Solution Brief

Data Center Graphics
AI Visual Inference

Intel® Data Center GPU Flex Series for AI Visual Inference

Growing industry requirements for cost-effective, high-density AI visual inference at large scale are met by the Intel Data Center GPU Flex Series.

The AI visual inference market is being driven by analytics on video data from security and site monitoring cameras, live broadcast streams and over-the-top (OTT) video. GPUs are forecasted to play an increasingly important role in this market segment.

Organizations that implement these systems need to maximize the amount of inference throughput per dollar spent. The Intel® Data Center GPU Flex Series provides the high throughput and low total cost of ownership (TCO) that customers need for:

- **Safety and Security —** The development of smart infrastructure is an area of high growth as public authorities and others create ways of applying data analytics to protect safety and improve efficiency, to public safety, road safety, site security and worker safety use cases. Common usages include detecting people in dangerous locations such as railway crossings or roadways, analyzing pedestrian and vehicle traffic to optimize space planning and identifying whether people are wearing required protective gear such as face masks or helmets.
- **Library indexing and compliance —** to extract information from large collections of video data to facilitate querying their contents. For example, a movie collection could be searched for specific plot elements, while video monitoring footage could be analyzed for specific behaviors of interest, without hours of manual review.
- **AI-guided video enhancement and encoding —** to improve the quality of video or lower the cost of its delivery. Older content can be upscaled to 4K in real time at far lower cost than specialized manual processes and at far better quality than the bicubic interpolation mechanisms employed by many 4K televisions. Likewise, UHD content can be downscaled to HD or SD for lower-cost distribution channels.
- **Retail Video Analysis —** use of artificial intelligence to improve customer journey through a retail story such as in self-checkouts adopting computer vision. Also supporting automated solutions for store efficiency measures such as loss prevention or inventory management.

The Flex Series GPUs are designed to complement Intel Xeon® Scalable processors in the same system to efficiently handle diverse, complex workloads. The Intel Xeon platform provides high per-core performance; an enhanced memory subsystem and acceleration technologies; and the flexibility of a wide range of core counts, clock speeds and features.

In addition to reducing platform silos between the GPU and CPU, the combination also inherits the Intel Xeon processor's immense software ecosystem advantage, with support across popular and open-source tools, APIs, frameworks and applications.

SUPPORTING STAT

UP TO **256 TOPS INT8**

(Tera Operations per second) per PCIe card

Open standards architecture

Code developed for GPUs under proprietary programming models such as CUDA lacks portability to other hardware, creating a siloed development practice that locks organizations into a closed ecosystem. By contrast, the Flex Series GPU supports a unified, open, standards-based software stack combined with oneAPI cross-architecture programming so developers can build high-performance AI applications and solutions that run seamlessly on Intel CPUs and GPUs, utilizing oneAPI-optimized deep learning frameworks — OpenVINO™, PyTorch and TensorFlow.

Open-standards code development based on oneAPI benefits from a large open ecosystem that includes open-source tools, APIs and drivers. That flexibility helps organizations reduce the complexity, cost and time requirements to bring new services and solutions to market, enabling engineers and programmers to innovate instead of maintaining code.

