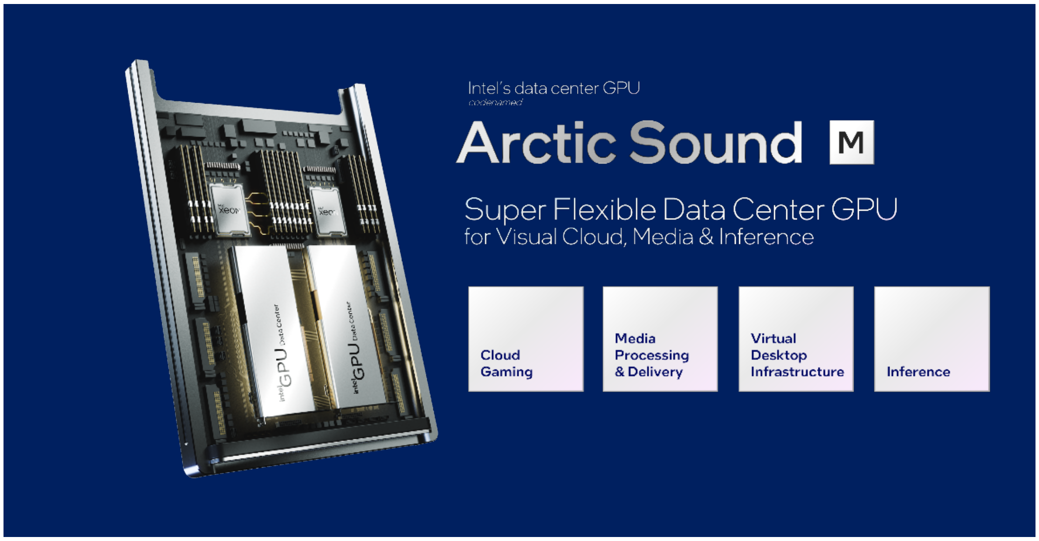
intel FACT SHEET

Intel's Data Center GPU, Code Named Arctic Sound-M

Intel® Xeon® Scalable Processor is the gold standard for cloud gaming, media processing and delivery, virtual desktop infrastructure, and inference – doing the heavy lifting for today's media consumption. With the current explosive growth in density and workload complexity, all these segments will bring unique workloads that now require processing pixels, inference, and analytics, rendering new content, and sending those pixels back to the client device to be viewed, or for further analysis. These tasks are currently being done by separate discrete products in the cloud.

Today, at its inaugural Intel Vision event, Intel shared further details on its data center GPU, code named Arctic Sound-M (ATS-M). Available in two configurations, ATS-M is a versatile GPU with leadership transcode quality and performance targeting 150 trillion operations per second (TOPS). A single solution built to flexibly handle a wide range of workloads, without compromising on performance or quality, while optimizing total cost of ownership (TCO).



Intel's data center GPU
codenamed
Arctic Sound M

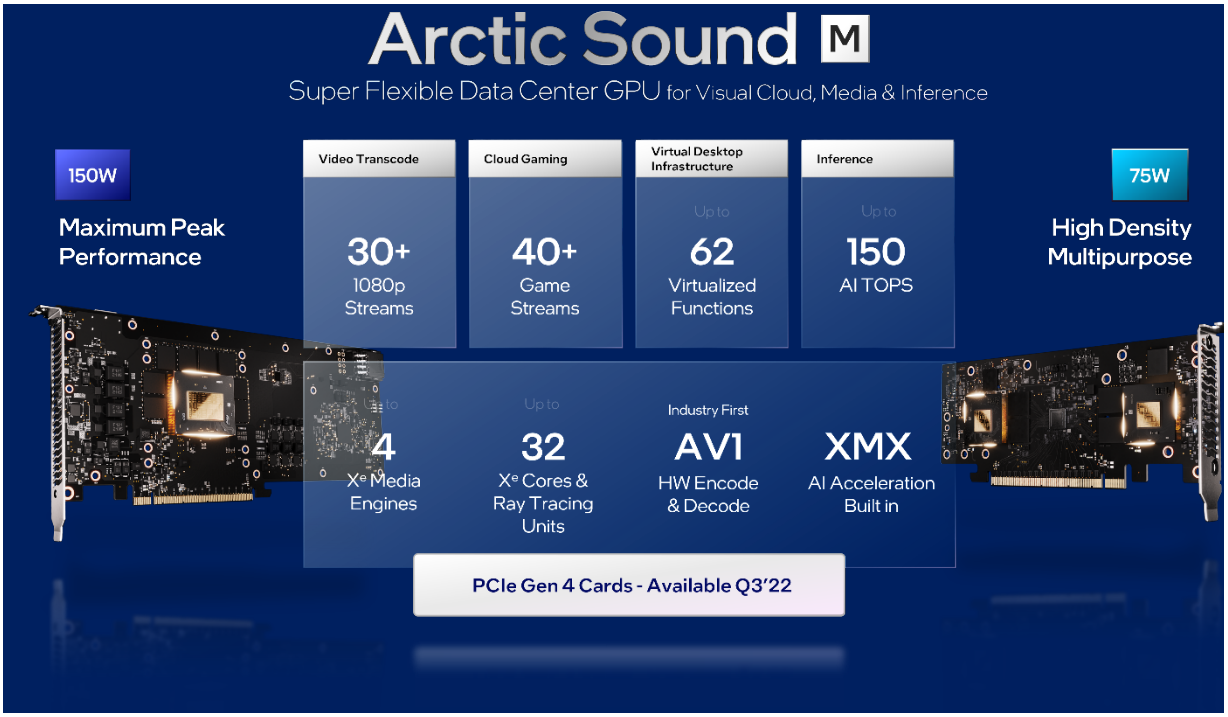
Super Flexible Data Center GPU
for Visual Cloud, Media & Inference

