SOLUTION BRIEF

SUPERIOR MEDIA PROCESSING AND DELIVERY SOLUTION BASED ON SUPERMICRO SERVERS W/ INTEL® DATA CENTER GPU FLEX SERIES

Supermicro Systems with Intel® Data Center GPU Flex Series

SYS-620C-TN12R

SYS-220BT-HNTR

SYS-420GP-TNR

SYS-530MT-H8TNR

TABLE OF CONTENTS

Executive Summary

The global health crisis of COVID-19 has played a significant role in driving the surge in video streaming, and this trend seems set to persist. Today, video is responsible for more than 80% of all internet traffic worldwide.

As the demand for this online video content continues to grow, so does the need for reliable and efficient transcoding and streaming infrastructure. In addition, consumers are getting accustomed to viewing broadcast-quality videos using a wide variety of devices with high resolutions. As a result, service providers must find new ways to optimize their existing solutions and ease the total cost of ownership; while meeting consumers' demands for more sophisticated content.

With Supermicro systems equipped with Intel® Data Center GPU Flex Series, providers can efficiently fit and scale to more subscribers with a smaller data center footprint, thereby reducing equipment and facility costs—without compromising quality.

The Importance of Transcoding in Video Streaming

Over the past few years, streaming quality has skyrocketed across industries due to the ongoing pandemic, increased accessibility to high-end video, and better bandwidth for Over-the-Top (OTT) and Video-on-Demand (VOD) delivery. Further, organizations increasingly use video to disseminate information internally and externally, relying on live-streaming media for webinars and company meetings. Cloud Gaming is another fast-growing segment in the video streaming domains, presenting service providers and game publishers with new business opportunities and rapid technological changes.

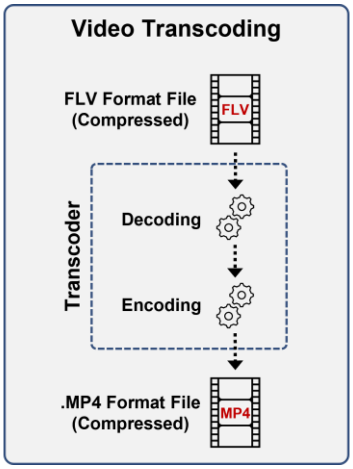


Figure 1 - Video Transcoding Workflow

As technology becomes more accessible, new streaming content providers, both large and small, are emerging. Media or video content, either on-demand streaming or live streaming, needs to be adjusted based on the device characteristics of viewers. The original video must be encoded or compressed to reduce the size of Raw video files. Also, to match the viewers' devices, the stream must be decoded or converted to a supported resolution, frame rate, video codec, and network bandwidth. This process is referred to as video transcoding, which minimizes bandwidth and delivery infrastructure usage.

Video transcoding is a core technique for streaming because it affects the streaming service for both the service provider and the users. The video service provider must decide how to transcode video content into multiple representations and store them, which further increases the operational cost of the service provider. Since video transcoding is computation-intensive and consumes considerable resources, it will significantly affect the service provider's operating cost.

Media Delivery Architecture and Workflow

To provide viewers with high-quality video content in real-time so much goes on behind the scenes. Media processing and delivery comprise three essential processes: video capture, video transcoding, and streaming delivery. Video content is initially captured and stored with a particular format, spatial resolution, frame rate, and bit rate. Then, the video is uploaded to streaming servers. Next, a streaming server must transcode the original video based on the client's network bandwidth, device resolution, frame rate, and video codec. Finally, video streams are distributed via CDNs server to end devices.