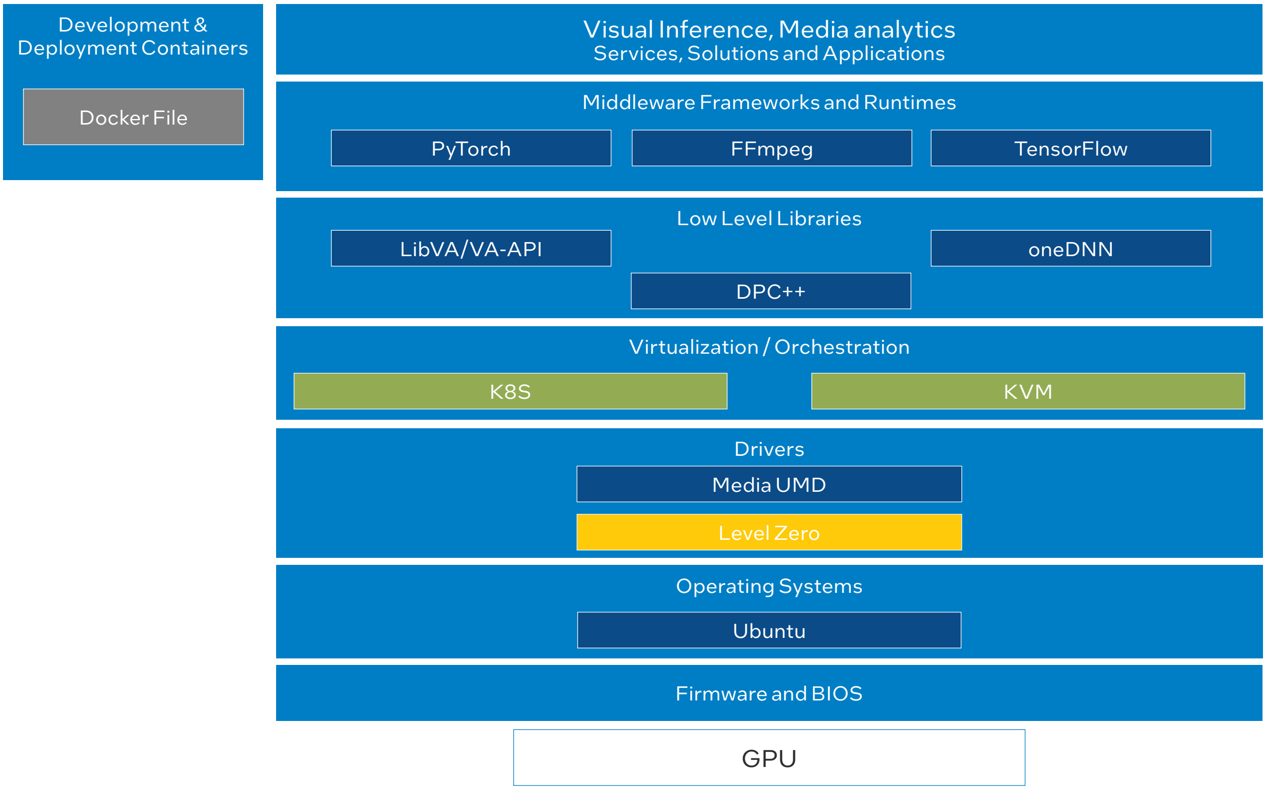
Industry ecosystem

Extending the benefits of its standards-based open architecture and optimized software stack, the Flex Series GPU draws on a broad ecosystem of service providers, independent software vendors (ISVs), original equipment manufacturers (OEMs) and others to support a wide range of media use cases.

These companies draw on the published oneAPI specification to integrate the Flex Series GPU with their technologies for AI visual inference. The openness and transparency of the programming model also encourages uptake by the open-source community, creating a virtuous cycle to further enhance the software stack, which supports popular programming languages including Python and C/C++. Inference, vision and media APIs are up-streamed to frameworks provided by the optimized industry ecosystem, helping streamline deep learning development and integration into existing solutions.

Intel® Data Center GPU Flex 170 & Flex 140

Intel is enabling the software ecosystem to take full advantage of the underlying hardware's capabilities at processing and delivering AI inference capabilities. This work helps ensure that software standards, frameworks and open-source technologies — as well as the customers that use them — can easily build solutions that increase throughput on a combination of CPUs and GPUs. To reach that goal, Intel invested substantially to enable programmability for media processing and delivery across CPU and GPU architectures.


Development & Deployment Containers
Docker File
Visual Inference, Media analytics
Services, Solutions and Applications
Middleware Frameworks and Runtimes
PyTorch
FFmpeg
TensorFlow
Low Level Libraries
LibVA/VA-API
oneDNN
DPC++
Virtualization / Orchestration
K8S
KVM
Drivers
Media UMD
Level Zero
Operating Systems
Ubuntu
Firmware and BIOS
GPU

Intel® oneAPI tools

Open-source performance libraries integrate decode and inference workloads into composite applications, providing performance optimization across Intel architecture. These integrated pipelines enable development of classification, detection, tracking and segmentation workloads for use cases in areas such as smart infrastructure and media analytics.

