그래프코어 소프트웨어 스택: 확장을 위한 개발

소개

새로운 프로세서 설계를 위한 소프트웨어는 애플리케이션을 배포하고 성능을 최적화하는 데 매우 중요합니다. 영국에 본사를 둔 스타트업인 그래프코어는 애플리케이션 가속화를 위한 실리콘 공급업체로서 소프트웨어에 상당한 비중을 두고 있으며, 전체 엔지니어링 인력의 절반 정도가 여기에 투입되고 있습니다. 그래프코어의 IPU(Intelligence Processing Unit)는 알고리즘의 표현식을 방향성 그래프(directed graph)로서 사용하며, 포플러(Poplar) 소프트웨어 스택은 모델과 알고리즘을 이들 그래프로 변환해 실행합니다.

이 소프트웨어는 AI 및 병렬 컴퓨팅을 위한 칩 채택을 단순화하며 이는 그래프코어의 성공에 핵심적인 역할을 하고 있습니다. 이 백서는 그래프코어의 소프트웨어가 제공하는 이점들에 대해 살펴보고, 이들 기능들이 어떻게 IPU에서 실행되는 애플리케이션의 개발 및 설치 시간을 단축하는지에 대해 밝힙니다.

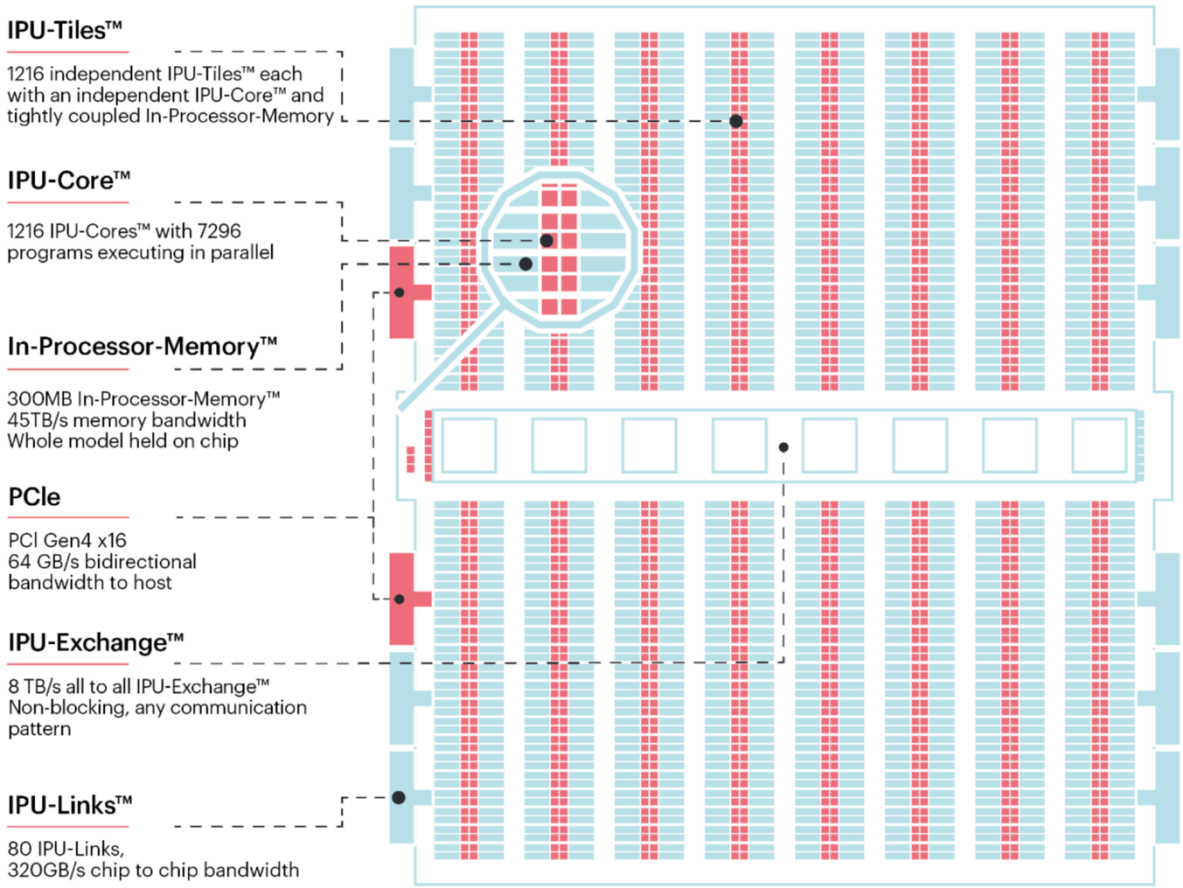
IPU(지능처리장치, INTELLIGENCE PROCESSING UNIT)의 개요