Cloud Gaming | Media Processing & Delivery | Virtual Desktop Infrastructure | Inference

More: For more information about Intel's data center GPU, codenamed Arctic Sound-M, including the Business Insights Vision presentation and news, please visit the Intel Newsroom.

Multi-purpose GPU for Data Center Workloads

Intel's data center GPU comes in two configurations. The 150W option has 32 Xe cores in a three-quarter length full height PCIe Gen4 card. The 75W option offers a low profile PCIe Gen4 card with two GPUs for 16 Xe cores. Both configurations come with four Xe media engines, the industry's first AV1 hardware encoder and accelerator for data center, GDDR6 memory, ray tracing units, and built-in XMX AI acceleration.



Arctic Sound M

Super Flexible Data Center GPU for Visual Cloud, Media & Inference

150W

Maximum Peak Performance

75W

High Density Multipurpose

Video Transcode	Cloud Gaming	Virtual Desktop Infrastructure	Inference
30+ 1080p Streams	40+ Game Streams	Up to 62 Virtualized Functions	Up to 150 AI TOPS

Up to 4 X^e Media Engines	Up to 32 X^e Cores & Ray Tracing Units	Industry First AV1 HW Encode & Decode	XMX AI Acceleration Built in

PCIe Gen 4 Cards - Available Q3'22

The Industry's only Data Center GPU Using Open-Source Software

Intel's data center GPU is supported by a full solution stack offering developers an open-source software stack for streaming media, cloud gaming, and inference, with broad support for AVC, HEVC, VP9, as well as APIs, framework, and the latest codecs.

oneAPI is the productive, smart path for accelerated computing, freeing developers from the economic and technical burdens of proprietary programming models. It is an open alternative to proprietary language lock-in that enables the full performance of the hardware with a complete, proven set of tools that complement existing languages and parallel models, allowing developers to design open, portable codes that will take maximum advantage of various combinations of CPUs and GPUs.

Cloud Gaming

The fast growth of the global cloud gaming market continues, with a projected CAGR of approximately 43.2 percent through 2026, when it will have a value of about $3.2B[1]. Intel supports a large user base with high quality gaming experience with support for both Windows and Android cloud gaming. With two form factors, the GPU also provides customers with the flexibility to choose

[1] "Insights on the Cloud Gaming Global Market to 2026 - Featuring Intel, Google and Microsoft Among Others." Research and Markets, January 4, 2022. https://www.globenewswire.com/news-release/2022/01/24/2371478/28124/en/Insights-on-the-Cloud-Gaming-Global-Market-to-2026-Featuring-Intel-Google-and-Microsoft-Among-Others.html

the right fit for their specific workload. Whether a user requires max peak performance, max density, or a converged cloud gaming solution that addresses both mobile and PC gaming on a single platform, the GPU delivers an excellent game streaming experience.

Full Stack Media Streaming Support

Intel's data center GPU provides the industry's first hardware accelerated AV1 encoder. Delivering a 30[2] percent bit-rate improvement without compromising on quality. [3]Ushering in a new generation of media streaming, it supports as many as eight simultaneous 4K streams or more than 30 1080p streams per card. In a server with four cards, this scales to 120 streams per node, and 13,000 streams per rack.

