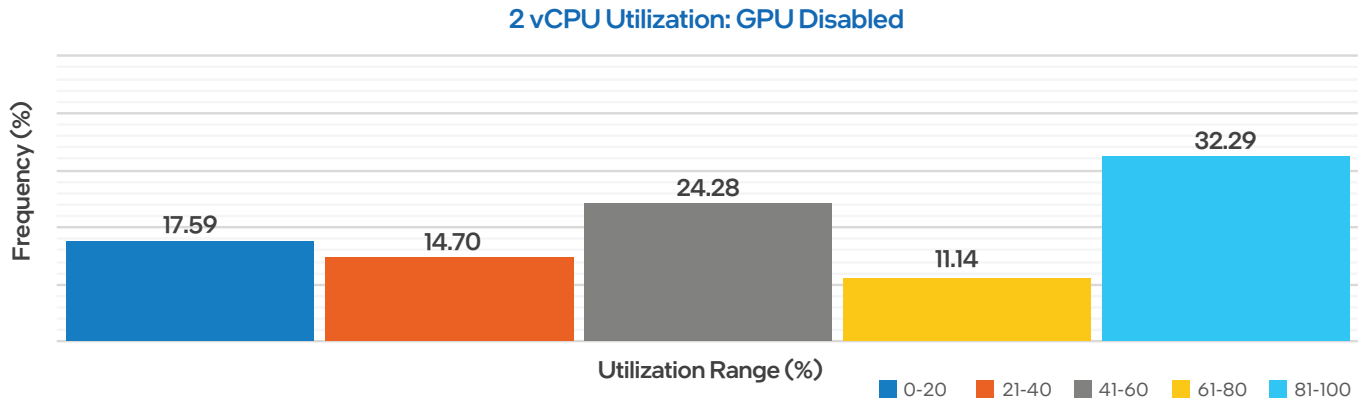


Figure 5. CPU Utilization histogram - CPU-only VM running PCMark10

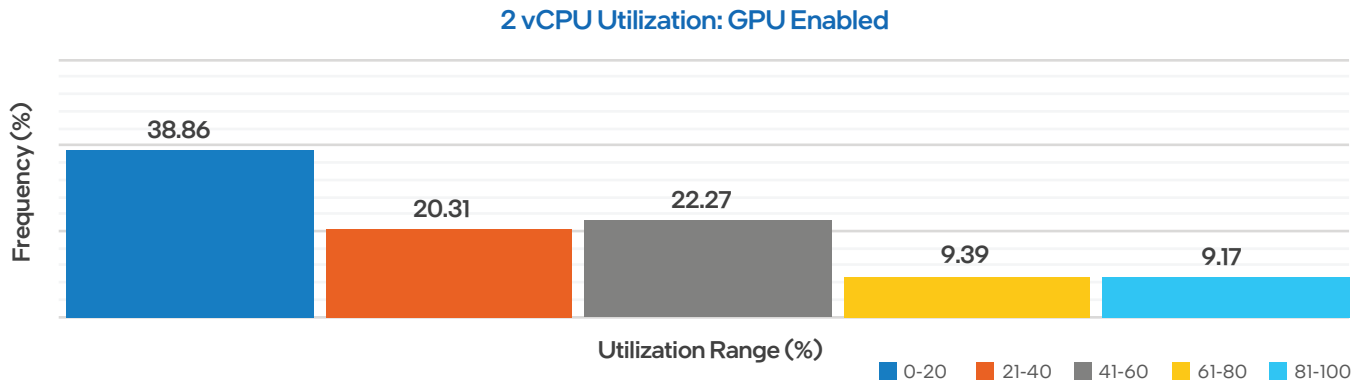


Figure 6. CPU Utilization histogram - VM with vGPU running PCMark10

The observations are more remarkable on running sustained GPU-centric workloads such as WebGL or media playback. Figures 7 and 8 show two VDI sessions (one with vGPU another CPU only) running the same WebGL application in the browser. Each VM is configured identically (8vCPUs and 16GB RAM) and the VDI agent used here is VMware Horizon hosting a single display at 1080p resolution.

The CPU-only VM is operating at close to 100% CPU utilization and struggling to deliver a 1080p VDI session at just 13 fps. On the other hand, the VM with vGPU delivers consistent VDI end user experience at ~30fps and offers considerable CPU resource headroom (CPU Utilization <10%).