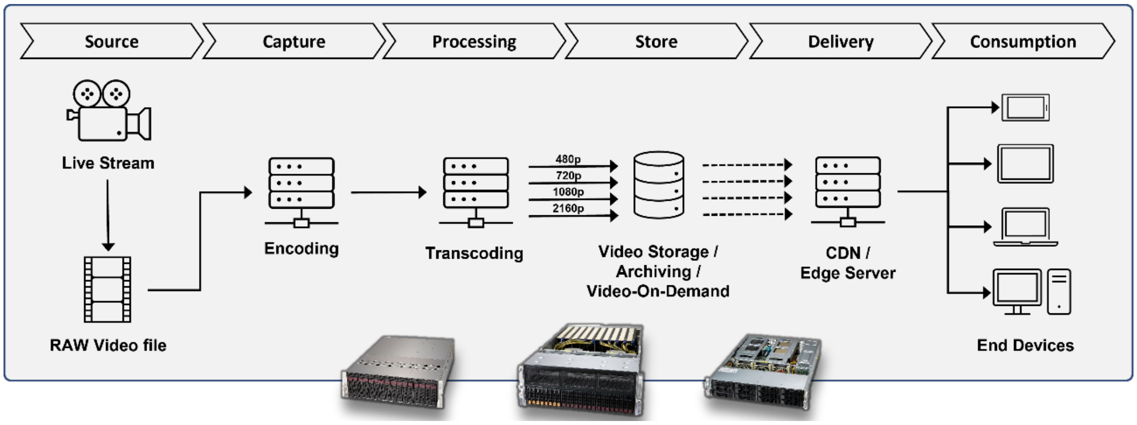


Figure 2 – Media Delivery Architecture and Workflow

The Challenges of Media Delivery

Rapid technological and network advances are improving content delivery but, at the same time, drastically increasing the streaming throughput. Therefore, a modern streaming solution infrastructure must be flexible and scalable. Determining the proper hardware to architect the media delivery workflow can take time and effort. Cloud service providers need to consider an extensible and easy-to-manage IT platform ecosystem, covering from the cloud to the edge.

The challenge of delivering video is aggravated by the users demanding access to their video streams anywhere, on any device. Videos must be transcoded effectively and efficiently to provide the desired user experience. Codecs are essential in delivering the best video quality while ensuring a smooth playback experience. Accordingly, a video streaming infrastructure requires leveraging the most advanced and new codecs, such as AV1, VP9, and HEVC.

New codecs allow service providers to deliver the same video quality with a smaller bitrate. However, these codecs are more compute-intensive than H264. With hundreds of millions of videos being uploaded by users every day, it would be prohibitively expensive for service providers to use CPU-only transcoding for all AV1 and HEVC videos. So, compute costs must be added as another dimension of an optimized and cost-effective solution. New hardware and technologies are demanded to support new streaming requirements.

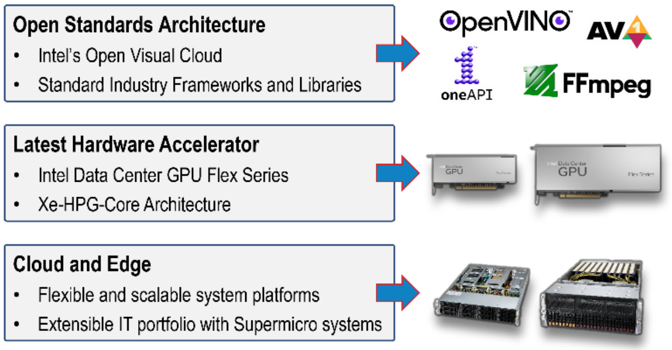
In addition to the above challenges, HEVC and h264 codecs include many patent-protected video compression techniques. Organizations must license the patents from their creator or representative to use these codecs. However, we're talking about several thousand patents from just a few dozen companies. Consequently, codec royalty payments are pricey and difficult to understand.

Key Challenges in Media Delivery Solutions	
Total Cost of Ownership	• Hard-and-fast hardware platforms; are expensive to manage and scale. • Hardware accelerators with high power consumption, low transcoding density, and high bit rate demands. • Siloed environments and high royalty fees.
User Experience	• High latency • Network and data congestion • Poor video quality

Supermicro Solutions for Media Processing and Delivery

Partnering with Intel®, Supermicro offers qualified platforms to get service providers ready to tackle the Media Processing and Delivery challenges.