- **Intel Distribution of OpenVINO Toolkit**, powered by oneAPI, accelerates pre-trained deep learning models deployed across multiple types of Intel architectures. It also facilitates programming end-to-end applications with hardware acceleration for multi-compute functions using common multimedia frameworks such as GStreamer and FFmpeg. Multiple data types — such as FP32, BF16, FP16 and INT8 — support diverse application requirements, such as balancing between accuracy and performance. A full range of GPU operations (e.g., tensor/array, math, neural networks) support diverse models for AI inference.
- **Intel oneAPI Deep Neural Network Library (oneDNN)** enables developers to maximize productivity and performance of deep learning frameworks on CPUs and GPUs. The OpenVINO toolkit builds on oneDNN (which is part of that software package) and is also integrated into popular frameworks providing optimizations for TensorFlow and PyTorch.
- **Compute Library for Deep Neural Networks (clDNN)**, part of the OpenVINO toolkit, is a performance library for accelerating deep-learning inference on Intel Graphics. It provides optimized building blocks for implementing convolutional neural networks (CNNs) with C and C++ interfaces.
- **Intel oneAPI Video Processing Library (oneVPL)**, is a performance library to optimize media transcode performance across integrated and discrete GPUs with a single codebase. oneVPL provides a video-focused API for fast video decoding, encoding and processing.

Analyze application performance with Intel® VTune™ Profiler

Accelerate application compute-intensive tasks by identifying the most time-consuming parts of GPU code and optimizing GPU offload schema and data transfers for SYCL, OpenCL code, Microsoft DirectX or OpenMP offload code. Analyze GPU-bound code for performance bottlenecks caused by microarchitectural constraints or inefficient kernel algorithms.

High-efficiency codecs

Even as large-scale data storage has become progressively cheaper, bandwidth to access that data remains expensive. Improved compression enables media processing and delivery providers to reduce those bandwidth requirements for lower operating costs.

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 GPU brings the highest quality real-time video scalable to any modern device at any bandwidth. It enables delivery of commercial or non-commercial user-generated content with low computational footprint, optimized for internet streaming. It does all this at 30% better compression with no degradation in streaming quality, reducing the cost per stream.[1]

In addition to AV1, the GPU also supports existing HVEC, AVC and VP9 workloads. Using AV1 can deliver cost savings relative to AVC and HEVC by avoiding the significant royalties payable for those codecs to the MPEG Licensing Authority (MPEG LA).[2] Providers can maximize quality for the channels with the greatest viewership and highest-profile content using the SVT-AV1 software encoder developed by Intel in cooperation with the Alliance for Open Media. Drawing on the open-source Scalable Video Technology (SVT) project for core libraries, the encoder is highly optimized for Intel Xeon Scalable processors, providing outstanding performance and power efficiency on the same servers that host Intel Data Center GPUs.

These codecs can be accessed using standard frameworks, such as FFmpeg or GStreamer, or with oneVPL, which provides additional access to more controls and parameters. Both the hardware and software encoders provide a range of performance/quality presets so providers can make TCO-oriented adjustments according to the requirements of specific use cases.

Higher performance with lower total cost of ownership (TCO)