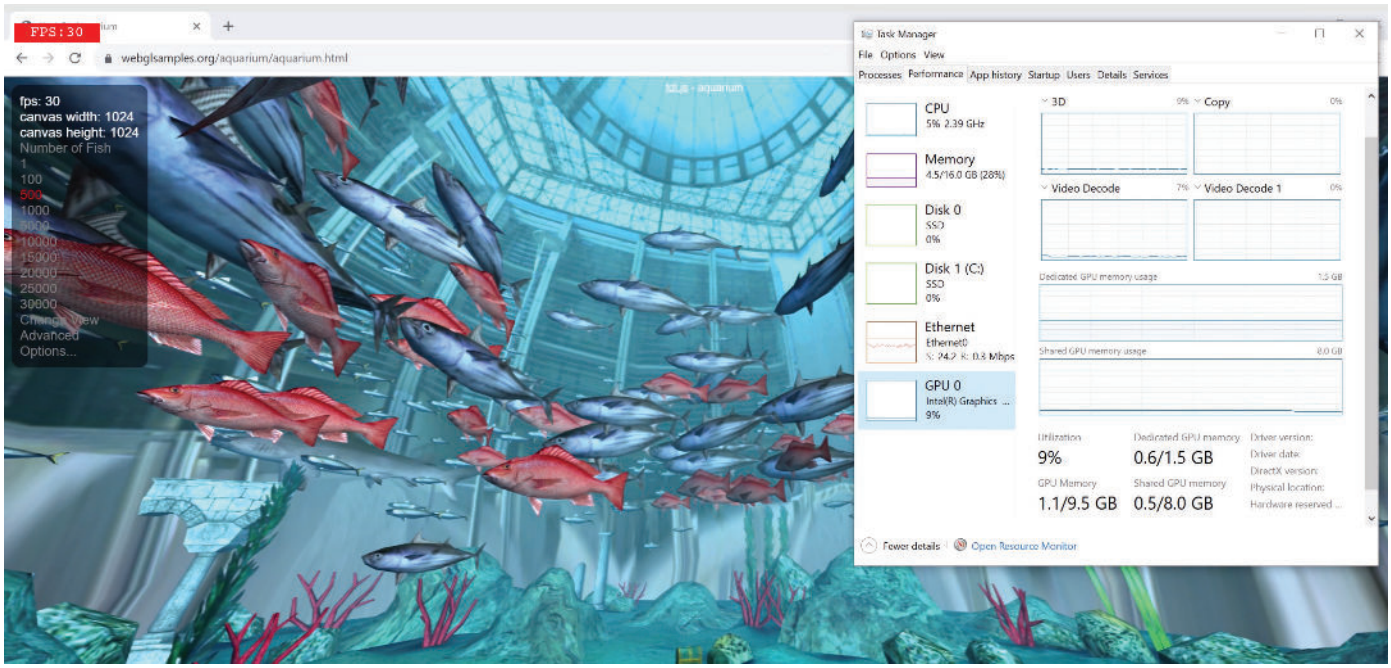


Figure 7. vGPU assigned VM delivering a VDI session running WebGL

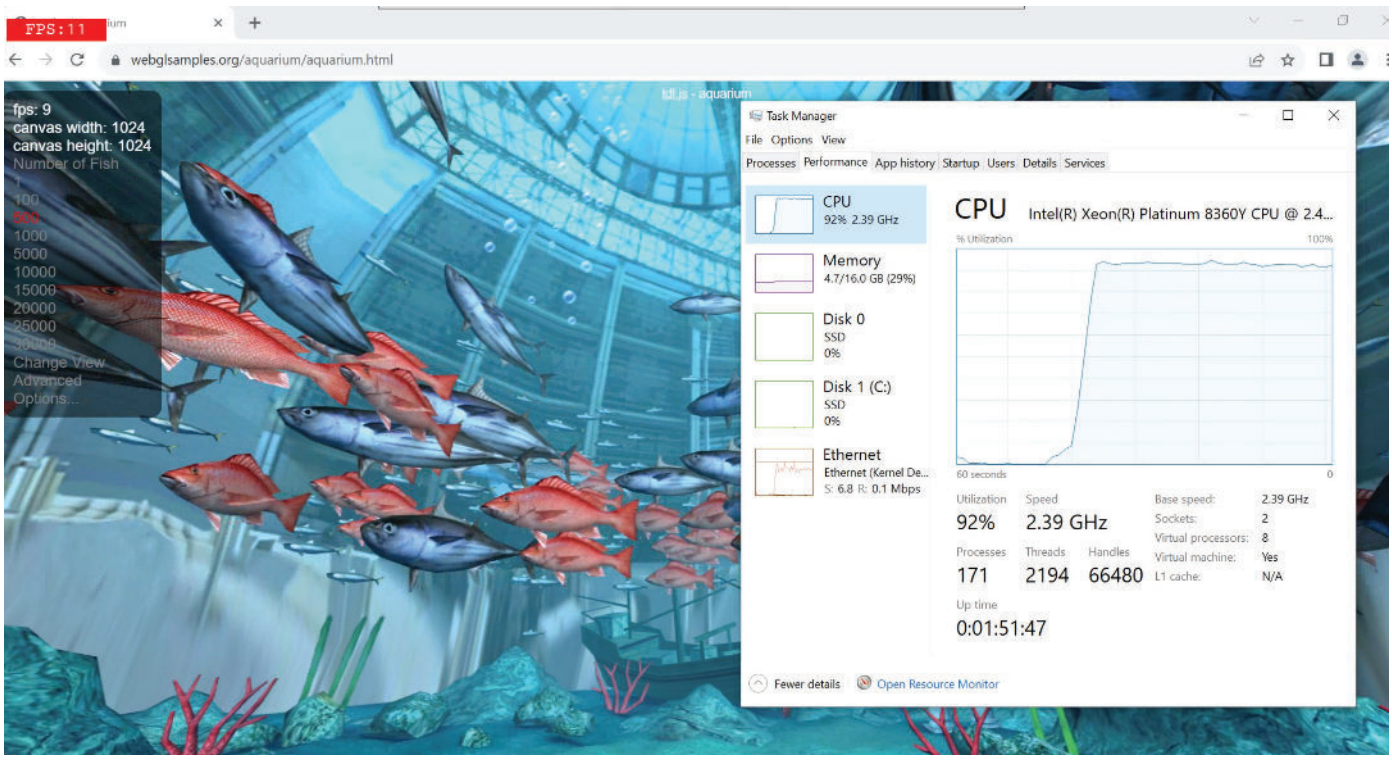


Figure 8. CPU only VM delivering a VDI session running WebGL

Key takeaway

Initial results indicate that virtual machines with Flex Series 140 vGPU deliver high quality graphics performance, consistent FPS of remoted frames and latencies in addition to significant CPU resource offload (~80% in 8vCPU VM) depending on the workloads. Ongoing testing demonstrates great end-user-experience of each VDI session at higher density of VDI sessions per server using lower vCPU configurations per VM.

