White paper
Cisco public

Media Analytics Benchmark Performance on Cisco UCS X210c M7 with X440p PCIe Node Featuring Intel Data Center GPU Flex Series

December 2023

Contents

Executive summary

Cisco reports that globally, business and consumer video accounts for 80 percent of all internet traffic. The demand for more sophisticated content produces upward pressure on costs to providers for both processing infrastructure and delivery bandwidth. Hence, high processing performance plays a pivotal role in meeting cost requirements by driving up the density of streams per server.

Intel Data Center GPU Flex Series is a general-purpose data-center graphics processor optimized for media stream density and quality, with server-class reliability, availability, and scalability.

In this paper, our primary goal is to highlight the performance of the Cisco UCS® X440p PCIe GPU node configured with two Intel® Data Center GPU Flex 140 cards for media analytics usage models. A Cisco UCS X440p PCIe GPU node paired with a Cisco UCS X210c Compute Node was installed on a Cisco UCS X9508 Chassis.

Performance data shown in this white paper is obtained using the Intel Deep Learning Streamer Pipeline Zoo framework. The pipe bench command line utility is used to measure the performance of a pipeline under different scenarios (object detection, object classification, decoding, etc.).

Overview of Cisco UCS X-Series Modular System

The Cisco UCS X-Series is designed with modularity and workload flexibility at top of mind. The Cisco UCS X-Series Modular System begins with the Cisco UCS X9508 Chassis engineered to be adaptable and future ready. It is an open system designed to deploy and automate faster in concert with a hybrid-cloud environment.

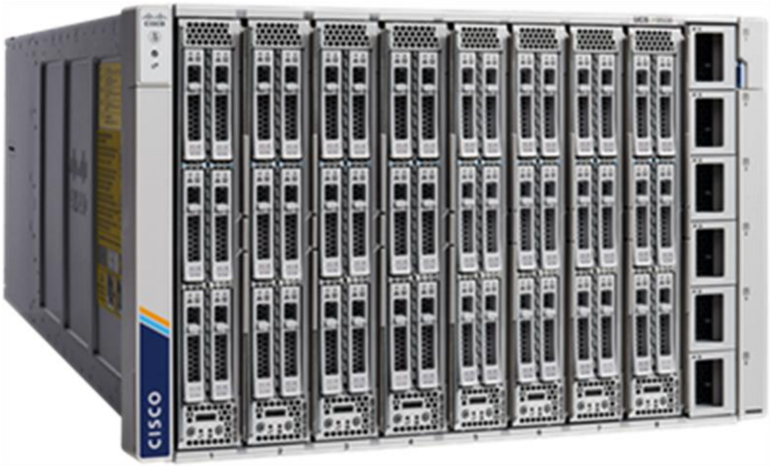


Figure 1.
Cisco UCS X9508 Chassis front view (populated)

The Cisco UCS X-Series Modular System simplifies your data center, adapting to the unpredictable needs of modern applications while also providing for traditional scale-out and enterprise workloads. It reduces the number of server types to maintain, helping to improve operational efficiency and agility as it helps reduce complexity. Powered by the Cisco Intersight® cloud-operations platform, it shifts focus from administrative details to business outcomes with hybrid-cloud infrastructure that is assembled from the cloud, shaped to your workloads, and continuously optimized.

Cisco UCS X210c Compute Node

The Cisco UCS X210c M7 Compute Node is the second generation of compute node to integrate into the Cisco UCS X-Series Modular System. It delivers performance, flexibility, and optimization for deployments in data centers, in the cloud, and at remote sites. This enterprise-class server offers market-leading performance, versatility, and density without compromise for workloads. Up to eight compute nodes can reside in the 7-Rack-Unit (7RU) Cisco UCS X9508 Server Chassis, offering one of the highest densities of compute, I/O, and storage per rack unit in the industry.

With the 4th Gen Intel Xeon® Scalable Processors, with 50 percent more cores per socket over previous generations, and advanced features such as Intel Advanced Matrix Extensions (AMX) and In-Memory Analytics Accelerator (IAA), many applications will see significant performance improvements with the Cisco UCS X210c M7 Compute Node.

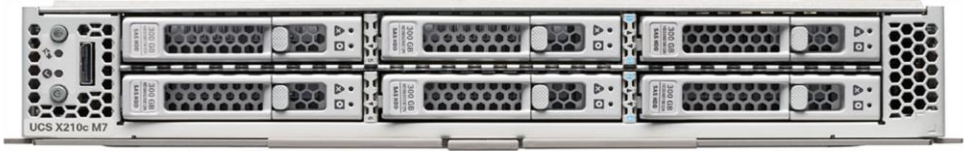


Figure 2.
Front view of Cisco UCS X210c M7 Compute Node

Cisco UCS X440p PCIe Node

With the introduction of PCIe nodes and X-Fabric technology, the UCS X-Series now supports workloads that require GPUs. The Cisco UCS X-Fabric Technology solution is a combination of two products: the Cisco UCS X9416 X-Fabric Module, which provides a PCIe Gen 4 fabric, and the Cisco UCS X440p PCIe Node, which hosts the GPUs. PCIe expansion nodes connected to compute nodes through Cisco UCS X-Fabric Technology can support applications such as artificial intelligence, machine learning, and virtual desktop infrastructure. This unique architecture dramatically simplifies the addition, removal, or upgrade of GPUs to compute nodes. Now you can easily and independently manage the different lifecycles of CPU and GPU components.

