제품 요약

인텔® AI Engines
5세대 인텔® 제온® 스케일러블 프로세서

5세대 인텔® 제온® 스케일러블 프로세서와 인텔® AI Engines로 전체 AI 파이프라인 성능 강화

65% 의 데이터 센터 AI 추론 **인텔® 제온® 프로세서로 실행**[1]

최대 **14배 더 높은** 실시간 개체 감지 추론 성능 (SSD-ResNet34), AMX BF16이 있는 5세대 인텔® 제온® 프로세서, 3세대 인텔® 제온® 프로세서 대비[2]

최대 **9.9배 더 높은** 실시간 자연어 처리 추론 (BERT-large) 성능과 7.7배 더 높은 성능/와트, AMX BF16이 있는 5세대 인텔® 제온® 프로세서, 3세대 인텔® 제온® 프로세서 대비[3]

최대 **8.7배 더 높은** 배치 권장 시스템 추론 성능(DLRM)과 6.2배 더 높은 성능/와트, 5세대 인텔® 제온®, 3세대 인텔® 제온® 프로세서 대비[4]

AI는 데이터 전처리와 클래식 머신러닝부터(ML) 자연어 처리와 이미지 인식 같은 딥러닝 용도에 이르는 광범위한 워크로드와 이용 사례에 사용됩니다. 인텔® 제온® 스케일러블 프로세서는 전체 AI 파이프라인에 강력한 컴퓨팅 성능을 제공합니다. 이 프로세서에는 ML, 데이터 분석 및 딥러닝 등 특정 AI 워크로드에 최적화된 내장형 가속기가 포함되어 있습니다.

전사에 걸친 AI를 위한 성능 내장

AI는 널리 퍼져 있으며, 다양한 중요 워크로드에 폭넓게 사용됩니다. 핵심 엔터프라이즈 애플리케이션부터 자동 음성 도우미까지, 클래식 ML과 딥러닝은 비즈니스가 완수되는 방법의 기본적인 구성 요소가 되어가고 있습니다. AI를 대규모로 사용하려면 데이터 전처리부터 학습을 거쳐 배포까지 이어지는 긴 개발 파이프라인이 필요합니다. 각 단계마다 자체적인 개발 툴체인, 프레임워크 및 워크로드가 있으며, 모두 고유한 병목 지점을 만들고 컴퓨팅 리소스에 각기 다른 부담을 더합니다. 인텔® 제온® 스케일러블 프로세서에는 전체 파이프라인을 실행하고 AI 성능을 전반에 걸쳐 높이기 위해 즉시 사용 가능한 가속기가 내장되어 있습니다.

인텔® Accelerator Engines는 특정 목적을 위해 만든 내장형 가속기로, 가장 까다로운 새로운 워크로드를 지원합니다

5세대 인텔® 제온® 스케일러블 프로세서는 범용 컴퓨팅에 탁월하고, 현재의 여러 중요 AI 워크로드를 지원하는 기초로 계속 사용될 것입니다. 이런 프로세서에는 딥러닝 추론과 학습을 CPU에서 가속하도록 설계된 내장형 AI 가속기인 인텔® Advanced Matrix Extensions (인텔® AMX)가 있습니다. 따라서 많은 경우에 분리형 가속기의 추가 비용과 복잡성을 배제할 수 있습니다. 인텔® 제온® 프로세서의 최신 세대는 고객 SLA를 일반적으로 준수하는 파라미터가 200억(20B) 개보다 적은 큰 언어 모델(LLM)에 매우 적합합니다.[5] 인텔® AMX는 전이 학습과 미세 조정에도 탁월하므로 모델 학습을 짧게는 (몇 시간이나 며칠이 아닌) 4분 만에 마칠 수 있으며, 하드웨어를 추가할 필요가 없습니다. 데이터 센터 추론의 65%가 인텔® 제온® 프로세서에서 실행되므로, 고객은 GPU 인프라로 전환하는 복잡한 과정을 거치지 않고 범용 AI를 위한 기존 아키텍처를 이용할 수 있습니다.

5세대 인텔® 제온® 스케일러블 프로세서와 인텔® Accelerator Engines로 지금 경험하는 미래 혁신

인텔® 제온® 프로세서를 워크로드에 온프레미스, 클라우드 또는 에지 중 어디에서 사용하든, 인텔® Accelerator Engines가 내장된 인텔® 제온® 프로세서의 도움으로 비즈니스 수준을 한 차원 더 높일 수 있습니다. 따라서 더 강력한 데이터 보호와 인프라 이용 개선 등 광범위한 이점을 얻을 수 있습니다.