Virtualization Overhead

Intel Flex Series GPU SR-IOV Technology grants each Virtual Function (VF) independent access to GPU resources via a VF driver. The submission of workloads from each VF driver is controlled by a GPU microcontroller called GuC. It supports a flexible and programmable scheduling schema. Each VF is granted time division multiplexed access to the entire GPU for a configurable duration of time before its contexts get preempted and switched out in favor of the workload contexts of the next VF in the round-robin.

The scheduler can be dynamically programmed to provide fixed QoS scheduling to each VF which guarantees dedicated GPU Time Quanta for each VF (whether active or inactive), or flexible scheduling optimized for GPU Utilization where each VF can concede some or all its allotted time quantum if it does not have any work to do.

Figures 9 and 10 depict the performance of SRIOV for the Knowledge Worker workloads of PCMark10 on scaling from single VM (V1 profile) to Multi-VM (V3 and V6 profile) configuration. Notice that in the multi-VM scenario, each VM produces a score that is quite comparable to single VM case which results in the cumulative score of ~3x, ~6x for V3 and V6 profiles respectively when comparing them with the V1 profile score.

PCMark10: HW Encode (H264)	V1 Profile	V3 Profile				V3 Cumulative/ V1 Performance
	VM1	VM1	VM2	VM3	V3 Cumulative	
App startup	9316	9610	9150	9608	28368	3.05
Video Conferencing	7436	7228	7094	7282	21604	2.91
Web Browsing	7491	7424	7385	7533	22342	2.98
Spreadsheets	4211	4180	4169	4135	12484	2.96
Writing	5424	5335	5283	5222	15840	2.92

Figure 9. Comparing 3VF vs 1VF performance

PCMark10: HW Encode (H264)	V1 Profile	V6 Profile							V6 Cumulative/ V1 Performance
	VM1	VM1	VM2	VM3	VM4	VM5	VM6	V6 Cumulative	
App startup	9316	9375	9061	9440	9337	9449	9155	55817	5.99
Video Conferencing	7436	6913	6974	7021	7032	7112	6938	41990	5.65
Web Browsing	7491	7531	7482	7312	7476	7636	7457	44894	5.99
Spreadsheets	4211	4149	4164	4132	4138	4113	4144	24840	5.90
Writing	5424	5329	5213	5214	5113	5264	5220	31353	5.78

Figure 10. Comparing 6VF vs 1VF performance

vGPU scheduler switching overhead i.e., the latency to de-schedule the current VF from the GuC software queue and resubmit workloads of the next VF is ~30 microseconds. This is trivial when compared to typical context execution time on Flex Series 140 GPU as indicated by the table to the right.

Application	Context Execution Time (msec)
1080p WebGL Render	0.3 - 1.0
1080p Media Decode	0.8 - 1.5
1080p Media Encode	1.0 - 3.0

Key Takeaway

The tested performance of Intel Flex Series GPU SR-IOV virtualization scales linearly with increasing number of Virtual Functions for typical knowledge worker VDI use cases. The overhead of vGPU Time-Division Multiplexing is minimal and it is observed in extended validation of VDI use cases that this accounts for <1% of the overall time under test.

Encode Performance

End user experience of VDI sessions predominantly depends on the number of remoted frames (fps) at the client-side and its consistency across the duration of the workload set.

Figures 11 and 12 plot the moving average ratio (in %) of Encoder FPS (client-side) to the framebuffer capture FPS (server-side) as observed on two different vGPU profiles (V1 and V6). This ratio helps indicate if GPU hardware encoding can keep up with the GPU rendering of framebuffers. The ideal value for this ratio should be >=100%. The moving average curve indicates low coefficient of variation which translates to good consistency in user-experience as well.