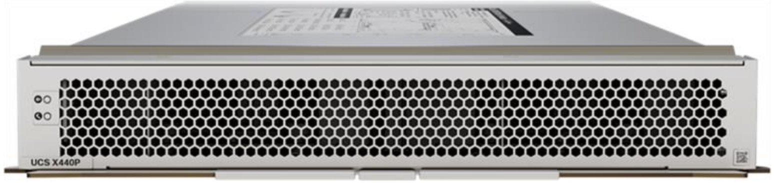


Figure 3.
Front view of Cisco UCS X440p PCIe Node

The Cisco UCS X440p PCIe Node supports both Intel and NVIDIA GPUs. Supported Intel GPUs options include:

- Up to four Intel Data Center GPU Flex 140s
- Up to two Intel Data Center GPU Flex 170s

The Cisco UCS X9508 Chassis has no midplane. This innovative design provides fewer obstructions for better airflow. The vertically oriented X210c compute nodes and the X440p PCIe nodes connect directly to horizontally oriented X-Fabric modules. This enables seamless upgrades to future compute nodes, resource nodes, and X-Fabric modules without requiring any forklift upgrades.

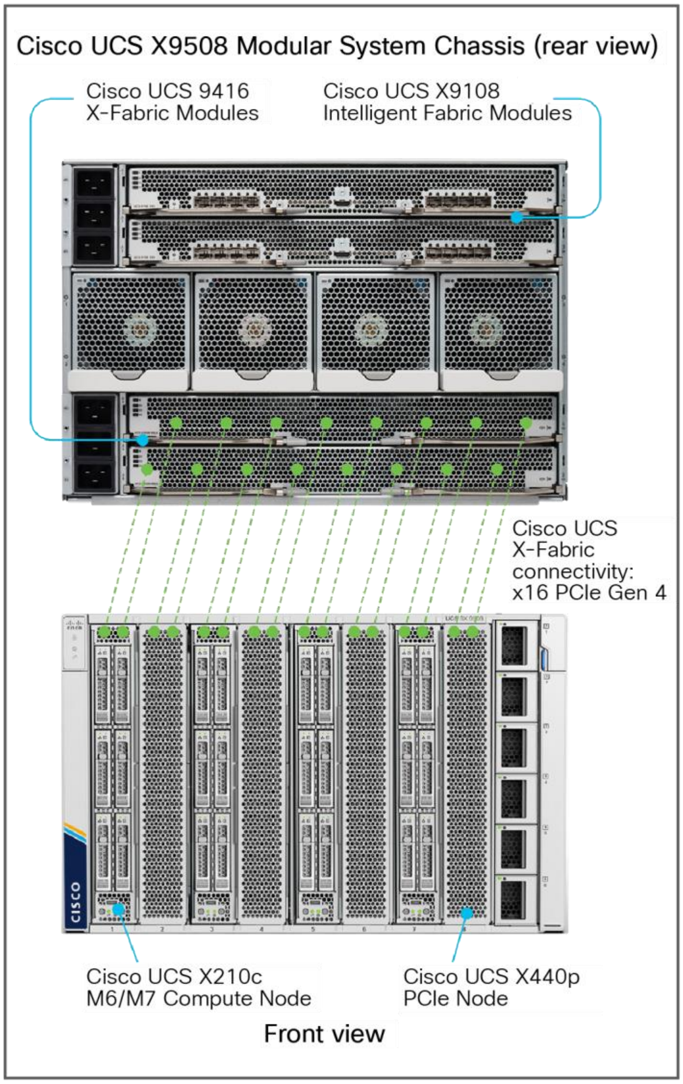

Cisco UCS X9508 Modular System Chassis (rear view)
Cisco UCS 9416 X-Fabric Modules
Cisco UCS X9108 Intelligent Fabric Modules
Cisco UCS X-Fabric connectivity: x16 PCIe Gen 4
Cisco UCS X210c M6/M7 Compute Node
Cisco UCS X440p PCIe Node
Front view

Figure 4.
Cisco UCS X-Fabric connectivity

The Cisco UCS X9508 Chassis has eight node slots, up to four of which can be X440p PCIe nodes when paired with Cisco UCS X210c or Cisco UCS X410c compute nodes. Please refer to **Figure 5**, below.