- An extensible IT portfolio offering fitting platforms for the cloud and the edge.
- Latest accelerator support, offering advanced codecs and technologies.
- Open standard architecture with standard industry frameworks and libraries.



Open Standards Architecture
- Intel's Open Visual Cloud
- Standard Industry Frameworks and Libraries

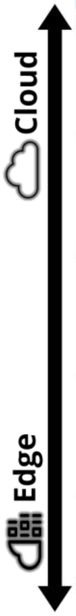
Latest Hardware Accelerator
- Intel Data Center GPU Flex Series
- Xe-HPG-Core Architecture

Cloud and Edge
- Flexible and scalable system platforms
- Extensible IT portfolio with Supermicro systems

Figure 3 – Media Solution Platform Stack

Extensible IT Portfolio with Supermicro Systems

Deciding whether a streaming media workflow should use cloud servers, edge servers, or a mix of both comes down to several factors. Encoding and transcoding can span across the streaming platform depending on specific workloads. For example, while the processing of live streaming for social networks can operate acceptably from the cloud, more latency-sensitive and high-performance workloads, such as game streaming (cloud gaming), should be placed at the edge.



Supermicro system		Intel Data Center GPU Flex Series	Target Workloads
3U MicroCloud **SYS-530MT-H8TNR**		**1 x Flex Series 140** (per node) **Total: 8 x GPUs per system**	Cloud Computing Web Hosting, VM Online Game Server Hosting Social Networking
4U GPU SuperServer **SYS-420GP-TNR**		**10 x Flex Series 140 or 170**	Maximum Transcoding Density High Performance Computing Rendering / VDI AI/Deep Learning Training
2U CloudDC **SYS-620C-TN12R**		**6 x Flex Series 140 or** **4 x Flex Series 170**	Visual Inferencing Optimized Web Server, Firewall Application Data Center Optimized, Value IaaS CDN, Edge Nodes
2U 4-Node BigTwin **SYS-220BT-HNTR**		**2 x Flex Series 140** (per node) **Total: 8 x Flex Series 140 per system**	Media Delivery Optimized All-Flash NVMe Hyperconverged Application Accelerator High-Performance File System
2U 2-Node SuperServer **SYS-210GP-DNR**		**3 x Flex Series 140 or 170** (per node) **Total: 6 x GPUs per system**	Cloud Gaming Optimized High Performance Computing AI/Deep Learning Training Media/Video Streaming

Figure 4 - Supermicro Systems for Media Processing and Delivery

With an extensible IT portfolio, Supermicro offers a choice of systems to build an optimized streaming platform. The systems provide a tradeoff between the maximum number of accelerators, rack density, and cost-effectiveness. Systems like the SYS-530MT-H8TNR, provide essential performance, low power consumption, and balanced workload density for the cloud. While systems equipped with the Intel® Xeon® Scalable Processors, such as the SYS-420GP-TNR and the SYS-220BT-HNTR, offer high performance and additional compute resources for critical workloads and optimized application acceleration.

In addition to having flexible, scalable, and power-efficient designed systems, Supermicro offers reliable and manageable systems. Featuring redundant power and cooling systems, Supermicro systems have proven reliability. Remote management, IPMI support, and Redfish 1.8 are standard, and the systems also support TPM 2.0 and Root-of-trust security.