소프트웨어를 논의하기에 앞서, 주요 하드웨어에 대한 기본적인 설명이 도움이 될 수 있습니다. 그래프코어의 IPU(Intelligence Processing Unit)는 오늘날 CPU, GPU 및 기타 AI 프로세서와는 근본적으로 다릅니다. IPU와 포플러 소프트웨어는 엄격하게 머신러닝(machine-learning) 툴 세트는 아닙니다. IPU는 유연하고 확장성이 뛰어난 파인 그레인드(fine-grained) 병렬 프로세서로서 다양한 계산 집약적 알고리즘에 고성능을 제공하도록 설계되었습니다. IPU와 포플러는 함께 금융, 초고성능 컴퓨팅(HPXC: High Performance Computing), 로봇 공학 및 데이터 과학 등에서 활용할 수 있으며, 머신 인텔리전스 워크로드를 지원하는 그래프-프로그래밍 플랫폼을 형성합니다.

IPU 설계 목표는 대부분 ASIC 및 GPU에서 볼 수 있는 기존 가속화 아키텍처의 한계를 벗어난 문제를 해결하는 것이었습니다. 일반적으로 이들 칩은 밀집 선형대수학(dense linear-algebra) 워크로드에 최적화되어 있습니다. 이는 일부 CNN(Convolutional Neural Networks)에 구현될 수 있지만, 다양한 계산, 통신 또는 데이터 액세스 패턴을 가진 애플리케이션에는 적합하지 않습니다.

IPU는 1,216개의 상호 연결된 프로세싱 타일로 구성되어 있습니다. 각 타일에는 자체 코어 및 로컬 온다이(on-die) SRAM 메모리가 있어 모델과 데이터가 IPU에 상주할 수 있으며, 이에 따라 메모리 대역폭과 지연시간이 크게 향상됩니다. 이들 타일은 8 TB/s 온다이 패브릭("IPU-Exchange")을 통해 상호 연결되며, 또한 320 GB/s의 "IPU-Links"를 통해 칩 간(chip-to-chip) 패브릭을 형성합니다.

그림 1: 그래프코어 인텔리전트 프로세서



그래프코어의 IPU에는 1,216개의 "타일"이 있으며 각 타일은 계산 코어와 메모리를 탑재하고 있습니다. 이들 타일은 모두 고속, 저지연시간 "익스체인지(Exchange)" 패브릭으로 연결되어 있으며, 해당 칩에서 확장해 수천 개의 칩에 멀티 칩(multi-chip) 병렬 처리를 실행할 수 있도록 합니다.

출처: 그래프코어

기본적으로 전체 시스템(흔히 많은 IPU로 구성됨)은 계산 및 통신 등 2개의 동기식 페이즈(synchronous phase)로 실행됩니다. IPU를 대상으로 하는 애플리케이션은 계산 그래프로서 표현됩니다. 계산은 그래프의 정점(vertices)에서 실행되며 그 결과는 그래프와 상호 연결하는 간선(edge)에 따라 인접 정점(adjacent vertices)에 전달됩니다. 통신 페이즈(communication phase)는 BSP(Bulk Synchronous Parallel) 연산으로 실행되며, 이를 통해 각 타일의 온다이 SRAM 메모리의 데이터가 연결된 타일의 메모리로 효율적으로 전송됩니다. 계산 명령 이외에도 각 IPU 코어에는 BSP 모델의 통신 페이즈를 위한 전용 타일 레벨 명령 세트가 있습니다.