The Cisco UCS X440p PCIe Node allows you to add up to four GPUs to a Cisco X210c M6, X210c M7, or X410c M7 compute node with Cisco UCS X-Fabric Technology. This provides up to 16 GPUs per chassis to accelerate your applications. If your application needs even more GPU acceleration, up to two additional GPUs can be added using the optional GPU front mezzanine on a UCS X210c or UCS X410c compute node.

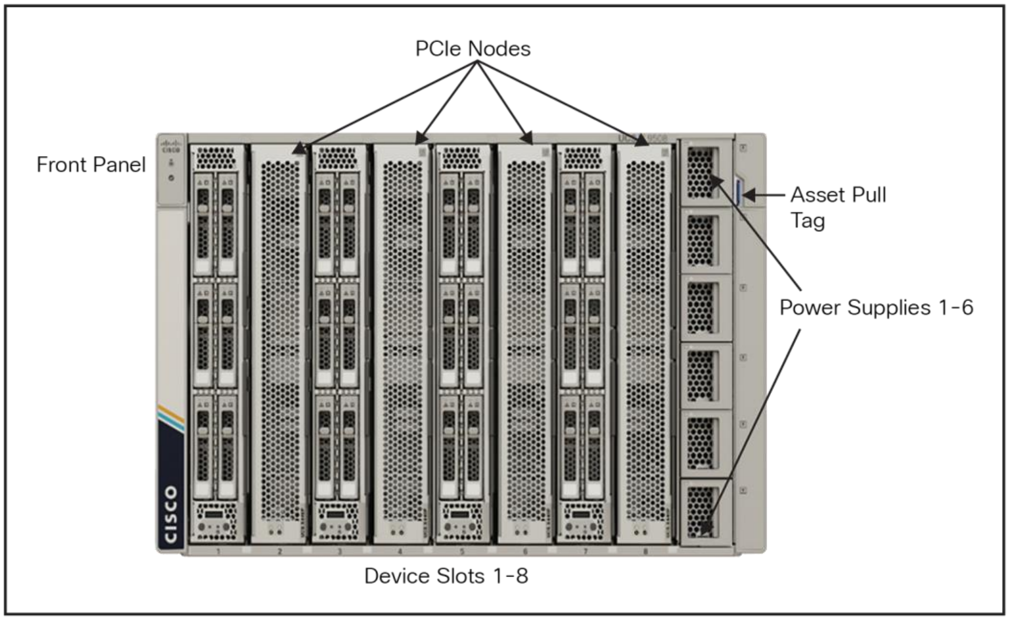


Figure 5.
Cisco UCS X9508 Chassis front view with PCIe Nodes (populated)

Introduction to media analytics

Media analytics is the analysis of audio and video streams to detect, classify, track, identify, and count objects, events, and people. The analyzed results can be used to take actions, coordinate events, identify patterns, and gain insights across multiple domains: retail store and events facilities analytics, warehouse and parking management, industrial inspection, safety and regulatory compliance, security monitoring, and many more.

Media analytics pipelines transform media streams into insights through audio and/or video processing, inference, and analytics operations across multiple IP blocks.

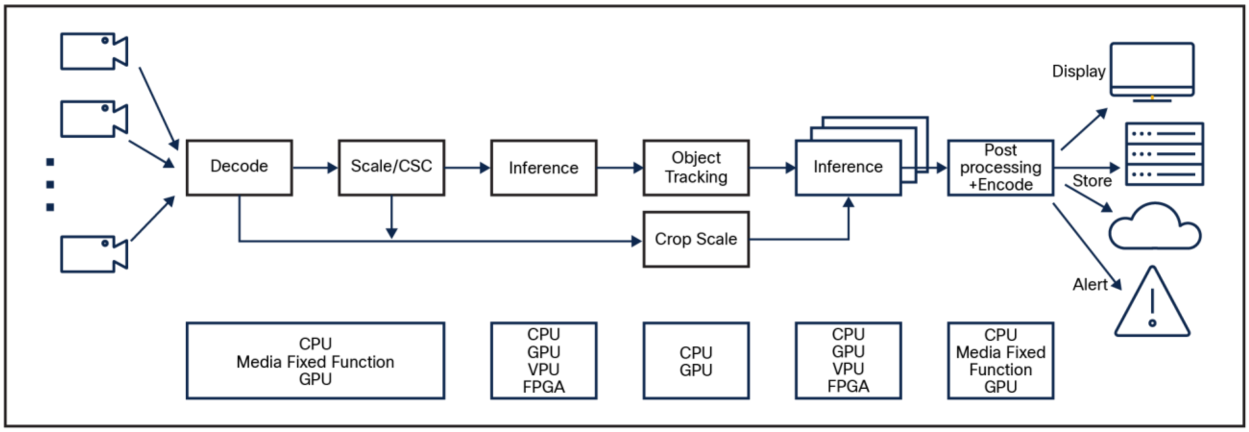


Figure 6.
Example of a media analytics use case

As an illustration, in Figure 6, above, video (also referred to as media) originates from a camera feed on the left. The camera feed initially undergoes decoding, which can be thought of as a process of taking a video, often compressed in a specific format to save space, and unpacking it for analytics and processing. After decoding, the pipeline continues to do scaling. For example, a video recorded in high definition can be adapted for playback on a smartphone with a smaller screen or a larger TV screen without blurriness or pixelation.