Media streaming and delivery software stacks lean on Intel® oneVPL to decode and encode acceleration for all the major codecs including AV1. Media distributors can choose from the two leading media frameworks FFMPEG or GStreamer, both enabled for acceleration with oneVPL on Intel CPUs and GPUs. Intel also offers the Open Visual Cloud - a set of open-source software stacks for media, analytics, graphics, and immersive media, optimized for cloud native deployment enabled to run within the FFmpeg and GStreamer frameworks.

Full Stack Virtual Desktop Infrastructure Support (VDI)

Virtual Desktop Infrastructure (VDI) and Desktop as a Service (DaaS) have been booming amid increased usage of remote workplace in the past two years (over 11 percent[4] growth in a recent study). Modern operating systems and applications are more graphics demanding and the display resolution has gone up; GPUs can offload the rendering and encoding to improve user experience with faster response time and higher frame rate. The freed-up CPU cycles can turn into higher performance of the user application workloads.

Intel's data center GPU also offers flexible vGPU scheduling policies (Fixed, Flexi, H/W Utilization Optimized Time-slice Scheduler), where the administrator has the capability to individually tweak each VM's execution quantum on the GPU. Current competitive offerings only allow tweaking this value as a global setting across all VMs.

Unlike current offerings, Intel does not charge any additional software licensing fees for hardware based SRIOV, helping to reduce overall cost of implementing virtualization for providers.

Full Stack Media Analytics Support

[2] Performance under claims and disclaimers

[3] H.264/AVC claim based on https://engineering.fb.com/2018/04/10/video-engineering/av1-beats-x264-and-libvpx-vp9-in-practical-use-case/.

[4] https://www.gminsights.com/industry-analysis/virtual-desktop-infrastructure-vdi-market?gclid=EAIaIQobChMIyKDg3bLD9wIVF5BoCR2P1gbvEAAYAiAAEgK7YfD_BwE

Every inference performed in the visual media space also requires video decode and pre-processing before being handed off to the AI model. The 75W option, with its two GPUs per card, has balanced compute and decode capabilities and is not media bound. This allows for excellent scaling of media analytics workloads, providing improved density and reduced costs for customers.

In parallel to oneVPL accelerating decoding and encoding of media streams, oneDNN (oneAPI Deep Neural Network library) delivers AI optimized kernels enabled to accelerate inference modes in TensorFlow or PyTorch frameworks, or with the OpenVINO model optimizer and inference engine to further accelerate inference and speed customer deployment of their workloads. This combination of AI and media software and stack can run seamlessly across Xeon and Intel's data center GPU.

Intel data center GPU, Arctic Sound-M has been validated with over fifteen designs from industry leading partners and will launch in the third quarter of 2022.

Notices and Disclaimers:

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

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.

[2] GPU Quality data is measured by Intel 02/09/2022 on pre-production Intel Data Center GPU (150) with IFWI ATS_M1_512_B0_EU_008_IFWI_21WW50_01_GS1463_PC9446A_HEADLESS_OP1040_P2SB_DS_14GT_LRG_TR_DS, on a 1-node dual socket Intel ® 3rd generation Xeon Gold 6336Y microcode 0x11 on an Intel customer reference server board (Coyote pass) 128GB 32-ch 3200 MT/sec, HT enabled, Turbo On, Ubuntu 20.04 LT2 kernel 5.10.54, gfx-driver-master-10017, pre-production FFMPEG. CPU Quality data is measured by Intel 02/09/2022 on an Intel SawTooth Pass Server 1-node, 2-socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz with HT on, Turbo on, CPU Microcode 0x4D, Windows 10 Enterprise LTSC 64-bit, 240GB 15-ch, DDR4-2666 DRAM. See command lines at pre-release version of https://github.com/intel/media-delivery. For "2 second delay" data collected with TU1 (Quality mode) using 27 broadcast /gaming sequences, details at https://dgpu-docs.intel.com/devices/iris-xe-max-graphics/guides/media.html, for "ultra low delay" data collected using VBENCH sequences, details at http://arcade.cs.columbia.edu/vbench. PSNRY-BDRATE quality metrics compared to x264-medium (tune PSNR, 1-pass) measured on Intel Xeon 6336Y.

All product plans and roadmaps are subject to change without notice.

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

Intel technologies may require enabled hardware, software or service activation.

Statements that refer to future plans or expectations are forward-looking statements. These statements are based on current expectations and involve many risks and uncertainties that could cause actual results to differ materially from those expressed or implied in such statements. For more information on the factors that could cause actual results to differ materially, see our most recent earnings release and SEC filings at www.intc.com.