Advanced Accelerators: Intel® Data Center GPU Flex Series

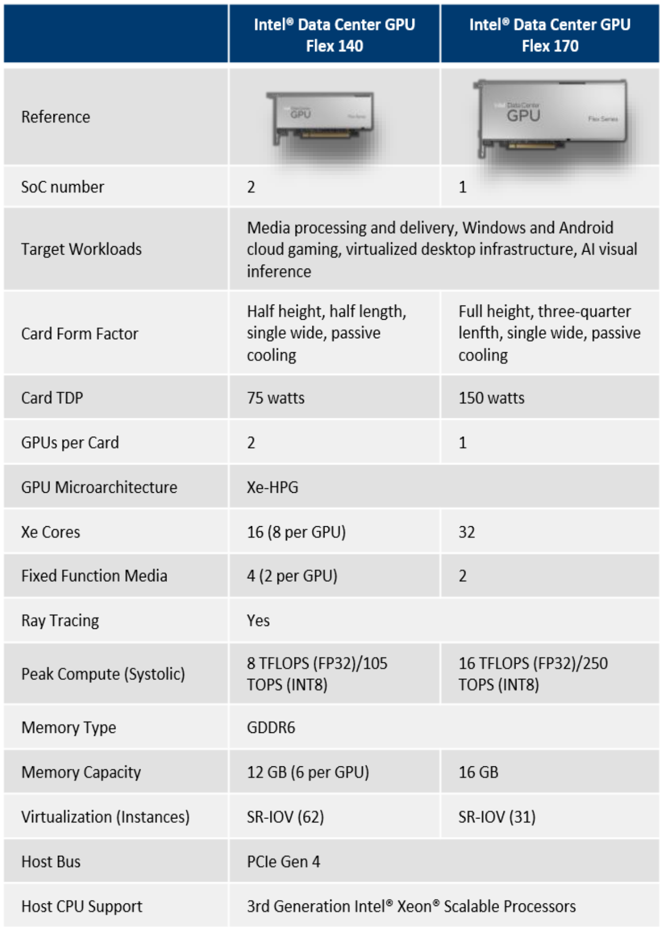
The development of GPUs in the last few years has progressed from specialized graphics chips to general-purpose computing devices. GPUs offer higher efficiency for parallelizable operations, allowing for higher compute per dollar compared to CPUs. This power consumption challenge is also faced by video streaming solutions.

The new Intel® Data Center GPU Flex Series performs high-performance and high-quality transcoding. The GPU instances deliver higher throughput, meaning a lower cost per video. In addition, it is possible to process the latency-sensitive workload faster and provide a better customer experience.

This accelerator has exciting new media features—royalty-free AV1 Hardware Encoding support and an advanced software bitrate controller to boost hardware encoding quality.

The Flex Series comes in two flavors: The Flex Series 140 and the Flex Series 170. These GPUs have four classes of video accelerator engines, ensuring typical transcode operations can facilitate pipeline execution to minimize latency. Multiple accelerator units allow the concurrent execution of various frames to maximize throughput.

Intel® has significantly innovated to develop these two data center GPU accelerators to provide the best operations and performance for Cloud Gaming and Virtual Desktop Infrastructure. Also, the AV1 codec has been demonstrated to provide lower bandwidth requirements to transmit higher quality streaming graphics. Being royalty-free versus HVEC codecs, the use of Intel Data Center GPUs significantly reduces costs for media delivery providers.



	Intel® Data Center GPU Flex 140	Intel® Data Center GPU Flex 170
Reference		
SoC number	2	1
Target Workloads	Media processing and delivery, Windows and Android cloud gaming, virtualized desktop infrastructure, AI visual inference	
Card Form Factor	Half height, half length, single wide, passive cooling	Full height, three-quarter lenfth, 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)	SR-IOV (62)	SR-IOV (31)
Host Bus	PCIe Gen 4	
Host CPU Support	3rd Generation Intel® Xeon® Scalable Processors	

Figure 5 - Intel(R) Data Center GPU Flex Series Specs

Open Standards Architecture

Intel's discrete graphics accelerators are well integrated into open-source media frameworks such as FFmpeg and Intel®'s oneAPI Video Processing Library (oneVPL). These popular frameworks allow both complex pipeline support and extreme customization of accelerator control. In addition, to make these tools even more accessible to Linux developers, Intel provides build scripts in Docker on the latest Linux kernels.