고객 성공 사례: 인텔® 제온® 스케일러블 프로세서를 사용한 실제 가속

Tencent Cloud는 인텔® 제온® 스케일러블 프로세서를 사용하여 실시간 음성 합성을 실현합니다.

자세히 알아보기 ›

Gunpowder는 4세대 인텔® 제온® CPU를 사용해 Google Cloud C3 인스턴스를 실행하여 렌더링 성능을 가속합니다.

사례 읽기 ›

인텔® Accelerator Engines는 가상 및 물리적 CPU 이용률을 높이고 코어당 솔루션 라이선싱을 줄이는 데에도 도움이 될 수 있습니다. 무엇보다도 이런 내장형 가속기는 애플리케이션 성능을 높이고 비용을 줄이고 플랫폼 수준 효율을 개선합니다.

인텔® Advanced Matrix Extensions로 딥러닝 가속

인텔® AMX는 딥러닝 학습 및 추론에 사용되는 5세대 인텔® 제온® 스케일러블 프로세서의 최신 발전 기술입니다. 자연어 처리, 추천 시스템 및 이미지 인식 같은 워크로드에 적합한 인텔® AMX는 고객이 5세대 인텔® 제온®(AMX BF16 포함)으로 3세대 인텔® 제온® 프로세서 대비 최대 7.2배 더 높은 실시간 개체 분류 추론 성능과 5.3배 더 높은 성능/와트를 실현하는 데 도움이 됩니다.[6]

또한 인텔® AMX는 AI 모델을 위한 워크로드 부스트를 제공하며, 더 많은 고객이 이미 실행 중인 플랫폼에서 SLA를 충족할 수 있도록 지원합니다. 5세대 인텔® 제온® 스케일러블 프로세서는 향상된 터보 주파수 (5단계의 터보 비율)를 추가함으로써 고성능 컴퓨팅 및 AI를 포함하여 벡터 및 행렬 연산 친화적인 워크로드를 개선할 수 있습니다.

인텔® AMX는 CPU 코어의 인텔® Advanced Vector Extensions 512(인텔® AVX-512)보다 더 높은 처리량(연산/주기)으로 행렬 곱셈 연산 성능을 개선합니다.[7] 이를 통해 딥러닝 학습 워크로드를 더 빨리 완료하기가 쉬워지고 더 많은 고객이 이미 비즈니스를 운영하는 데 사용하는 플랫폼에서 SLA를 준수할 수 있습니다.

자연어 처리와 생성형 AI 지원

인텔® AMX가 탑재된 5세대 인텔® 제온® 스케일러블 프로세서는 추가 하드웨어 없이도 자연어 처리 성능을 크게 높입니다. 인텔® 라이브러리는 TensorFlow와 PyTorch에 최적화되고 통합되므로, 개발자는 내장형 AI 가속의 이점을 즉시 얻을 수 있습니다. 개발자는 다른 하드웨어 환경에서 코드를 더 쉽게 마이그레이션(긴 시간과 많은 비용이 모두 소요될 수 있는 프로세스)할 수도 있습니다.

인텔® AMX가 탑재된 5세대 인텔® 제온® 스케일러블 프로세서는 딥러닝 추론과 학습을 가속하여 SLA를 준수하면서 총소유비용(TCO)의 균형을 맞추는 데에도 도움이 됩니다. 이 프로세서는 실시간 사용자 행동 신호와 시간 및 위치 같은 추가적인 상황 특성을 감안하는 딥러닝 기반 추천 시스템을 사용하여 이를 실현합니다.

5세대 프로세서는 인간 중심 컨텐츠를 모방하는 생성형 AI 모델도 실행하여 큰 언어 모델과 텍스트-이미지 생성을 지원합니다. 보다 집약적인 생성형 AI 작업에는 특정 목적을 위해 만든 인텔® Gaudi® AI 가속기, 인텔® Data Center GPU 및 기타 하드웨어 구성요소를 사용하여 CPU의 기능을 확장할 수 있습니다.

인텔® AVX-512로 ML을 더 빠르게

인텔® 제온® 프로세서는 웹 사이트를 위해 SSL 암호화를 해싱하고, 거대한 데이터베이스를 압축하고, 제약 연구, 칩 설계 또는 Formula 1 엔진 시뮬레이션을 실행할 수 있습니다.

여러 세대에 걸쳐 개선된 인텔® AVX-512를 통해, 인텔® 제온® 스케일러블 프로세서는 각 클럭 사이클에 연산을 더 많이 처리하고 병렬 처리 애플리케이션의 성능을 개선할 수 있습니다. 인텔® AVX-512 명령어 집합 아키텍처(ISA)는 AI, HPC, 네트워킹 및 스토리지에 걸친 다양한 워크로드의 성능을 높이기 위해 만든 확장 기능을 포함합니다.