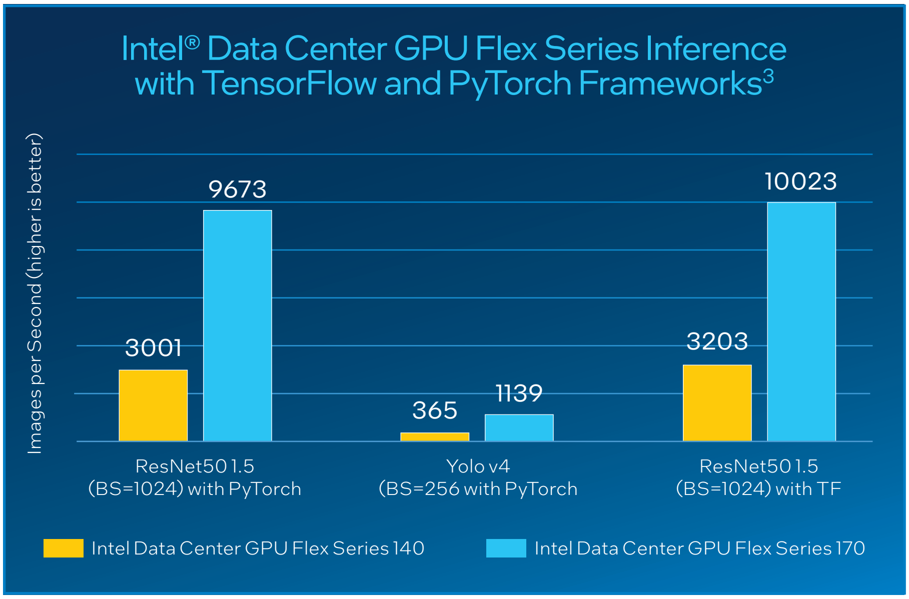
Solution providers have a strategic imperative to optimize TCO while meeting customer demands for more sophisticated functionality. By handling a given workload with less infrastructure, the GPU supports growing workloads with smaller data center footprints, helping reduce capital expenditure (CapEx) associated with equipment and facilities costs. High performance per watt helps drive TCO down further by reducing operational expenditure (OpEx).

The Intel Flex Series GPUs show impressive AI visual inference performance. The test scenario consists of decoding multiple HEVC video input streams, resizing them and performing inference on them. This type of pipeline is common across use cases that combine inputs from arrays of cameras, as in manufacturing quality inspection, crowd management, smart city and many others.

Intel extensions for both PyTorch and TensorFlow simplify running popular inference benchmarks on both the Intel Data Center GPU Flex Series 170 and 140. The Flex Series 170 GPU's 32 Intel Xᵉ cores typically deliver superior performance for low batch sizes and peak batch size inference compared to the 16 Intel Xᵉ cores provided by the Flex Series 140. Media-bound workloads are an exception, where the Flex Series 140 GPU's greater number of media engines may provide better performance. Intel AI inference library and framework support both CPU and GPU, with minimal code changes.

Intel® Xᵉ architecture

Built on the Intel Xᵉ architecture, the GPU 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.



For workloads and configurations visit www.Intel.com/PerformanceIndex.

The future for AI visual inference

The industry is in an accelerated state of adaptation in the effort to deliver breakthrough capabilities and experiences from edge to cloud for AI visual inference — while keeping their eyes squarely on the bottom line. The industry transition replacing proprietary, specialized technologies with those based on open standards is a key contributor to this balance, along with innovation moving forward. The Flex Series GPU contributes to this transition with a seamless hardware and software AI visual inference solution, untethering the AI visual programming environment from restrictive proprietary environments.

Learn more about the Intel® Data Center GPU Flex Series at www.intel.com/FlexSeriesGPU

[1] Cost savings based on streaming x264 at 4 Mbs (1.21 GB/hr) and 30% bit rate efficiency for AV1 at 2.7 Mbs (0.81 GB/hr). Host HW information: CPU: Intel(R) Xeon(R) Platinum 8360Y CPU, max turbo frequency 3.50 GHz, processor base frequency 2.40 GHz. Physical RAM memory size: 64 GB (4x16 GB DIMM DDR4 3200MHz). GPU Information: 2 Intel Data Center GPU Flex Series cards (Device 56c1 rev 05). Model: 75W 128EU. OS: Ubuntu 20.04.4 LTS (focal). Kernel: 5.10.54+prerelease2828 (agama-ci-prerelease-366). Display driver version: 22.2.1.10510-main. OneVPL: libvpl-dev version 2022.0.3.0.10510-main

[2] MPEG LA Current Programs, AVC (https://www.mpegla.com/programs/avc-h-264/) and HEVC (https://www.mpegla.com/programs/hevc/).