통합 교환-통신(exchange-communication) 패브릭은 데이터 및 모델 병렬 처리 모두 — 그래프 컴파일러에 의해 실행됨 — 를 위해 BSP를 지원하도록 설계되었으며, 수천 대의 노드로 확장 가능합니다. 그래프코어는 IPU 아키텍처의 중요한 차이점이 희소 데이터 및 그래프의 효율적인 처리 성능이라고 밝히고 있습니다. 이는 성능을 향상시키는 동시에 총 메모리 요구 사항을 줄입니다.

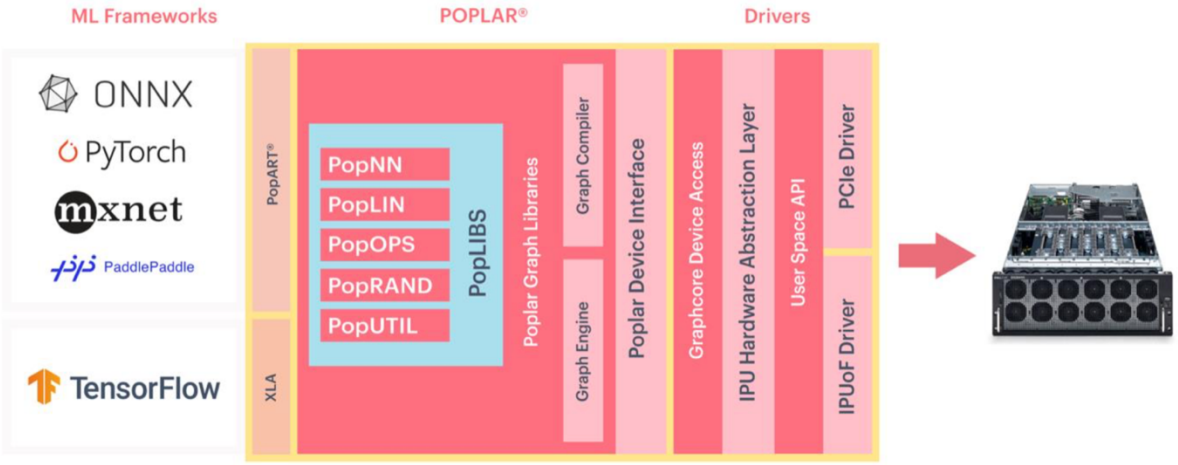
그래프코어의 포플러 SDK(Poplar SDK)는 워크플로우의 컴파일 및 최적화를 자동화함으로써 이 고유한 아키텍처가 야기한 프로그래밍 과제를 해결하기 때문에, 수작업으로 조정된 명령 레벨 프로그래밍 없이 IPU 패브릭의 파인 그레인드 병렬 처리를 활용할 수 있습니다. 직접 하드웨어 액세스도 지원됩니다.

그래프코어 소프트웨어 플랫폼

개요

그래프코어는 자체 칩을 위한 소프트웨어 개발로 2가지 과제를 해결하고 있습니다. 1) 고수준 프레임워크에 표현된 심층 신경망 또는 dDNN 등과 같은 기존 머신러닝 소프트웨어를 손쉽게 최적화 및 실행하고 2) IPU 인프라에서 실행할 수 있도록 완전히 새로운 파인 그레인드 병렬 워크플로우를 연구 및 개발할 수 있도록 합니다. 두 번째 기능은 회사 전략의 핵심으로서, 그래프코어가 보다 폭 넓은 시장 부문에 진출할 수 있도록 했습니다.