The next step of inference in video files is making educated guesses based on visual data. This process is used for tasks like recognizing objects, identifying faces, and understanding actions.

On the far right, the results are displayed on a TV, stored for future reference, stored in the cloud, or used to generate alerts based on suspicious events, such as someone jumping over a fence at a home.

The diagram above also depicts which hardware component (GPU vs. CPU or others) processes each pipeline block.

Role of Intel GPUs in media analytics

Intel GPUs have dedicated hardware that enables fast, energy-efficient encoding and decoding of video compliant with the industry standards such as AVC, HEVC, and AV1. Intel Data Center GPU Flex Series comes in two flavors:

- Intel Data Center GPU Flex 140
- Intel Data Center GPU Flex 170

The Intel Data Center GPU Flex 140 is a low-profile, 75W option with two 8 Xe-core GPUs. The Intel Data Center GPU Flex 140 is a small form factor GPU that provides more video decode/encode engines and is ideal for pipelines that involve lighter AI models, such as simple object detection. Due to its higher number of video engines, the Flex 140 GPU can support high number of streams.

The Intel Data Center GPU Flex 170 is a 150W option featuring a single 32 Xe-core GPU for higher peak performance. This high-power, 150W adapter is optimal for pipelines that include more complex AI models, such as multiple object detection or multiple classification models. With its AI compute power, it can support the same number of video streams per adapter as the compute requirements increase.

These accelerators have four classes of video accelerator engines:

- Two engines (VDBOX) that accelerate video decoding and encoding.
- Two video enhancement engines (VEBOX) and two scale and format converters (SFCs) that accelerate video scaling, color space conversion, denoising, deinterlacing and more.
- One rendering engine that provides distributed execution units combined with media samplers.

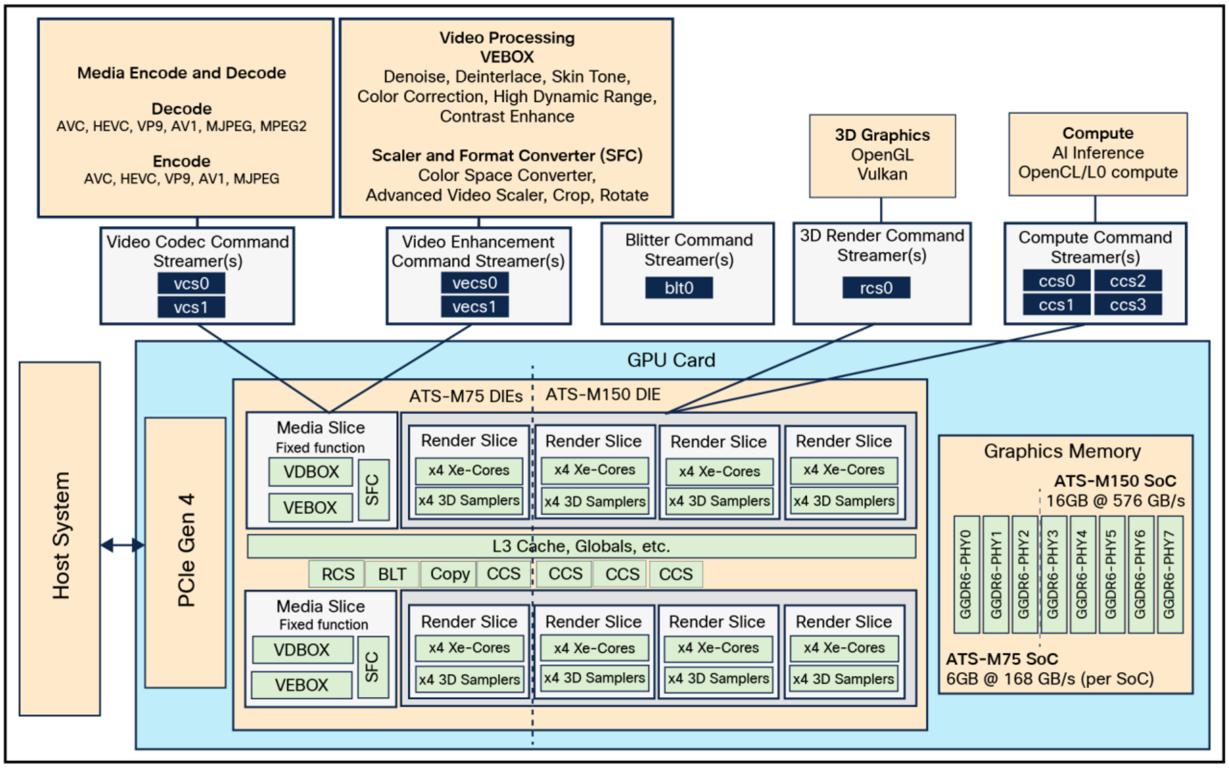


Figure 7.
Overview of Intel Data Center GPU Flex 140 (ATS-M75) and Intel Data Center GPU Flex 170 (ATS-M150) SoC

Overview of Intel benchmark tools

Intel Deep Learning Streamer

Intel DL Streamer is a streaming media analytics framework, based on the GStreamer (open source) multimedia framework, for creating complex media analytics pipelines.