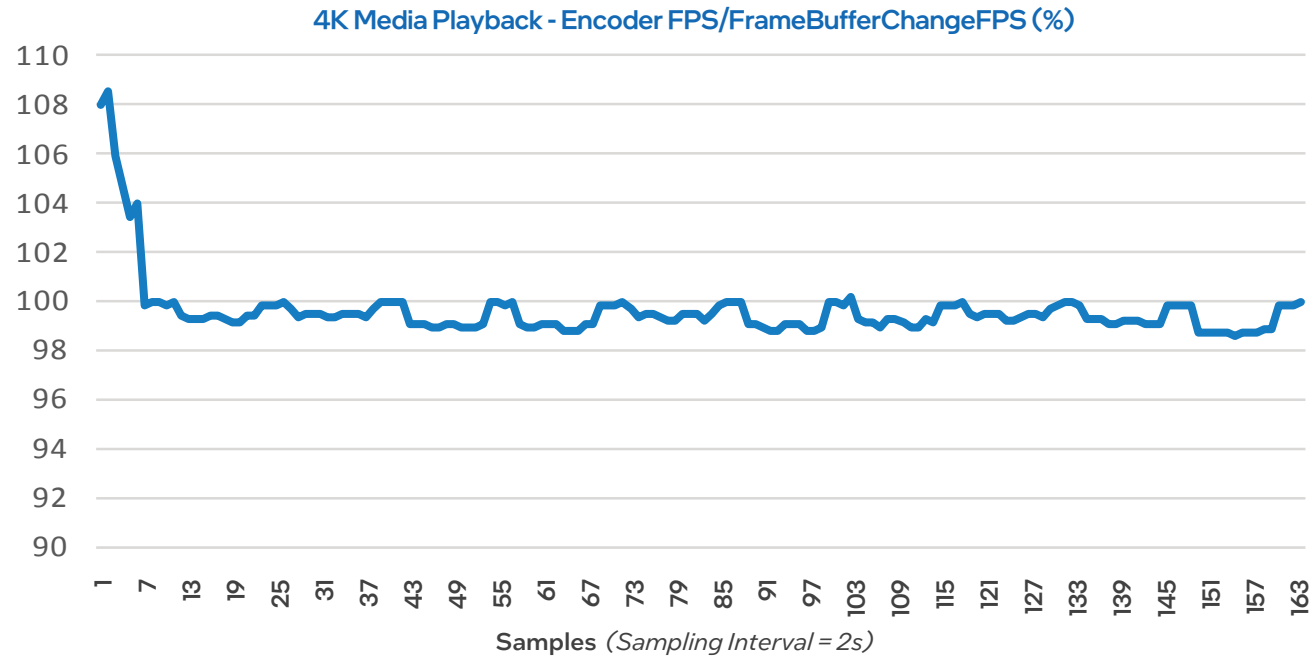


Figure 11. V1 profile – 4K @ 30 fps VDI session

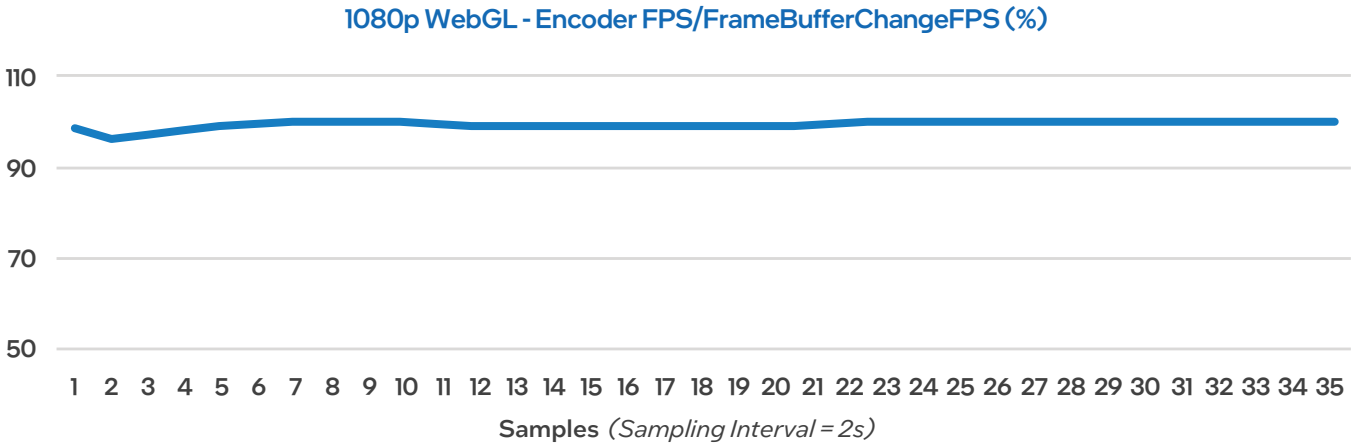


Figure 12. V6 profile – 1080p @ 30 fps VDI session

The latency observed in encoding each captured frame (on server side) is another key factor for end user experience. Figures 13 and 14 show consistent latency of ~6msec and 2.5msec for 4K and 1080p VDI sessions, respectively. In contrast, for a VM configuration with 4vCPUs and no GPU (see Figures 15, 16) the Encode latencies for 4K and 1080p VDI sessions running basic WebGL content are ~53msec and ~18msec. Do note that the latency charts also indicate a higher standard deviation in the encoder latency which translates to inconsistent end user experience.

Display Resolution of VDI session	Encode Latency VM config A: 4 vCPUs, 8GB RAM with Flex vGPU	Encode Latency VM config B: 4vCPUs, 8GB RAM – no vGPU	Config A vs Config B
1080p	~2.5msec	~18msec	~0.14x
4K	~6msec	~53msec	~0.11x