그림 2: 그래프코어 소프트웨어 플랫폼



그래프코어 소프트웨어 플랫폼은 인기 있는 많은 AI 프레임워크에서 모델을 가져와서 IPU에서 실행할 수 있도록 코드를 최적화합니다. 이는 새로운 신경망과 그래프로 표현될 수 있는 기타 알고리즘을 개발하기 위해 최적화된 코드의 커스텀 정점(custom vertices) 또는 애플릿(applets)의 작성 또한 지원합니다. 출처: 그래프코어

그림 2에서 볼 수 있듯이, 포플러는 그래프 및 요소 컴파일러(element compiler), 최적화된 라이브러리 그리고 런타임 관리 및 스케줄링을 위한 그래프 엔진으로 구성되어 있습니다. 머신러닝 프레임워크는 ONNX(Open Neural Network eXchange)을 위한 PopART(Poplar Advanced Run Time) 인터페이스는 물론, 텐서플로우(TensorFlow) 기반 모델을 위한 XLA(Accelerated Linear Algebra) 컴파일러를 통해 포플러 스택을 지원합니다. 그래프코어는 2020년 말까지 파이토치(PyTorch)에 대한 직접 지원이 제공될 것이라고 밝혔습니다.

이제 그래프코어의 소프트웨어 플랫폼이 어떻게 IPU를 위한 애플리케이션 개발을 지원하는지 살펴볼 것입니다. 특히, 프레임워크, 컴파일러, 런타임 지원 및 라이브러리 등 주요 구성 요소들에 대해 알아보고, 성능 및 채택을 가속화할 수 있는 가능성을 생각해 보도록 하겠습니다.

개방형 프레임워크(OPEN FRAMEWORKS)를 이용한 모델 포팅(Porting) 및 개발

새 AI 액셀러레이터를 평가하려면, 딥러닝 과학자는 먼저 텐서플로우나 파이토치와 같은 표준 프

레임워크를 이용해 기존 DNN을 포팅, 훈련 및 테스트한 다음, 그 결과(훈련 시간 및 정확성)를 기존 플랫폼과 비교합니다. 그래프코어의 소프트웨어 부서는 IPU에서 해당 신경망을 간편하게 훈련 및 실행해 신규 솔루션의 프로토타입을 작성할 수 있도록 지원합니다. 개발자들은 커스터마이징된 레이어 유형과 새로운 라이브러리 함수로 표준 네트워크 레이어를 확장할 수 있으며, 이는 향후 오픈소스 프레임워크에 포함 가능합니다. 해당 기능을 통해 기존 레이어도 확장 또는 향상시킬 수 있습니다.

포플러 DNN 프레임워크 범위를 벗어난 새로운 애플리케이션을 위해 그래프코어는 IPU를 위한 맞춤형 프레임워크를 개발했습니다. 그래프 프레임워크(Graph Framework)는 기존 프레임워크와 경쟁하기 위한 것이 아닙니다. 반대로, 이들 그래프 아키텍처에서 완전히 새로운 병렬 워크로드를 실행되도록 하는 데 이용할 수 있습니다.

그래프코어는 내년에 오픈소스에 자사 소프트웨어를 공개할 예정이며 IPU SDK 다운로드는 이미 포플러 라이브러리 소스 코드를 포함하고 있습니다. 오픈소스 전략은 IPU 플랫폼에 대한 연구 개발에 대한 적극적인 업계 참여를 확대함으로써 커뮤니티가 새 라이브러리와 레이어를 개발하고 배포할 수 있도록 하므로 그래프코어에게 매우 중요한 방향성을 제시합니다.

실행 최적화: 포플러 컴파일러, 라이브러리 및 그래프 엔진

그래프 및 그래프 엘리먼트 컴파일러

개발 프로세스는 그래프 컴파일러(Graph Compiler)로 시작되며, 배포를 위한 모델을 개발하고, IPU에서 실행되는 그래프 엘리먼트(Graph Elements, 커널 또는 "코드 조각(codelets)")를 가져옵니다. 여기에서는 계산을 컴파일링(그래프 정점)하고 BSP 통신(그래프를 위한 간선)을 구현하기 위해 코드를 생성하는 2 단계로 진행됩니다.