Intel DL Streamer makes media analytics easy:

- Write less code and get industry-leading performance.
- Quickly develop, optimize, benchmark, and deploy video & audio analytics pipelines in the cloud and at the edge.
- Analyze video and audio streams, create actionable results, capture results, and send them to the cloud.
- Leverage the efficiency and computational power of Intel hardware platforms.
- Intel DL Streamer includes:
    - Intel DL Streamer Pipeline Framework for designing, creating, building, and running media analytics pipelines. It includes C++ and Python APIs.

Intel DL Streamer provides over two dozen samples, demos, and reference applications for the most common media analytics use cases such as action recognition, face detection and recognition, draw face attributes, audio event detection, vehicle and pedestrian tracking, human pose estimation, metadata publishing, smart city traffic and stadium management, intelligent ad insertion, single- & multi-channel video analytics pipeline benchmarks, and other use cases.

Intel DL Streamer Pipeline Zoo

The Intel DL Streamer Pipeline Zoo is a media analytics tool optimized for Intel hardware. It includes the Pipebench benchmark tool for downloading pipelines and their dependencies for performance measurements.

The term "pipeline" denotes a sequence of tasks, such as decoding, scaling, and inference, that you want to execute on provided input media.

Pipelines are organized according to the task they perform (what types of input they accept and what types of output they generate). Figure 8 provides an example of pipeline for object detection.

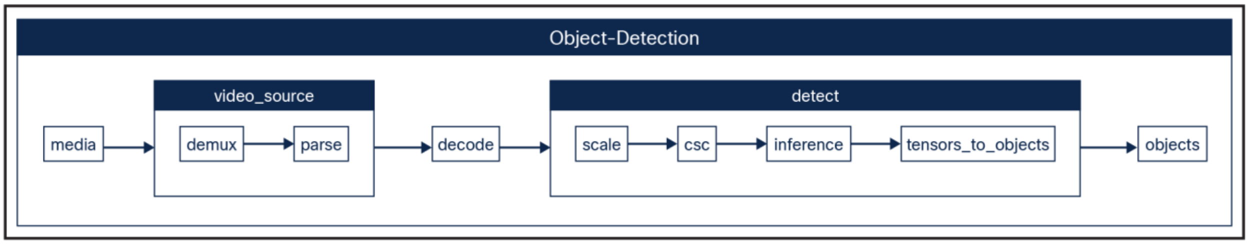


Figure 8.
Example of media analytics pipeline for Object detection

Performance test configuration

The test configuration is as follows:

- One Cisco UCS X440p PCIe Node was paired with Cisco UCS X210c M7 Compute Node installed on a Cisco UCS X9508 Chassis.
- 2 x Intel Data Center GPU Flex 140 cards were installed on Cisco UCS X440p PCIe Node, which hosts the GPUs.
- AVC OD, HEVC OD, AVC OC, HEVC OC, AVC OD+OC, and HEVC OD+OC workload tests were performed using Intel DL Streamer Pipeline Zoo.
- For this white paper, a pipebench command line utility was used to measure the performance of a pipeline under different scenarios (object detection, object classification, decoding, etc.).

Benchmark test case description

AVC vs. HEVC

H.264 (AVC) and H.265 (HEVC) are both standards for video compression used in recording and distributing digital video.

HEVC offers superior compression efficiency, resulting in smaller file sizes or better quality at the same size, making it ideal for high-resolution content such as 4K and 8K video. It also provides better video quality with fewer artifacts. AVC, on the other hand, enjoys broader compatibility with older devices and software. HEVC is preferred for modern applications where efficiency and high quality are paramount, while AVC remains a practical choice for legacy systems and devices due to its wider support.

Object detection and object classification

Object detection is like finding multiple hidden treasures in a picture, marking each treasure's exact spot with a bounding box. Meanwhile, classification is like putting a single label on the whole picture or each treasure you have found.

For example, in a self-driving car application, object detection identifies pedestrians, cars, and traffic signs in the road scene with bounding boxes around each one. After that, object classification labels these detected objects as " pedestrian," " car," or (for example) " stop sign." These tasks are often combined in complex computer vision systems to fully understand the visual scene.

Planned test cases

A media analytics pipeline is composed of media plus inference components connected in a generic sense. Table 1 lists the MA use cases included for this testing and the AI inference models used.

Table 1. Media analytics performance test cases for different pipelines using various AI inference models.