새 프로세서 세대에서는 터보 비율 단계가 네 개에서 다섯 개로 늘어나 터보 성능이 향상됩니다. 따라서 인텔® AMX와 인텔® AVX-512를 이용하는 특정 HPC 및 AI 워크로드의 터보 주파수가 개선됩니다.

더 적은 단계는 더 빠른 처리를 의미

수학은 매우 스마트하고 아주 명쾌할 수 있습니다. 5세대 인텔® 제온® 스케일러블 프로세서의 인텔® AVX-512는 여러 스마트하고 아름다운 수학을 사용해 일반적인 컴퓨팅 작업을 더 적은 단계로 압축하고 결합하고 융합합니다. 간단한 예를 들어보겠습니다. CPU에 3 x 3 x 3 x 3 x 3을 계산하라고 명령할 수 있는데, 이 계산에는 다섯 클럭 사이클이 소요됩니다. 아니면 CPU가 한 사이클에 수행할 수 있는 3^5을 계산하라는 명령어를 만들 수도 있습니다. 인텔® AVX-512는 이 논리를 AI에서 가장 까다로운 몇몇 작업을 포함한 수백 개의 워크로드별 작업에 적용합니다.

8개 단위로 세면 1개 단위로 세는 것보다 훨씬 빠릅니다.

인텔® AVX-512에서 "512"는 이런 명령어가 클럭 사이클마다 CPU가 처리할 수 있는 비트 수를 늘리는 두 번째 방법과 관련이 있습니다. 40년 전에는 16비트 PC가 아주 대단하다고 인식되었지만, 곧 32비트 컴퓨터로 대체되었습니다. 이제는 스마트폰이 64비트로 작동합니다. 비트 수는 CPU가 1 클럭 사이클마다 어드레싱할 수 있는 레지스터(CPU가 데이터를 보관하는 메모리 슬롯)의 수를 의미합니다. 이름이 시사하듯, 인텔® AVX-512는 레지스터 수를 512 비트로 확장합니다. 애플리케이션이 인텔® AVX-512를 이용할 경우, 레지스터 수를 늘려 CPU의 기본 64비트 속도보다 최대 8배까지 더 빠르게 실행됩니다. 96까지 하나씩(1, 2, 3...) 세는 것과 8 단위(8, 16, 24...)로 세는 것의 차이와 비슷합니다.

필요한 성능을 낮추면서 더 강력한 AI를 실행하는 엔진

인텔® AI Engines가 탑재된 인텔® 제온® 스케일러블 프로세서는 하드웨어 리소스를 덜 필요로 하므로, AI 워크로드를 실행하는 데 더 강력하고 에너지 효율이 높은 솔루션을 제공합니다.

가속기 엔진이 내장된 인텔® 제온® 스케일러블 프로세서는 지금의 까다로운 AI 워크로드와 관련된 TCO를 낮추고 투자 수익(ROI)을 개선하는 등 워크로드 결과를 개선하는 데에도 도움이 될 수 있습니다.

인텔® 제온® 프로세서를 사용하면 AI가 사실상 자동으로 더 빨라집니다.

인텔® 제온® 스케일러블 프로세서에서, AI 가속은 CPU의 명령어 집합 아키텍처(ISA)에 내장됩니다. 다시 말해, 이 CPU는 그것을 이용할 수 있는 어떤 소프트웨어에나 사용할 수 있게 준비되어 있습니다. 인텔 소프트웨어 엔지니어들은 오픈 소스 AI 툴체인을 계속 최적화하고 이런 최적화를 커뮤니티에 다시 전달하고 있습니다. 예를 들어 TensorFlow 2.9는 기본적으로 인텔® oneAPI Deep Neural Network Library(인텔® oneDNN) 최적화와 함께 제공됩니다. 최신 에디션을 다운로드하면 TensorFlow가 인텔® 최적화를 자동으로 이용합니다.

데이터 과학자와 개발자는 인텔® 제온® 스케일러블 프로세서의 ISA에 내장된 가속기를 모두 이용하는 무료 오픈 소스 인텔® 배포판, 라이브러리 및 개발 환경을 다운로드하여 AI 파이프라인의 다른 애플리케이션에 사용할 수 있습니다. 데이터 과학자와 AI 개발자는 도구를 다시 코딩하고 인텔® AVX-512에 사용할 수 있게 다시 컴파일링할 필요가 없습니다. 자동으로 되기 때문입니다.