런타임 배포 및 최적화의 핵심인 그래프 컴파일러(Graph Compiler)는 지난 5년 여 간 개발 중이었습니다. 이는 코드 재사용을 최적화하고 데이터 이동을 최소화하며 데이터 지역성(locality)을 활용합니다. 이 접근 방법은 대형 연속 메모리 액세스 패턴에 더 적합한 다른 아키텍처에 부담을 주는 희소 모델에서 IPU가 뛰어난 성능을 발휘할 수 있도록 지원합니다. 그래프코어는 특히 많은 IPU 전반에 작업이 배포되는 경우, IPU 프로그래밍을 단순화하도록 이 컴파일러를 설계했습니다. 중요한 것은 이 컴파일러가 데이터를 관리하거나 병렬 처리를 모델링해야 하는 개발자들의 부담

을 줄여 준다는 것입니다.

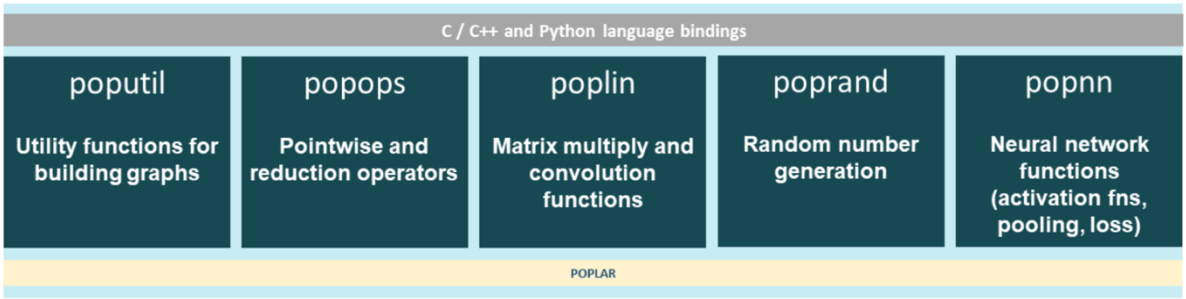
가령 서버 내 16개 개별 ASIC이나 GPU에 프로그래밍하는 대신, 포플러 그래프 컴파일러는 단일 "멀티-IPU"를 대상으로 하기 때문에 개발자들은 자체 데이터 및 알고리즘에 집중할 수 있습니다. 이는 대부분의 다른 액셀러레이터를 능가하는 확실한 강점입니다.

포플러 엘리먼트 컴파일러는 백엔드로서, 코드 계산을 요소(element)로 컴파일하여 정점에서 실행합니다. 따라서 워크플로우는 다음과 같은 형태로 이루어집니다.

텐서플로우 프론트엔드 -> XLA -> 포플러 그래프 컴파일러 ->포플러 엘리먼트 컴파일러

IPU에는 유연한 전체 명령어 세트가 있기 때문에 타일들은 기본적으로 모든 코드 또는 알고리즘을 실행할 수 있으며, 그래프 엘리먼트는 LLVM 기반 컴파일러를 이용해 C/C++로 작성하거나 IPU 어셈블리로 직접 작성할 수 있습니다. 모든 포플러 라이브러리(Poplar Libraries)는 이 컴파일 툴 세트를 이용해 개발됩니다.

그림 3: 그래프코어의 포플러 라이브러리



C/C++ 및 Python 언어 바인딩				
poputil	popops	poplin	poprand	popnn
그래프 작성을 위한 유틸리티(Utility) 함수들	점별(pointwise) 및 리듀스(reduce) 연산자	행렬 곱셈(matrix multiply) 및 컨벌루션(convolution) 함수	난수(random number) 생성	신경망 함수들 (activation fns, pooling, loss)
포플러				

그래프코어는 공통적으로 필요한 함수 및 연산자들을 위해 50개 이상의 프리미티브(primitive)를 개발했습니다. 사용자들은 신규 라이브러리를 추가해 새로운 워크로드가 개발될 경우 이를 지원하고, 오픈소스 커뮤니티에 기여할 수 있습니다.