Use Case	Description	Pipebench Pipeline Used
AVC OD	AVC decoding with object detection only	od-h264-ssd-mobilenet-v1-coco
HEVC OD	HEVC decoding with object detection only	od-h265-ssd-mobilenet-v1-coco
AVC OC	AVC decoding with object classification	oc-h264-full_frame-resnet-50-tf
HEVC OC	HEVC decoding with object classification	oc-h265-full_frame-resnet-50-tf
AVC OD+OC	AVC decoding with object classification and detection	oc-h264-ssd-mobilenet-v1-coco-resnet-50-tf
HEVC OD+OC	HEVC decoding with object classification and detection	oc-h265-ssd-mobilenet-v1-coco-resnet-50-tf

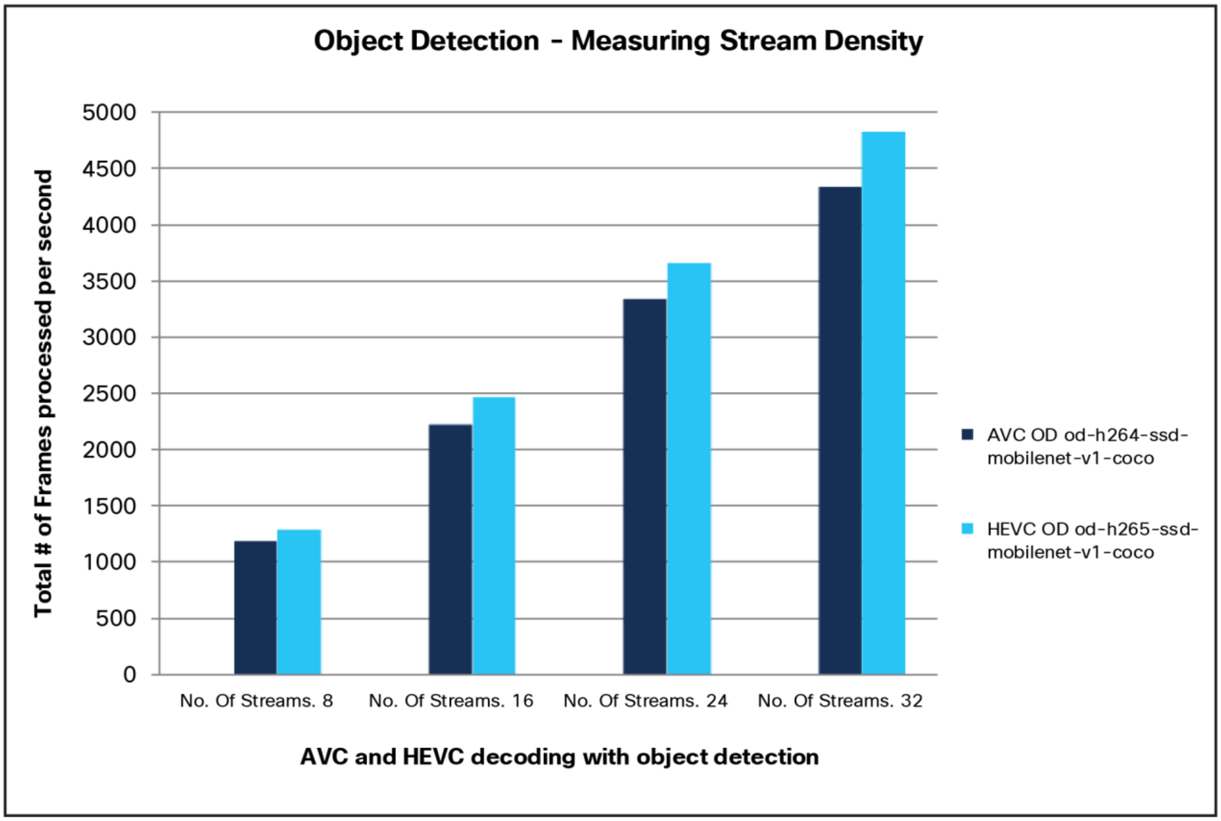
Media analytics performance results

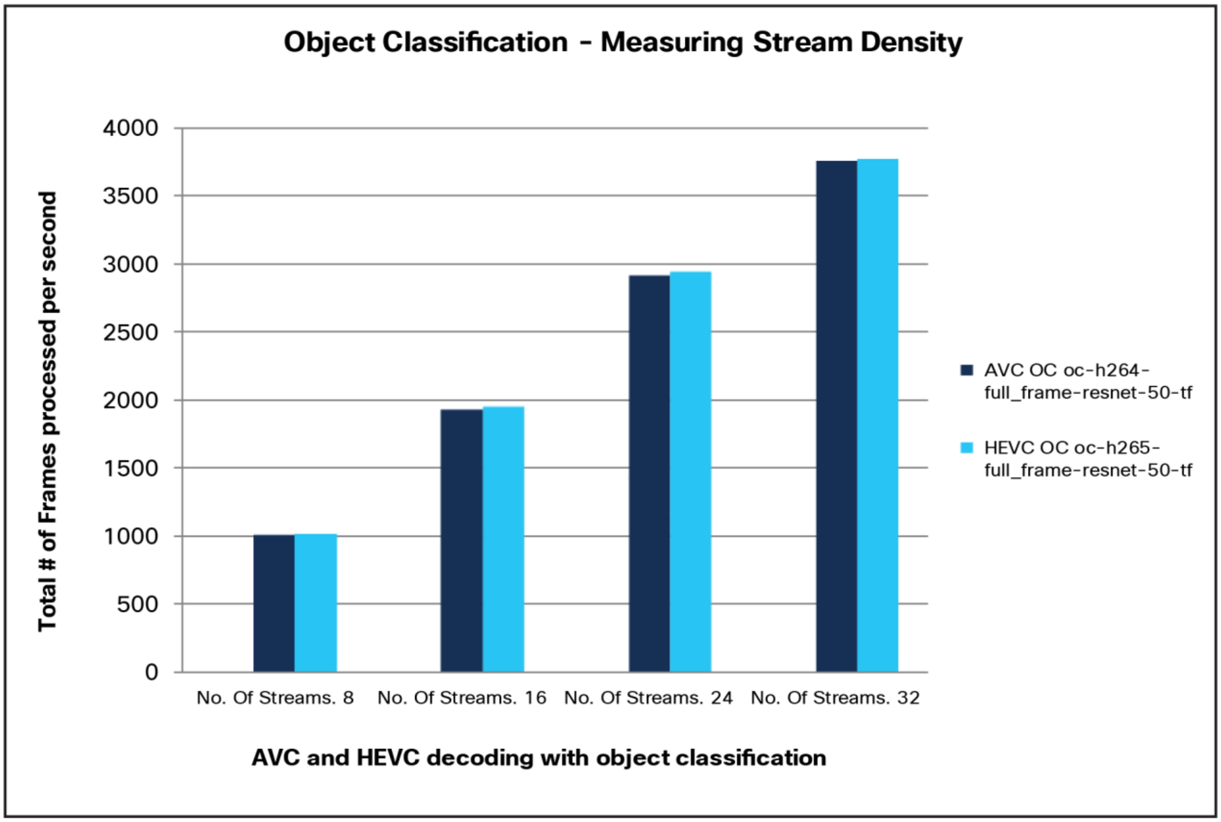
