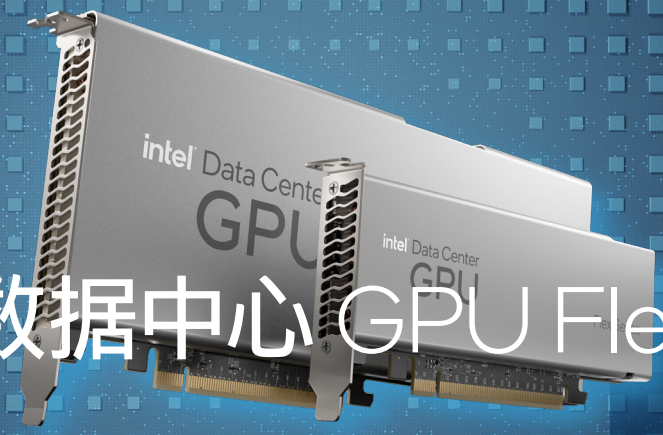
产品简介

加速计算
系统和显卡



英特尔® 数据中心 GPU Flex 系列

英特尔® 数据中心 GPU Flex 系列是面向智能视觉云的灵活、强大且开放的 GPU 解决方案。

越来越多的媒体处理和交付、AI 视觉推理、云游戏和桌面虚拟化任务在数据中心进行。但是这种迅速增长的势头因行业依赖专有的许可编码模型（如面向 GPU 编程的 CUDA）而大大受到限制。基于 CUDA 的软件也仅限用于专有 GPU，无法移植到其他加速器架构或 CPU 上。由此产生的总体拥有成本（TCO）上升压力使专有 GPU 编程无法大规模进行。

英特尔® 数据中心 GPU Flex 系列克服了上述限制，同时还为视觉云工作负载提供了出色的计算密度和能效。该系列产品基于英特尔® Xᵉ HPG（高性能显卡）微架构打造，内置视觉处理和 AI 加速技术。其提供的功能和优势包括：

- **支持开放、灵活、基于标准的软件堆栈以及 oneAPI 统一编程**，其中包括用于构建高性能、跨架构媒体应用和解决方案的开源组件与库、工具及框架。这种开放的方法有助于生态系统摆脱使用专有编程模型带来的技术和经济负担。
- **开创性地在 GPU 内配置了基于硬件的开源 AV1 编码器**，在相同质量下将带宽提高 30%，从而每年每十万名观众节省 2,300 万美元，或者在相同带宽下提高流媒体质量[1]。

支持性数据

高达 **68 路** 720p30 游戏视频流（基于特定游戏）

单个英特尔® Flex 系列 170 GPU[2]

高达 **46 路** 720p30 游戏视频流（基于特定游戏）

单个英特尔® Flex 系列 140 GPU[1]

硬件规格

该系列将以两种 SKU 形式提供：英特尔® 数据中心 GPU Flex 系列 170（峰值性能更高）和英特尔® 数据中心 GPU Flex 系列 140（密度更高）。图形处理器拥有多达 32 个英特尔® Xᵉ 内核及光线追踪单元、多达 4 个英特尔® Xᵉ 媒体引擎，具备用于 AI 加速的英特尔® Xᵉ Matrix Extensions（英特尔® XMX），并且支持基于硬件的 SR-IOV 虚拟化。每卡配备两个图形处理单元 (GPU) 的 Flex 系列 140 利用英特尔® oneVPL Deep Link Hyper Encode 功能可满足业内的“一秒时延”要求，同时提供 8K60 实时转码能力。此功能适用于 AV1 和 HEVC HDR 格式。

	英特尔® 数据中心 GPU Flex 系列 140	英特尔® 数据中心 GPU Flex 系列 170
目标工作负载	媒体处理和交付、基于 Windows 和 Android 的云游戏、虚拟桌面基础设施、AI 视觉推理[2]	
显卡外形规格	半高、半长、单宽、被动散热	全高、四分之三长、单宽、被动散热
显卡 TDP	75 瓦	150 瓦
每卡 GPU 数量	2	1
GPU 微架构	Xᵉ HPG	
Xᵉ 内核数量	16 个（8 个/GPU）	32
Fixed Function Media	4（2 个/GPU）	2
光线追踪	是	
峰值算力（脉动阵列浮点运算）	8 TFLOPS (FP32)/105 TOPS (INT8)	16 TFLOPS (FP32)/250 TOPS (INT8)
内存类型	GDDR6	
内存容量	12 GB（6 GB/GPU）	16 GB
虚拟化（实例）[3]	SR-IOV（62 个）	SR-IOV（31 个）
操作系统	Linux（Ubuntu、CentOS、Debian）、Windows Server 2019/2022、Windows Client 10、Red Hat® Enterprise Linux	
主机总线	PCIe Gen 4	
主机 CPU 支持	第三代英特尔® 至强® 可扩展处理器	

按用例划分的软件堆栈

Flex 系列 GPU 支持开放、灵活、基于标准的软件堆栈和 oneAPI 跨架构编程。堆栈包括开源的组件与库、工具及框架。开发人员可以利用它们开发高性能、跨架构媒体应用和解决方案，满足广泛的用例需求。这种开放的方法消除了专有模型形成的障碍——使用专有模型，代码可移植性和采用跨多供应商的新架构的能力会受到限制。

通用软件功能集可集成到主流中间件和框架中，而堆栈能以经过验证的产品化容器或参考堆栈形式交付。开发人员可以在裸机上利用 Kubernetes 对这些容器进行编排，或者利用 SR-IOV 虚拟化技术及相关工具在虚拟机中进行编排，实现工作负载分配和管理。此工具集旨在加快上市速度，并支持在同一 GPU 上灵活部署多个工作负载。

英特尔通过开展行业合作、参与和资助多项计划以及参加各种标准组织为软件生态系统赋能，并为开源社区提供持续的领导力、投资和技术贡献。

注：oneDNN 指 oneAPI 深度神经网络库。oneDAL 指 oneAPI 数据分析库。oneVPL 指 oneAPI 视频处理库。oneVPL、oneDNN、oneDAL 和英特尔® VTune™ Profiler 包含在英特尔® oneAPI 基础工具套件内（各个工具可单独下载）。面向英特尔® 架构优化的 TensorFlow 和 PyTorch 包含在英特尔® AI 分析工具套件内。

进一步了解英特尔® 数据中心 GPU Flex 系列，请访问

https://www.intel.cn/content/www/cn/zh/products/docs/discrete-gpus/data-center-gpu/flex-series/overview.html

[1] 实际性能受使用情况、配置和其他因素的差异影响。更多信息请见英特尔的性能指标网页。

[2] 反映了英特尔® 数据中心 GPU Flex 系列的功能，这些功能将在产品完全成熟时提供。

[3] 虚拟机因用例而异。

性能测试结果基于配置信息中显示的日期进行的测试，且可能并未反映所有公开可用的安全更新。详情请参阅配置信息披露。没有任何产品或组件是绝对安全的。

英特尔并不控制或审计第三方数据。请您审查该内容，咨询其他来源，并确认提及数据是否准确。

具体成本和结果可能不同。

英特尔技术可能需要启用硬件、软件或激活服务。