Intel Data Center GPU Flex Series

Jeff McVeigh – VP & GM AXG Super Compute Group
Pixel Explosion

19+ Terapixels per second
of user generated content

7 Petapixels per second
of video content consumed worldwide

Pixel Uses Expand

Media Processing & Delivery
44B hours per week

Cloud Gaming
6.4B hours per year

AI Inference
4.8B+ hours per year

Virtual Desktop Infrastructure
166M Instances per year

Xeon gold Standard
GP pixel processor

Current Accelerator Solutions Limited...

Flexible Pixel Powerhouse
Built with a
Open SW Stack

Proprietary Closed SW Stack
Flexible Performance

Open Software Stack
Narrow Performance

Intel Data Center GPU Flex 140

Intel Data Center GPU Flex 170

Shipping to customers now
Intel Data Center GPU

- Media Engines: 4
- Power Envelope: 75W
- Ray Tracing Units: 16
- Xe HPG Architecture
- PCIe
- Xe cores: 16
Xe media engine

Pure fixed function hardware
Supported by popular video software

Up to 8k60 12-bit HDR decode
Up to 8k 10-bit HDR encode

VP9  AVC  HEVC  AV1
Media Processing & Delivery

Software Stack

Industry’s Biggest Media Frameworks Supported

Dedicated oneAPI Libraries – oneVPL & oneDNN
Decide Throughput

AVC
HEVC
AV1
VP9

Intel Data Center GPU Flex140
Nvidia A10

*In Fastest Mode for HEVC & AV1 in 1080p
Transcode Performance

<table>
<thead>
<tr>
<th></th>
<th>Perf</th>
<th>Balanced</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>4K60</strong></td>
<td>8</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td><strong>1080p60</strong></td>
<td>36</td>
<td>28</td>
<td>16</td>
</tr>
</tbody>
</table>

Intel Data Center GPU Flex140  Nvidia A10

See [www.intel.com/performancedata](http://www.intel.com/performancedata) for workloads and configurations. Results may vary.
30% Distribution Cost Savings vs x264 Medium

Cost savings based on streaming x264 @4Mbs (1.21GB/hr) and 30% bitrate efficiency for AV1@2.7Mbs (0.81GB/hr)

See www.intel.com/performanceindex for workloads and configurations. Results may vary.
Xe HPG

Scalable High-Performance Graphics Architecture for visually immersive workloads

Xe core
16 Vector Engines
16 XMX Engines

Scalable Architecture

Dedicated Ray Tracing Units

Cloud Gaming Driver Optimized

intel.
Cloud Gaming
Software Stack

- Biggest 3rd Party Game Engines Supported
- Broad Driver Support
- Run Across CPU & GPU Seamlessly
Cloud Gaming Performance

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

Honor of Kings (王者荣耀)
720p60: 48  720p30: 40

Asphalt 9: Legends
720p30: 29  1080p30: 23

Knives Out
720p30: 60  1080p60: 28

Note: 720p60 and 1080p60 are rendered at 60fps and encoded at 30fps

See www.intel.com/performanceindex for workloads and configurations. Results may vary.
**Xe Matrix Extensions**

Dedicated built-in AI functionality
Accelerating most AI Data Types

- **128 FP16/BF16** ops/clock
- **256 INT8** ops/clock
- **512 INT4/INT2** ops/clock
Support for Biggest AI Frameworks

Effortlessly Scale Code from Xeon to Intel Flex Series GPU

TensorFlow, OpenVINO, PyTorch

oneDNN delivers AI-optimized kernels
Resnet50v1.5
SSD - MobileNet
BS1
BS256
BS1
BS256

Based on OV 2022.3 For workloads and configurations visit www.Intel.com/PerformanceIndex. Click on the Events tab and Innovation Event Claims. Results may vary.
AI Inference

For workloads and configurations visit www.Intel.com/PerformanceIndex. Click on the Events tab and Innovation Event Claims. Results may vary.
For workloads and configurations visit www.Intel.com/PerformanceIndex. Click on the Events tab and Innovation Event Claims. Results may vary.
Intel® DL Streamer
Media Analytics Framework

- DL Streamer AI elements
- Intel xPU optimized
- Community supported and Intel xPU optimized
- 3rd party

Intel® DL Streamer Performance Critical Section

- Pipeline construction API
- Pads and Properties API
- Process Metadata

- Gststreamer

- Intel® DL Streamer Performance Critical Section

- Pipeline

- Decode
- Interference
- Object Tracking
- Overlay
- Utility Elements

- oneAPI Base Kit, OpenVINO inference engine, OpenCV...