Media transcoding performance is optimized across integrated and discrete GPUs with Intel® oneVPL. In addition to video processing and delivery, oneVPL provides encoding, decoding, and streaming APIs for applications, including broadcasting, streaming, video-on-demand, and cloud gaming.

CDNs and other delivery providers continue to struggle with high delivery costs, despite the decreasing cost of large-scale data storage. Improved compression helps the media processing and delivery providers reduce operating costs.

AV1, a next-generation codec, is built into the Flex Series GPUs, bringing the highest quality real-time video, scalable to any modern device at any bandwidth. AV1 is a royalty-free codec that delivers commercial or non-commercial user-generated content with a low computational footprint optimized for internet streaming. In addition, the streaming quality is not compromised, and the cost per stream is reduced by 30% with no degradation in compression.

AVC, HEVC, and VP9 codecs are also supported along with AV1. It is possible to access these codecs using standard frameworks such as FFmpeg or Gstreamer, or with oneVPL, which allows access to more controls and parameters. With these encoders, providers can adjust TCO-related parameters based on performance/quality presets.

Optimized Platform with Supermicro and Intel®

To meet subscriber demands for more sophisticated content, media processing and delivery providers must optimize their TCO. Supermicro and the Intel Flex Series GPU support that purpose by increasing the density of streams supported per server without compromising quality.

Supermicro is first-to-market with solutions for Intel Data Center GPU Flex Series, so early testing for workload has been possible. Following Intel's guidelines, video streams with 1080p and 4K resolutions were benchmarked using FFmpeg. Further, using optimized Intel presets for video transcoding, H.264, HEVC, and AV1 codecs were tested.