[3] AI Inference (OV):

Intel Data Center GPU Flex Series 170: 05/06/2023, Intel Coyote Pass, Intel Xeon Platinum 8360Y CPU @ 2.40GHz, 2S-36C-72T, 128GB DDR4-3200, OS: Ubuntu 22.04, 5.15 packed with Agama-devel-627, GPU: Intel Data Center GPU Flex Series 170 x 1 , GPU Firmware: ATS_M150_512_C0_ES_106_IFWI_2023WW13_02_00_PC9865_GSC2084_OP1063_Headless_18GT_P2SB_DIS_ECC_EN.bin, **ECC:** enabled, Openvino: OV master (commit: 615177ae), data measured at PVT. 50 ~65 OC Temperature; **AI model Resolution:** Resnet v1.5 (224 x 224), YoloV4 (416x416) for **Data type:** INT8.

Intel Data Center GPU Flex Series 140: 05/06-2023, Intel Coyote Pass, Intel Xeon Platinum 8360Y CPU @ 2.40GHz, 2S-36C-72T, 128GB DDR4-3200, OS: Ubuntu 22.04, 5.15 packed with Agama-devel-627, GPU: Intel Data Center GPU Flex Series 140 x 1, GPU Firmware: ATS_M75_128_B0_ES_029_23WW08_02_GS2054_PC9815_OP1060_Headless_P2SB_DS_ECC_EN.bin, **ECC:** enabled, **Openvino:** OV master (commit: 615177ae), data measured at DVT2 2.2 reworked as PVT. 50 ~65 OC Temperature; **AI model Resolution:** Resnet v1.5 (224 x 224), YoloV4 (416x416) **Data type:** INT8.

AI Inference (IPEX, ITEX):

Intel Data Center GPU Flex Series 170: 06/27-2023, Intel Coyote Pass, Intel Xeon Platinum 8360Y CPU @ 2.40GHz, 2S-36C-72T, 128GB DDR4-3200, OS: Ubuntu 22.04, 5.15 packed with Agama-devel-602, GPU: Intel Data Center GPU Flex Series 170 x 1 , GPU Firmware: ATS_M150_512_C0_ES_113_IFWI_2023WW17_05_00_PC9865_GSC2097_OP1064_Headless_18GT_P2SB_DIS_ECC_EN.bin, **ECC:** enabled, **TensorFlow:** AI container with ITEX1.2.0 2023 Q1 public release build-in; **PyTorch:** AI container with IPEX_1.2.0 2023 Q1 public release build-in, data measured at PVT. 50 ~65 OC Temperature **AI model Resolution:** Resnet v1.5 (224 x 224) for TF/PT, YoloV4 (416x416) for PT **Data type:** INT8.

Intel Data Center GPU Flex Series 140: 06/29-2023, Intel Coyote Pass, Intel Xeon Platinum 8360Y CPU @ 2.40GHz, 2S-36C-72T, 128GB DDR4-3200, OS: Ubuntu 22.04, 5.15 packed with Agama-devel-602, GPU: Intel Data Center GPU Flex Series 140 x 1, GPU Firmware: ATS_M75_128_B0_ES_033_23WW18_02_GS2097_PC9866_OP1064_Headless_P2SB_DS_ECC_EN.bin, **ECC:** enabled, **TensorFlow:** AI container with ITEX1.2.0 2023 Q1 public release build-in; PyTorch: AI container with IPEX_1.2.0 2023 Q1 public release build-in, data measured at DVT2 2.2 reworked as PVT. 50 ~65 OC Temperature; **AI model Resolution:** Resnet v1.5 (224 x 224) for TF/PT, YoloV4 (416x416) for PT **Data type:** INT8.

No product or component can be absolutely secure.

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

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.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

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.

Intel is committed to respecting human rights and avoiding complicity in human rights abuses. See Intel's Global Human Rights Principles. Intel's products and software are intended only to be used in applications that do not cause or contribute to a violation of an internationally recognized human right.

1023/MH/MESH/349366-002US