* All trademarks, service marks and company names are the property of their respective owners.
Object Classification
- HEVC + Resnet50

Object Detection + Classification
- HEVC + SSD-MobileNet + Resnet50

Object Detection
- HEVC + SSD-MobileNet

For workloads and configurations visit www.Intel.com/PerformanceIndex. Click on the Events tab and Innovation Event Claims. Results may vary.
AI Visual Inference (Media + AI)

Object Classification
HEVC + Resnet50

Object Detection
HEVC + SSD-MobileNet

Based on OV 2022.2. For workloads and configurations visit www.Intel.com/PerformanceIndex. Click on the Events tab and Innovation Event Claims. Results may vary.
Virtual Desktop Infrastructure

Support Coming 1H 2023
HW based
SR-IOV
Virtual Desktop Infrastructure
Software Stack

No Additional Virtualization Costs
for GPU SW licensing
“The Intel Flex Series card using OpenVINO in Intel Developer Cloud performed ResNet50 inference on real world MRI data from our collaborators at Stanford several times faster than the PyTorch and ONNX runtimes on other vendor GPU platforms we were previously using. Additionally, by using the AUTO feature available in OpenVINO we were able to observe linear scaling across several Flex Series cards, further multiplying the speedup.”

Subutai Ahmad, CEO, Numenta

“In our testing, Intel’s Data Center Flex Series GPUs deliver compelling AI inferencing performance required for realtime solutions like Smart Cities and other industry verticals. They do this while (thanks to Intel’s OpenVINO) providing significantly easier AI development, deployment and management vs competitive discrete GPU options, making it a great choice for our customers.”

Ken Mills, CEO, EPIC IO

“There is tremendous interest among our clients in rolling out vision-based solutions at scale, and frictionless checkout is a great example of this. We are constantly looking to improve performance for our vision checkout solutions and Intel’s Data Center Flex Series GPUs are providing 3X better speed than existing discrete GPU alternatives. Powered by scalable OpenVINO inferencing software, we support multi-GPU edge servers powering multiple retail checkout stations. This is lowering total cost of ownership for operators and delivering outstanding user experience. UST Vision AI and our partner Xplorazzi are excited to take this to our clients and deliver meaningful outcomes.”

Alexander Varghese, Chief Operating Officer, UST

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Results may vary.
Open & Full SW Stack Approach
Built on the foundation of oneAPI for unified programming across architectures

Effortlessly Scale from CPU to GPU
Full Platform Compute for Middleware, Frameworks and Applications
**Intel Data Center GPU**

- **Flex 140**
  - 16 Xe cores
  - 16 Ray Tracing Units

- **Flex 170**
  - 32 Xe cores
  - 32 Ray Tracing Units

**AV1 Encoding & Decoding**

**ECC memory**

**Intel Xe Matrix Extensions**

- 75W
  - Half Height PCIe
- 150W
  - Full Height PCIe

**Up to 4 Xe Media Engines**

**Up to 256TOPS (INT8)**

**Xe HPG Based Arch**

**Hardware based SR-IOV**
### Media Processing & Delivery
- **5x** Media transcoding throughput vs Nvidia A10
- **2x** Media decoding throughput vs Nvidia A10
- **36 Streams** 1080p60
- **8 Streams** 4K60

### Cloud Gaming
- **Up to 68** 720p30 Game Streams (Flex 170)
- **Up to 46** 720p30 Game Streams (Flex 140)

### AI Inference
- **TensorFlow**
- **OpenVINO**
- **6 PyTorch**
- **130** Obj. Classification Streams per Card

### Virtual Desktop Infrastructure
- **Up to 62** Virtualized Functions

### System Designs
- Lenovo
- Cisco
- Dell Technologies
- H3C
- Hewlett Packard Enterprise
- Inspur
- Lenovo

### Support
- 3rd Gen Xeon Scalable Processors
- 4th Gen Xeon Scalable Processors
- Formerly code-named "Ice Lake"
- Formerly code-named "Sapphire Rapids"

*See www.intel.com/performancedata for workloads and configurations. Results may vary.*
Notices & Disclaimers

For notices, disclaimers, and details about performance claims, visit www.intel.com/PerformanceIndex. Or scan the QR code

From the landing page go to the Events tab and then to Intel Innovation 2022.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of the Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

Notices & Disclaimers

Performance varies by use, configuration and other factors. Learn more on the Performance Index site.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation. All product plans and roadmaps are subject to change without notice.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.