제품소개

가속 컴퓨팅 시스템 및 그래픽

인텔® 데이터센터 GPU Flex 시리즈

인텔 데이터센터 GPU Flex 시리즈는 유연하고 강력하며, 지능형 비주얼 클라우드를 위한 업계에서 가장 개방적인 GPU 솔루션입니다.

intel DATA CENTER GPU FLEX SERIES

미디어 처리 및 전달, AI 시각적 추론, 클라우드 게임 및 데스크탑 가상화가 데이터 센터에서 급증하고 있습니다. 이러한 급속한 성장은 GPU 프로그래밍을 위한 CUDA와 같은 독점적이고 라이선스가 부여된 코딩 모델에 대한 의존도가 높은 산업에서 주로 이뤄져 왔습니다. CUDA 기반 소프트웨어의 사용은 다른 가속기 아키텍처나 CPU로의 이식이 되지 않는 독점적인 GPU로 제한됩니다. 결과적으로 총 소유 비용에 대한 상승 압력은 독점 GPU 프로그래밍을 대규모로 유지하기 어렵게 만듭니다. 인텔® 데이터센터 GPU Flex 시리즈는 이러한 한계를 극복하는 동시에 시각적 클라우드 워크로드를 위한 뛰어난 컴퓨팅 밀도와 에너지 효율성을 제공합니다. 실리콘에 내장된 AI 및 시각 처리 가속화를 통해 GPU는 인텔® Xe HPG(고성능 그래픽) 마이크로아키텍처를 기반으로 합니다. 인텔® 데이터센터 GPU Flex 시리즈는 다음과 같은 기능과 이점을 제공합니다.

- **개방형 소스 구성 요소 및 라이브러리, 도구 및 프레임워크로 구성**되어 고성능의 교차 아키텍처 미디어 응용 프로그램 및 솔루션을 구축하는 **oneAPI 통합 프로그래밍과 함께 개방형의 유연한 표준 기반 소프트웨어 스택을 지원**합니다. 이 개방형 접근 방식은 생태계가 독점 프로그래밍 모델의 기술 및 경제적 부담에서 벗어날 수 있도록 도와줍니다.

- **GPU에서 업계 최초 하드웨어 기반으로 구현한 오픈 소스 AV1 인코더**는 동일한 품질의 대역폭을 30% 개선하여 연간 시청자 100,000명당 2,300만 달러의 비용을 절약하거나 동일한 대역폭에서 스트리밍 품질을 개선합니다.[1]

제품 관련 성능 수치

5X 경쟁사의 절반 수준의 전력으로 미디어 트랜스코드 처리량 제공

인텔 Flex 시리즈140 GPU와 NVIDIA A10 비교

HEVC 1080p60[1]

2X 경쟁사의 절반 수준의 전력으로 디코딩 처리량 제공

인텔 Flex 시리즈140 GPU와 NVIDIA A10 비교

across HEVC, AV1, AVC, VP9[1]

최 대 **68개** 일부 게임 스트림에서 720p30 스트림 달성 가능

Single Intel Flex Series 170 GPU[2]

최 대 **46개** 일부 게임 스트림에서 720p30 스트림 달성 가능

Single Intel Flex Series 140 GPU[1]

하드웨어 사양

본 제품은 최대 성능을 위한 인텔® 데이터센터 GPU Flex 시리즈 170과 최대 밀도를 위한 인텔® 데이터센터 GPU Flex 시리즈 140의 두 가지 SKU로 제공됩니다. 이 그래픽 프로세서는 최대 32개의 인텔® Xe 코어 및 레이 트레이싱 장치, 최대 4개의 인텔® Xe 미디어 엔진, 인텔® Xe Matrix Extensions(XMX)를 통한 AI 가속을 내장하고 있고, 하드웨어 기반 SR-IOV 가상화를 지원합니다. 인텔® 딥링크 하이퍼 인코딩(Deep Link Hyper Encode) 기능을 활용하는 2개의 GPU가 있는 Flex 시리즈 140은 8K 60 실시간 트랜스코딩을 제공하면서 업계의 1초 지연 요구 사항을 충족할 수 있습니다. 이 기능은 AV1 및 HEVC HDR 형식에 사용할 수 있습니다.