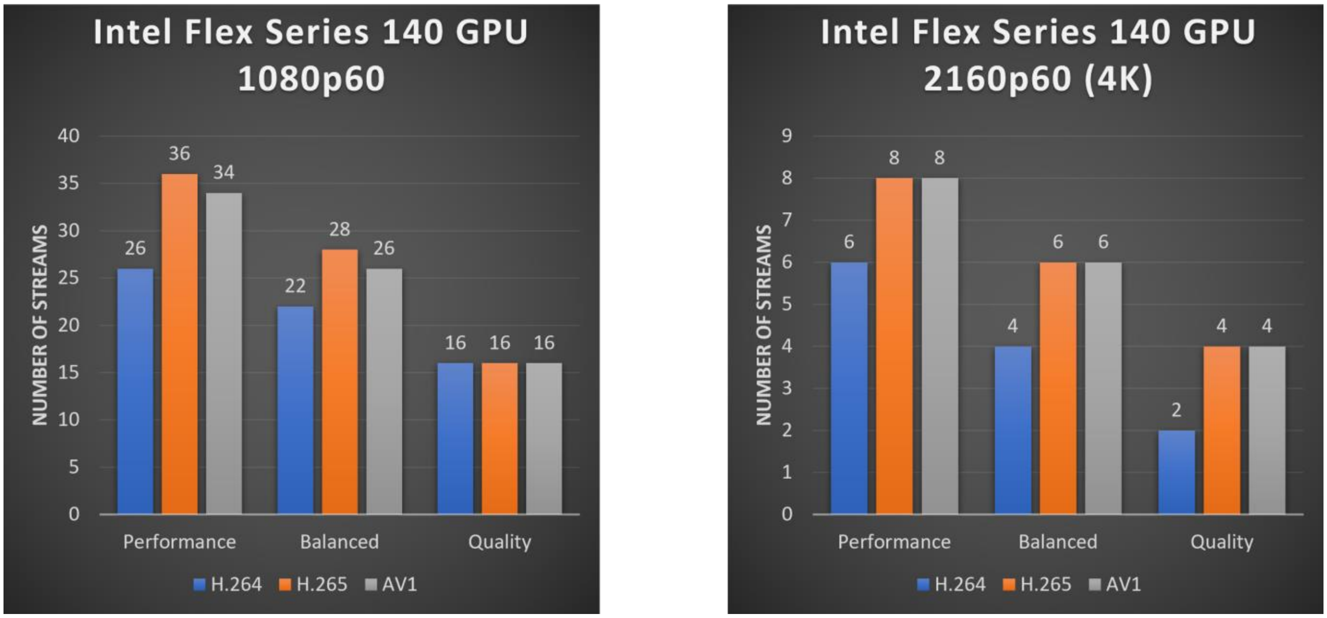


Figure 7 - Intel Flex Series GPU Transcoding Performance

As a result, a single Intel Flex Series 140 GPU can support as many as eight simultaneous 4Kp60 or 36 1080p60 streams per card. The accelerator provides this exceptional performance while keeping its power below 70 watts. The partnership between Supermicro and Intel allows for a range of configurations. For example, populating SYS-the 420GP-TNR systems with ten cards, 80 4Kp60 or 360 1080p60 streams can be delivered. More configurations are available for different Supermicro systems, giving service providers the right choice and allowing them to reduce their data center footprints and expenses associated with equipment and facilities.

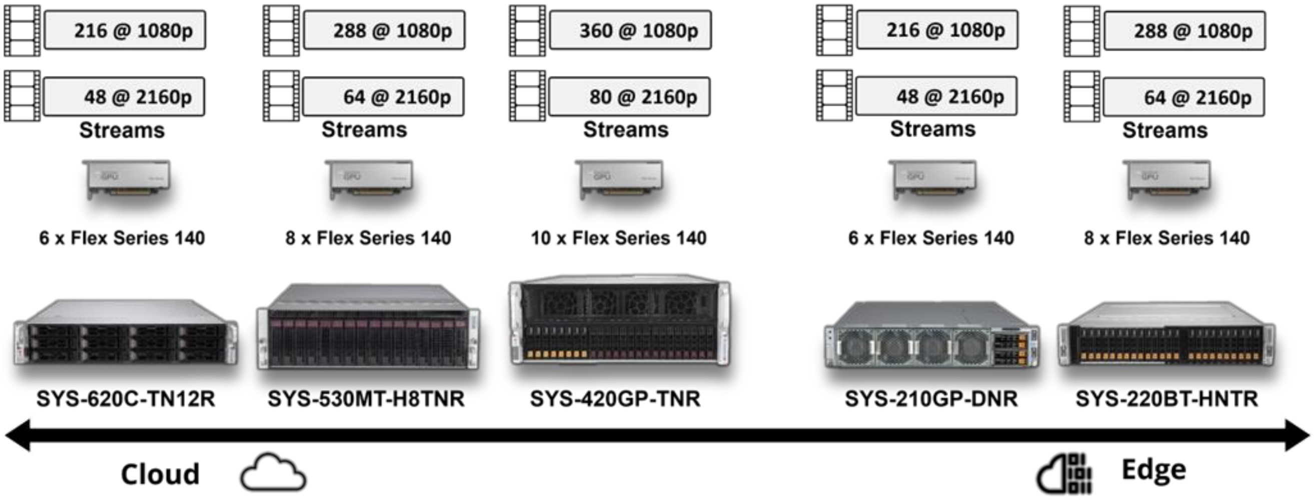


Figure 6 - Streaming density for various Supermicro systems with the Intel(R) Flex Series GPUs

Conclusion

Supermicro and the Intel Flex Series GPUs let providers serve more subscribers with a smaller data center footprint, reducing the costs associated with equipment and facilities. In addition, reducing the high performance per watt reduces the total cost of ownership (TCO).

Please contact your Supermicro sales representative for more information.

For More Information, please visit:
https://www.supermicro.com/en/accelerators/intel
https://www.supermicro.com/en/products/rackmount
https://www.intel.com/content/www/us/en/products/docs/discrete-gpus/data-center-gpu/flex-series/overview.html