Here are the performance benchmark results for media analytics pipeline measured on a Cisco UCS X440p PCIe GPU node configured with two physical Intel Flex GPU 140 cards paired with Cisco UCS X210c compute nodes. A media analytics pipeline comprises media and the connected inference components. For this white paper, we have measured the stream density results with various numbers of streams (8, 16, 24, and 32). Before running the performance measurement of the media analytics pipeline, the CPU frequency scaling governor was set to performance.

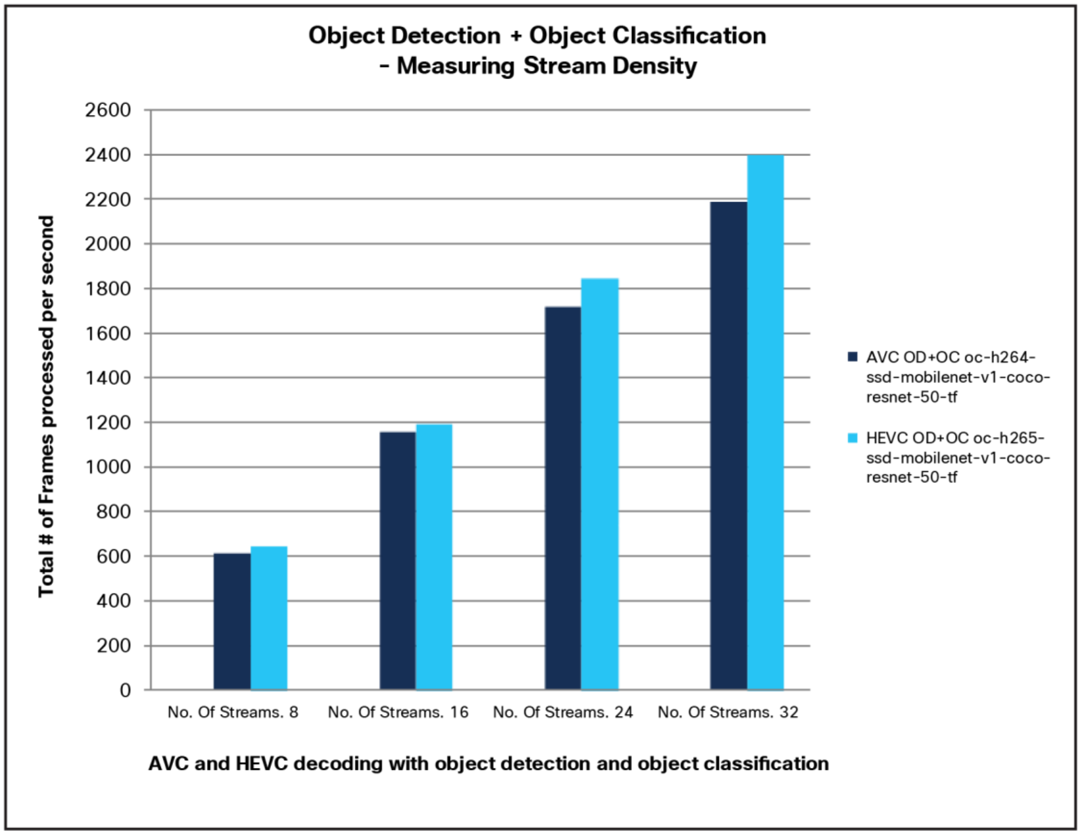
Table 2. Media analytics performance results for various pipelines