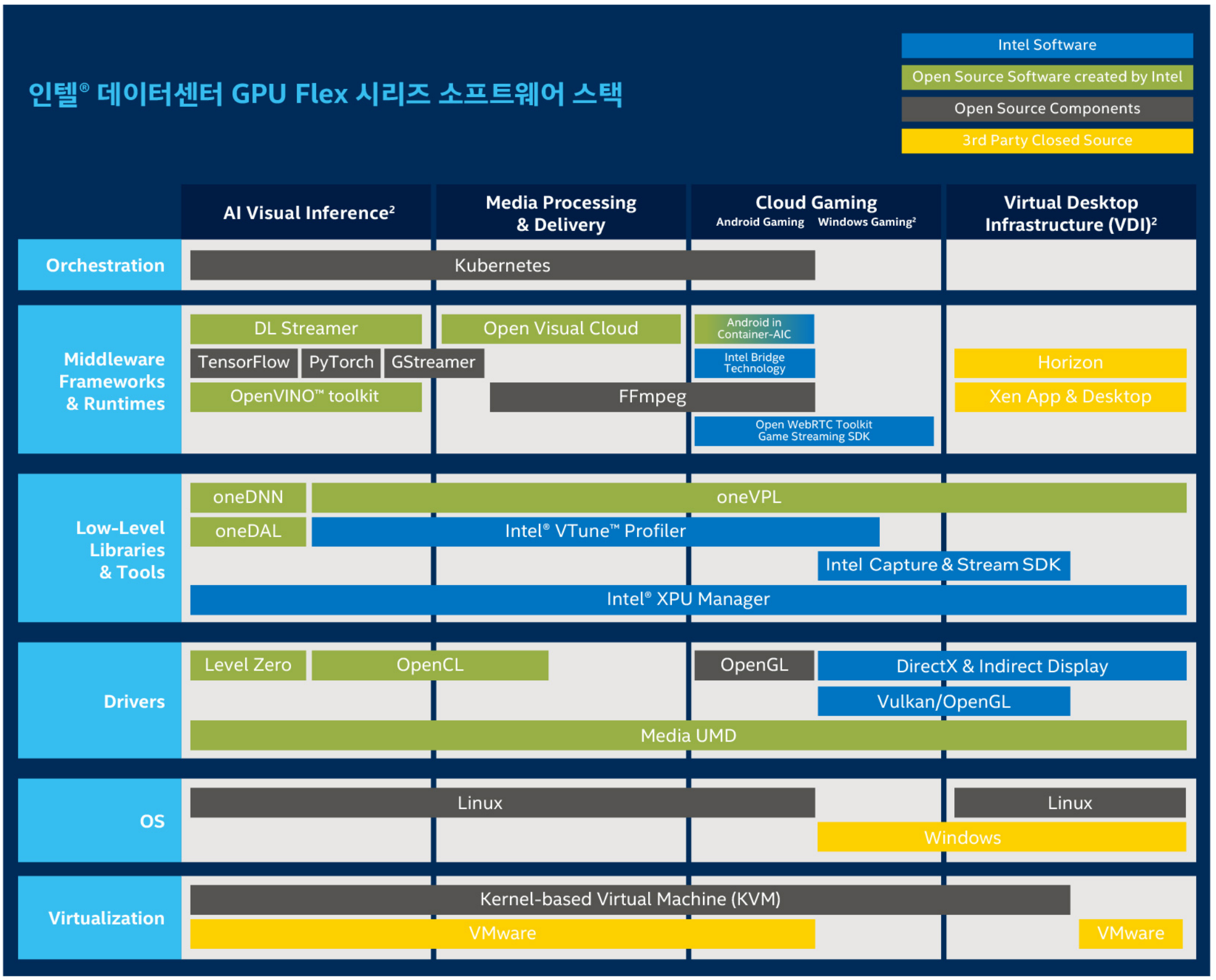
	인텔® 데이터 센터 GPU Flex 시리즈 140	인텔® 데이터 센터 GPU Flex 시리즈 170
대상워크로드	미디어 처리 및 전달, Windows 및 Android 클라우드 게임, 가상화된 데스크톱 인프라, AI 시각적 추론[2]	
카드 폼 팩터	Half height, half length, single wide, passive cooling	Full height, three-quarter length, single wide, passive cooling
카드 TDP	75와트	150와트
카드당 GPU	2개	1개
GPU 마이크로아키텍처	Xᵉ HPG	
Xe 코어	16개 (GPU 당 8개)	32개
고정 기능 미디어	4개 (GPU 당 2개)	2개
레이 트레이싱 지원	지원	
피크 컴퓨팅 (수축기)	8 TFLOPS (FP32) / 105 TOPS (INT8)	16 TFLOPS (FP32) / 250 TOPS (INT8)
메모리 유형	GDDR6	
메모리 용량	12 GB (6 per GPU)	16 GB
가상화 (인스턴스)[3]	SR-IOV (62)	SR-IOV (31)
운영체제	Linux (Ubuntu, CentOS, Debian), Windows Server 2019/2022, Windows Client 10, Red Hat® Enterprise Linux	
호스트 버스	PCIe Gen 4	
호스트 지원 CPU	3세대 인텔® 제온® 스케일러블 프로세서	