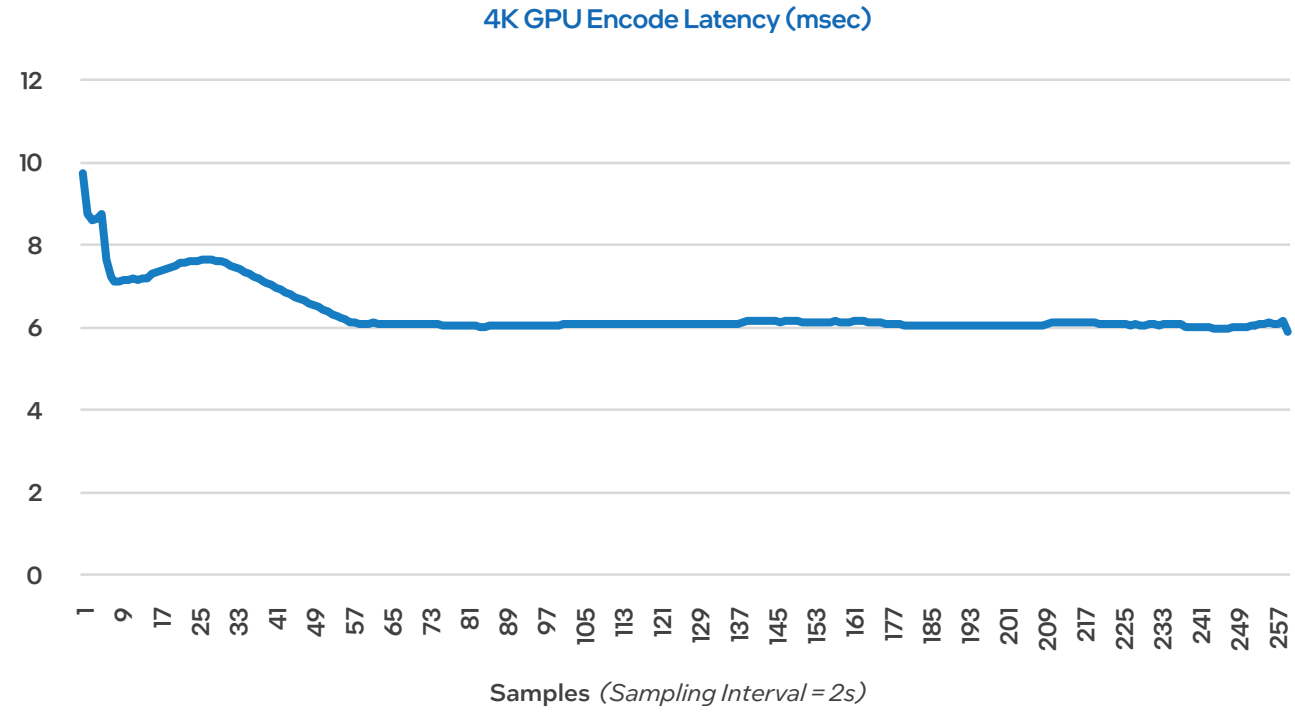


Figure 13. V1 profile – 4K @ 30 fps VDI session

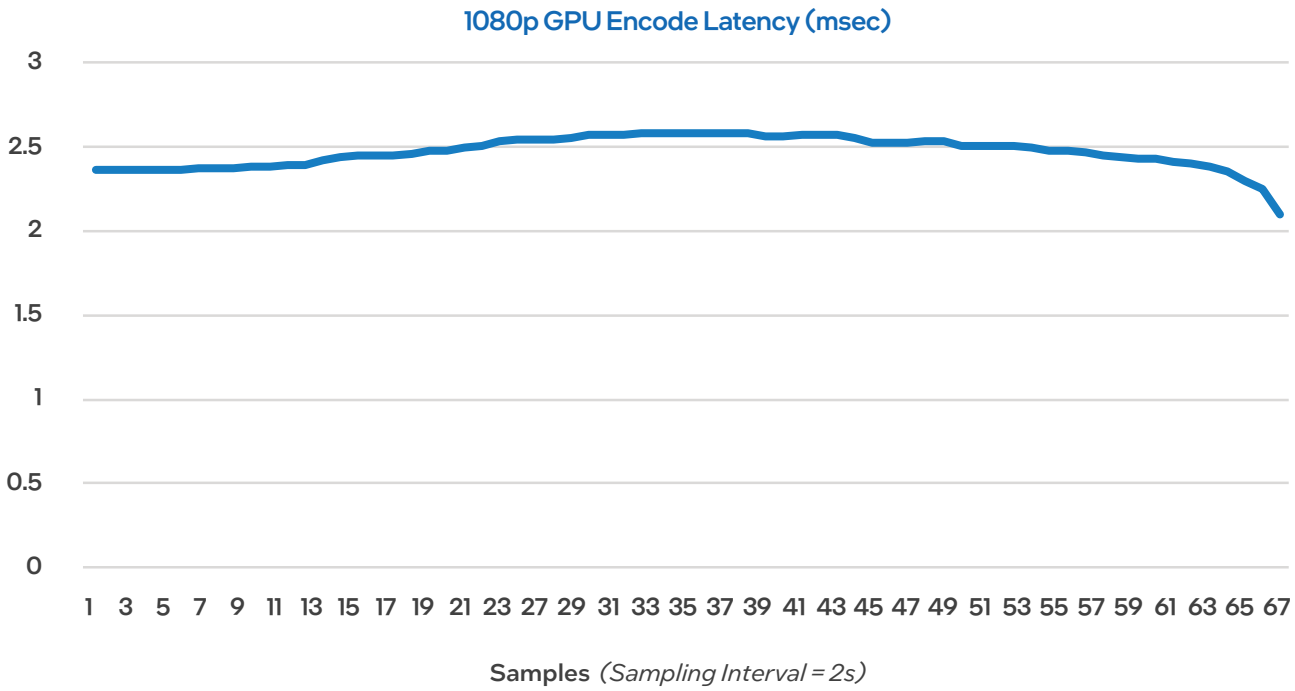


Figure 14. V6 profile - 1080p @ 30 fps VDI session

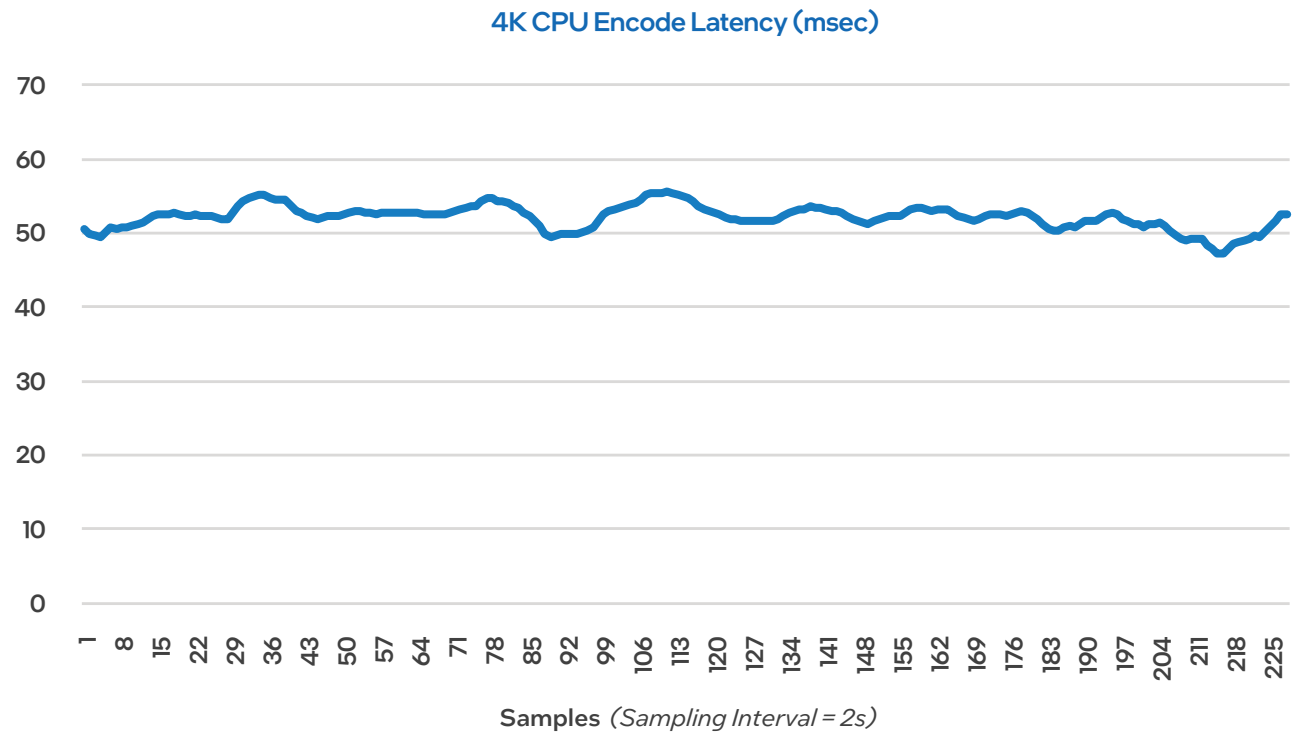


Figure 15. CPU-only (4vCPUs) VM – 4K @ ~4 fps VDI session

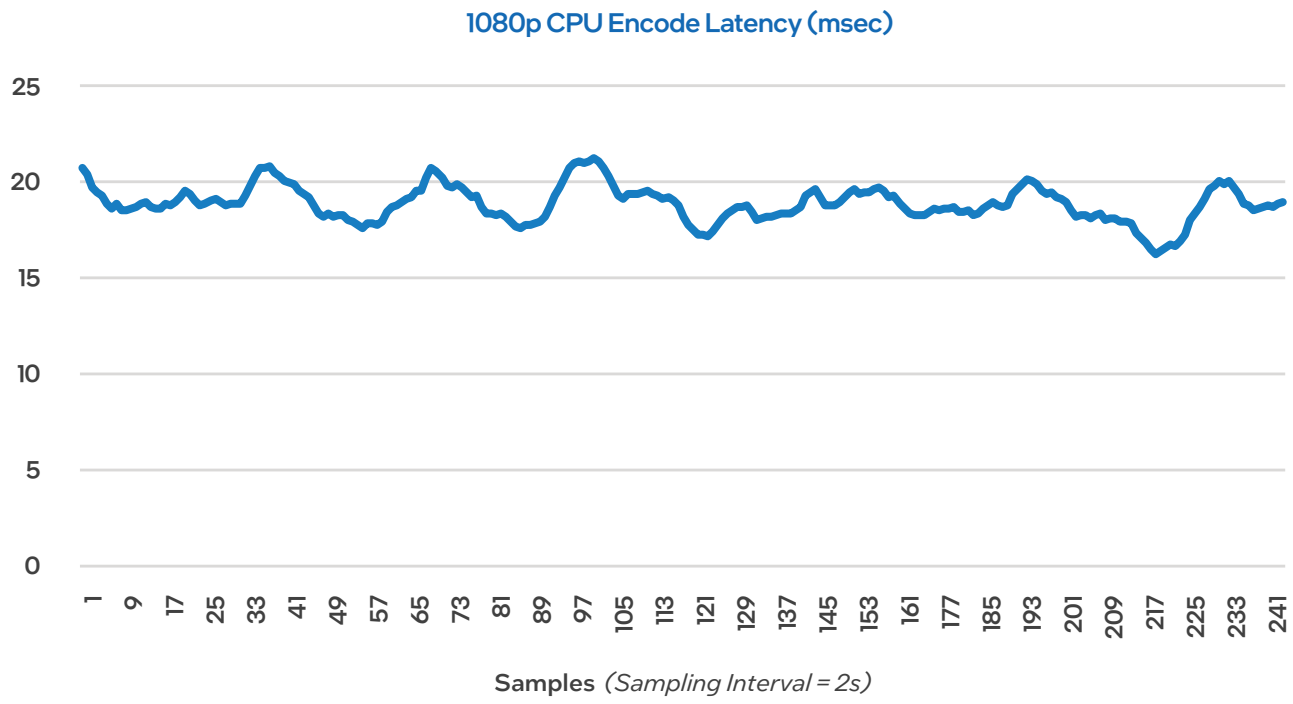


Figure 16. CPU-only (4vCPUs) VM – 1080p @ ~6 fps VDI session

In our testing, the client-side receives on average ~6fps for 1080p and ~4fps for 4K VDI sessions running WebGL application at close to 100% CPU utilization. Also, the client-side FPS highly depends on the type of workload that is currently executed – media playback ~13fps, WebGL ~6fps.

Display Resolution of VDI session	Encode FPS VM config A: 4 vCPUs, 8GB RAM with Flex vGPU	Encode FPS VM config B: 4vCPUs, 8GB RAM – no vGPU	Config A vs Config B
1080p	30fps	~6fps	~5x
4K	30fps	~4fps	~7.5x

Additional experiments show that a VM with 2vCPUs with no vGPU attached is unusable with the client-side observing ~2/3fps depending on the applications. Assigning a vGPU to the same VM configuration completely transforms the user experience and 1080p VDI sessions can be streamed to the client at 30fps.