Media	Workload	Frames Processed Per Second			
		8 Streams	16 Streams	24 Streams	32 Streams
AVC	Object detection	1186	2222	3348	4340
HEVC	Object detection	1293	2472	3655	4831
AVC	Object classification	1008	1921	2910	3752

Media	Workload	Frames Processed Per Second			
		8 Streams	16 Streams	24 Streams	32 Streams
HEVC	Object classification	1017	1948	2940	3767
AVC	Object detection + object classification	619	1161	1721	2189
HEVC	Object detection + object classification	651	1194	1849	2400




Object Classification - Measuring Stream Density
Total # of Frames processed per second
4000
3500
3000
2500
2000
1500
1000
500
0
No. Of Streams. 8
No. Of Streams. 16
No. Of Streams. 24
No. Of Streams. 32
AVC and HEVC decoding with object classification
■ AVC OC oc-h264-full_frame-resnet-50-tf
■ HEVC OC oc-h265-full_frame-resnet-50-tf


Object Detection + Object Classification - Measuring Stream Density
Total # of Frames processed per second
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
No. Of Streams. 8
No. Of Streams. 16
No. Of Streams. 24
No. Of Streams. 32
AVC and HEVC decoding with object detection and object classification
■ AVC OD+OC oc-h264-ssd-mobilenet-v1-coco-resnet-50-tf
■ HEVC OD+OC oc-h265-ssd-mobilenet-v1-coco-resnet-50-tf

Media analytics performance analysis summary

The test measures the stream density using eight streams on each graphics device. Stream density is the maximum number of streams that can be processed at the target-FPS in parallel.

- The number of frames processed per second (hence processing speed) linearly increases with the number of GPUs.
- When a combination of object detection and classification is performed by the GPU in parallel, it results in an increase in the GPU traffic. So, the number of frames processed per second is reduced.

Performance summary

The Cisco UCS X210c M7 Compute Node, paired with the Cisco UCS X440p PCIe Node using Intel Data Center GPU Flex 140 provide the highest level of performance for the modern generation video standards. Includes HEVC and AVC, while still supporting ultra-high density and optimized media processing, inferencing, and analytics operations.

For more information

For additional information about the Cisco UCS X-Series Modular System, refer to the following resources:

- https://www.cisco.com/c/en/us/products/servers-unified-computing/ucs-x-series-modular-system/index.html.
- https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/solution-overview-c22-2432175.html?ccid=cc002456&oid=sowcsm025665.
- https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-x-series-modular-system/x440p-specsheet.pdf.

For information on Intel Data Center GPU Flex series and Intel DL Streamer pipeline framework, refer to the following resources:

- https://www.intel.com/content/www/us/en/products/docs/discrete-gpus/data-center-gpu/flex-series/overview.html.
- https://github.com/dlstreamer/dlstreamer.

Appendix: Test environment

Table 3 lists the details of the server under test.

Table 3. Compute node and PCIe GPU node properties

Name	Value
Compute node	Cisco UCS X210c M7 Compute Node
CPUs	Two 2.0-GHz Intel Xeon Gold 6438Y+ processors (32 cores)
Total memory	1024 GB (16x 64G 4800 DIMMs)
VIC adapter	UCSX-ML-V5D200G-D: Cisco UCS VIC 15231 2x100/200G mLOM
SFF NVMe SSDs	6.4 TB 2.5-inch Intel D7-P5620 NVMe high-performance high-endurance SSDs (UCSX-NVME4-6400-D)
OS	Ubuntu 22.04.2 LTS
PCIe node	Cisco UCS X440p PCIe Node
GPUs	2 x Intel Data Center GPU Flex 140

Table 4 lists the server firmware version and the BIOS settings used for media analytics benchmark performance testing.

Table 4. Server firmware version and BIOS settings

Name	Value
Firmware version	Release 5.1(1.230052)
BIOS version	Release 5.1.1d
Fabric interconnect firmware version	Release 9.3(5)I42(3d)
BIOS settings	Set to Platform-default

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **https://www.cisco.com/go/offices**.

Printed in USA C07-4124153-00 12/23