사용 케이스별 소프트웨어 스택

Flex 시리즈 GPU는 oneAPI 아키텍처 간 프로그래밍을 통해 유연한 개방형 표준 기반 소프트웨어 스택을 지원합니다.
스택에는 오픈 소스 구성 요소와 라이브러리, 도구 및 프레임워크가 포함되어 있으므로 개발자는 광범위한 사용 사례를 충족하는 고성능, 아키텍처 간 미디어 애플리케이션 및 솔루션을 만들 수 있습니다. 이 개방형 접근 방식은 코드 이식성과 여러 공급업체에 걸쳐 새로운 아키텍처를 채택할 수 있는 능력이 제한된 독점 모델에 대한 장벽을 제거합니다.

공통 소프트웨어 기능 세트는 널리 사용되는 미들웨어 및 프레임워크에 통합되며 스택은 검증된 제품화된 컨테이너 또는 참조 스택으로 제공됩니다. 컨테이너는 워크로드를 할당하고 관리하는 도구와 함께 SR-IOV 가상화를 사용하여 베어 메탈 또는 VM에서 쿠버네티스(Kubernetes)로 오케스트레이션될 수 있습니다. 이 도구 세트는 출시 시간을 단축하고 동일한 GPU에서 여러 워크로드를 유연하게 배포할 수 있도록 설계되었습니다.

인텔은 업계 협력, 이니셔티브 및 표준 기구를 통해 소프트웨어 에코시스템을 활성화합니다. 또한 오픈 소스 커뮤니티에 지속적인 리더십, 투자 및 기술 기여를 제공합니다.



참고: oneDNN은 oneAPI 심층 신경망 라이브러리입니다. oneDAL은 oneAPI 데이터 분석 라이브러리입니다. oneVPL은 oneAPI 비디오 처리 라이브러리입니다. oneVPL, oneDNN, oneDAL 및 Intel VTune 프로파일러는 인텔® oneAPI Base Toolkit에 있습니다(개별 도구는 별도로 다운로드할 수 있음). 인텔에 최적화된 TensorFlow 및 PyTorch는 인텔® AI Analytics Toolkit에 있습니다.

인텔® 데이터센터 GPU Flex 시리즈에 대한 더 자세한 내용은 아래의 사이트에서 자세히 알아보십시요.

https://www.intel.co.kr/content/www/kr/ko/products/docs/discrete-gpus/data-center-gpu/flex-series/overview.html

intel®

1 성능은 사용, 구성 및 기타 요인에 따라 다릅니다. 더 자세한 내용은 성능 지수 사이트(Performance Index site)에서 알아보십시오.
2 제품이 완전히 상용화되면 인텔 데이터센터 GPU Flex 시리즈의 성능을 반영합니다.
3 VM은 사용 사례에 따라 다릅니다.