지금의 조직은 더 높은 전력 효율과 더 낮은 비용으로 인프라에서 더 높은 워크로드 성능을 얻어야 합니다. 인텔® 제온® 스케일러블 프로세서에 통합되고 특정 목적을 위해 만든 인텔® AI Engines는 비즈니스에 가장 중요한 AI 워크로드에서 가장 많은 것을 얻어내는 데 도움이 될 것입니다.

인텔® Accelerator Engines가 내장된 인텔® 제온® 스케일러블 프로세서가 비즈니스에 가장 중요한 AI 워크로드를 위해 무엇을 할 수 있는지 자세히 알아보십시오.

자세한 내용

인텔® 제온® 스케일러블 프로세서 기반 AI 및 딥러닝 ›

인텔® AVX-512 ›

인텔® AI 분석 툴킷 ›

인텔® 하드웨어 및 소프트웨어를 사용한 개발 ›

AI와 ML을 위한 인텔® 최적화로 지금 클라우드 또는 자체 인프라에서 AI 워크로드를 가속하기 시작하십시오.

자세한 내용 ›

---

intel® XEON®

1. 2022년 12월 현재 AI 추론 워크로드를 실행하는 전세계 데이터 센터 서버 설치 기반의 인텔 시장 모델링 기준입니다.
2. intel.com/processorclaims: 5세대 인텔® 제온® 스케일러블 프로세서에서 [A21]을 참조하십시오. 결과는 다를 수 있습니다.
3. intel.com/processorclaims: 5세대 인텔® 제온® 스케일러블 프로세서에서 [A19]를 참조하십시오. 결과는 다를 수 있습니다.
4. intel.com/processorclaims: 5세대 인텔® 제온® 스케일러블 프로세서에서 [A20]을 참조하십시오. 결과는 다를 수 있습니다.
5. 2023년 12월 기준 인텔 내부 모델링을 기반으로 합니다.
6. intel.com/processorclaims: 5세대 인텔® 제온® 스케일러블 프로세서에서 [A22]를 참조하십시오. 결과는 다를 수 있습니다.
7. https://edc.intel.com/content/www/kr/ko/products/performance/benchmarks/vision-2022/, 세션 벤치마크 41번, 42번. 결과는 다를 수 있습니다.

고지 및 면책 조항

성능은 사용, 구성 및 기타 요인에 따라 다릅니다. 자세한 내용: Performance Index 사이트.

성능 결과는 구성에 표시된 날짜의 테스트를 기반으로 하며 공개된 모든 업데이트가 반영되어 있지 않을 수도 있습니다. 구성 백업 상세 정보를 확인하십시오. 어떤 제품 또는 구성 요소도 절대적으로 안전할 수는 없습니다.

비용과 결과는 다를 수 있습니다.

워크로드 및 구성에 대해서는 www.intel.com/processorclaims에서 5세대 인텔® 제온® 스케일러블 프로세서를 참조하십시오. 결과는 다를 수 있습니다.

인텔 기술은 지원되는 하드웨어, 소프트웨어 또는 서비스 활성화가 필요할 수 있습니다.

© 인텔사. 인텔, 인텔 로고 및 기타 인텔 마크는 인텔사 또는 그 자회사의 상표입니다. 기타 명칭 및 브랜드는 해당 소유주의 자산일 수 있습니다.

인텔은 타사 데이터를 제어하거나 감사하지 않습니다. 정확성 평가를 위해서는 기타 소스를 참고해야 합니다.

가속기의 가용성은 SKU에 따라 다릅니다. 제품 세부 정보를 더 보려면 인텔® 제품 사양 페이지를 방문하십시오.

인텔® Advanced Vector Extensions(인텔® AVX)는 특정 프로세서 작업에 더 높은 처리량을 제공합니다. 여러 프로세서 성능 특성으로 인해, AVX 명령어를 이용하면 a) 일부분이 정격 주파수 미만으로 작동하고, b) 인텔® Turbo Boost Technology 2.0이 있는 일부분이 임의의 또는 최대 터보 주파수에 도달하지 못하는 원인이 될 수 있습니다. 성능은 하드웨어, 소프트웨어 및 시스템 구성에 따라 다르며, 자세한 내용은 다음 웹 페이지에서 확인할 수 있습니다. https://www.intel.co.kr/content/www/kr/ko/products/details/processors/core.html

인텔은 인권을 존중하고 인권 침해에 연루되지 않도록 하기 위해 노력합니다. 인텔의 글로벌 인권 원칙을 참조하십시오. 인텔® 제품과 소프트웨어는 국제적으로 인정되는 인권의 침해를 초래하거나 악화시키지 않는 애플리케이션에만 사용해야 합니다.