Apps	Client FPS, CPU Utilization VM config: 2 vCPUs, 8GB RAM – with Flex vGPU	Client FPS, CPU Utilization VM config: 2 vCPUs, 8GB RAM - no vGPU	Client FPS, CPU Utilization VM config: 4vCPUs, 8GB RAM - no vGPU	Client FPS, CPU Utilization VM config: 8vCPUs, 8GB RAM - no vGPU
WebGL	30fps, 10%	2fps, 100%	6fps, 98%	13fps, 95%
1080p media	30fps, 7%	4fps, 100%	13fps, 99%	30fps, 94%

Key takeaway

As per existing data, the end user experience of VDI sessions without vGPUs is shown to be inconsistent, highly workload dependent and in lower end 2vCPU configurations, nearly unfeasible for modern knowledge worker scenarios. CPU-only VDI sessions deliver lower client-side frame rates at higher encode latency and do not scale well to larger virtual display resolutions. 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 yet still 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.

Virtual GPU Density

Flex Series 140 GPUs support up to 12 VFs per GPU card (~1GB GPU Local Memory per VF). It also offers administrators the flexibility to increase the per VF memory allocation and assign different amounts of GPU memory to each user on a single Flex Series 140 card.

From a VDI solution perspective, there is no artificial limitation imposed on virtual display per VDI session. This is solely determined by the capabilities of the VDI remoting agent and its usage of Intel® oneAPI Video Processing Library (oneVPL). The maximum display resolution and number of concurrent displays that can be supported per VDI session depends on the availability of two key GPU resources: 1) GPU Local Memory 2) Encode capacity (i.e., GPU time slice).

Key Takeaway

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 30fps as determined by the VDI remoting agent.

GPU Utilization

During the evaluation of VDI use cases on Intel Flex Series 140 GPU, it is observed that that the overall GPU Engine Utilization required to drive six 1080p VDI sessions per card hovers around ~18% (see Figure 17). Additionally, to deliver twelve concurrent 1080p VDI sessions per card, the engine utilization increases to 30-35% (see Figure 18). Local memory utilization is ~35% in the V3 profile, and ~60% in the V6 profile.

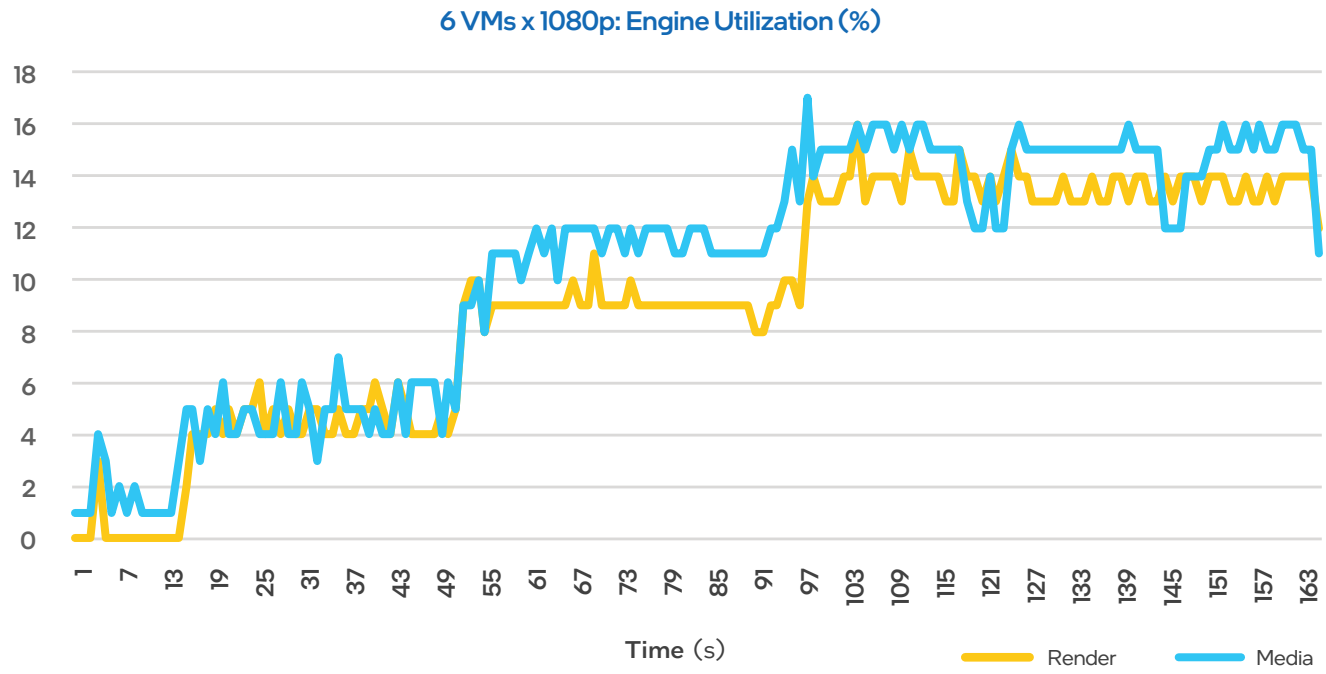


Figure 17. 6x1080p VDI sessions (V3 profile) – GPU Engine Utilization

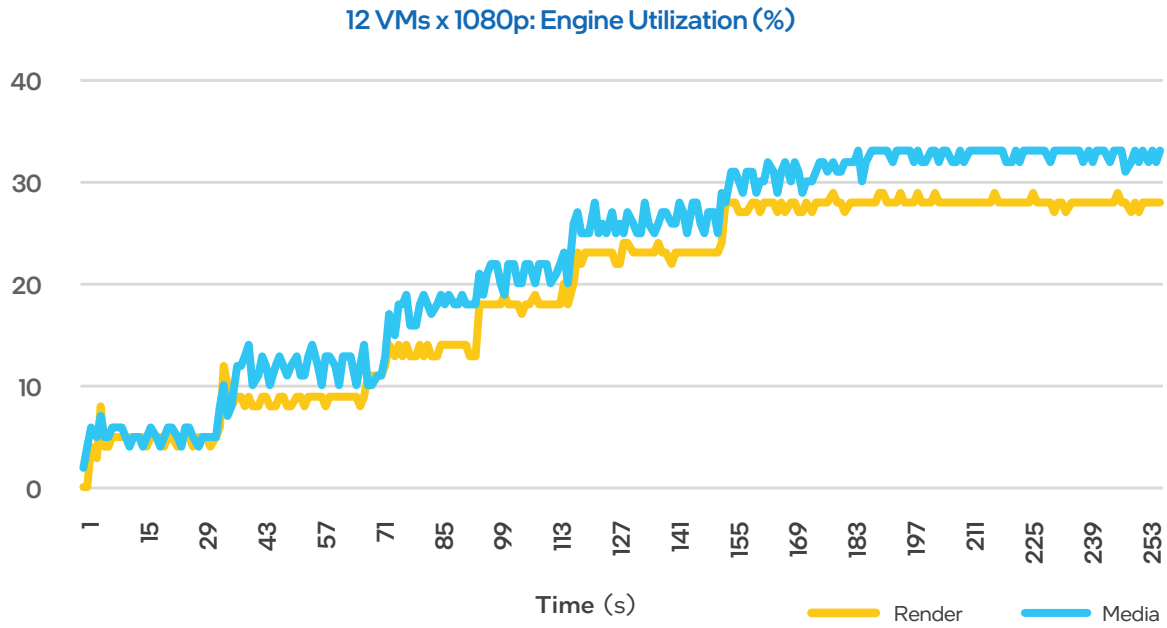


Figure 18. 12x1080p VDI sessions (V6 profile) – GPU Engine Utilization

Key takeaway

Intel Flex Series GPUs will offer flexible GPU resource assignment to VFs via SRIOV provisioning to ensure that adequate amount of GPU resources can be assigned to each VF whilst aiming to keep per VM and overall GPU utilization to an optimum level with sufficient resource headroom. 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 12x1080p sessions per card at just 30-35% GPU utilization and 60% Local Memory Utilization. Additional experiments reveal that the Flex 140 GPU can accommodate concurrent GPU-focused applications across various display resolutions.

Conclusion

The results highlighted in this whitepaper demonstrate that Intel® Data Center GPU Flex Series will deliver reliable performance for VDI use cases, especially for the high-density knowledge worker segment. Superlative graphics user experiences can be delivered to the end user with consistent framerates, low render and encode latencies at high quality whilst supporting modern display topologies. Each VDI session with a vGPU demonstrated sufficient GPU resource headroom, substantially reduced CPU utilization, thus improving VM density, the potential Total Cost of Ownership, and scalability of each VDI server.

SR-IOV based GPU virtualization will allow provisioning of flexible vGPU configurations at zero licensing costs. Intel Flex GPUs will support modifiable vGPU scheduling policies whereby the VDI administrator can dynamically rebalance the performance of all vGPUs while VMs and their VDI sessions are active.

Flex Series GPUs will offer strong encode performance per watt and will instantly transform a VDI data center in delivering highly performant, dependable, scalable, and future-proof VDI solutions at high user density.