출처: 그래프코어

포플러 라이브러리

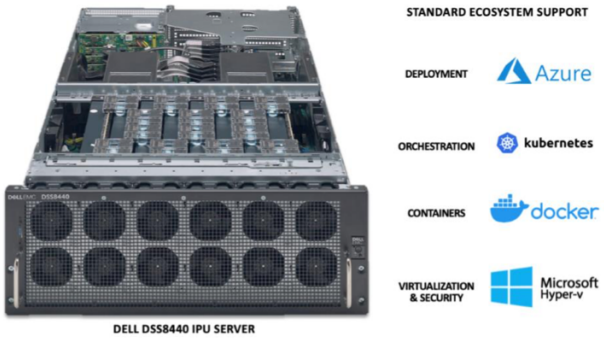
포플러 라이브러리는 선형 대수학, 공통 신경망 함수 및 머신 인텔리전스 모델에서 사용되는 기타 연산 등 공통의 연산을 대상으로 하는 50개 이상의 고도로 최적화된 프리미티브와 빌딩 블록을 포함하고 있습니다. 각 함수는 750개 이상의 고성능 컴퓨팅 요소 또는 코드 조각의 방대한 라이브러리에서 빌딩 블록을 결합하여 작동합니다. 포플러 라이브러리는 사용자 애플리케이션에서 높은 수준의 그래프 설명을 사용하며 IPU를 대상으로 하는 데 필요한 대용량 병렬 계산 그래프를 작성합니다. 해당 프로세스의 경우 IPU의 분산 프로세서 및 메모리 아키텍처를 목표로 하는 작업 및 데이터 파티셔닝이 포함됩니다. 포플러 라이브러리는 IPU에서 실행 가능하도록 세밀하게 조정된 각 연산을 위해 커스텀 레이아웃을 생성합니다.

라이브러리는 IPU 아키텍처에서 실행 가능한 신규 애플리케이션을 위한 혁신적인 커널 및 DNN 레이어를 퍼블리싱하고 실행할 수 있도록 지원합니다.

그래프 엔진

포플러 그래프 엔진(Poplar Graph Engine)은 IPU를 위한 런타임 지원을 제공합니다. 이러한 지원에는 호스트 CPU의 소프트웨어를 IPU에서 실행되는 소프트웨어와 연결하고 데이터 이동을 관리하며 I/O, 애플리케이션 로딩, 디버깅 및 프로파일링 등을 위해 IPU 장치를 관리하는 것 등이 포함됩니다. CPU에서부터 데이터 흐름을 최적화하는 것이 최대 처리량(throughput)을 달성하는 데 중요하며, 그래프 엔진은 데이터 I/O 파이프라인을 오케스트레이션함으로써 연속 실행이 이루어지도록 합니다. 그래프 엔진은 디버깅 및 프로파일링을 수행하는 IPU상의 맞춤형 하드웨어에 의해 지원됩니다.

그림 4: 그래프코어 에코시스템 지원



그래프코어는 스타트업에서 보기 드문 배포 기능에 대한 지원을 포함하고 있습니다. 출처: 그래프코어

도입: 업계 표준 툴 및 플랫폼

그래프코어 소프트웨어는 퍼블릭 클라우드 및 온프레미스 인프라 전반에서 생성 워크로드의 효율적 배포하도록 확장됩니다. 특히, 델 테크놀로지스(Dell Technologies)은 서버용 듀얼 IPU 그래프코어 C2 PCIe 카드를 판매하고 있으며 Microsoft Azure 및 Cirrascale은 IPU 클라우드 인스턴스(Azure의 프리뷰에서)를 지원합니다. 관리 및 오케스트레이션을 위해 그래프코어는 쿠버테니스(Kubernetes), 도커(Docker), 마이크로소프트 하이퍼-V(Microsoft Hyper-V) 등 업계 전반에서 표준으로 자리매김한 가상화, 보안 및 오케스트레이션 툴을 지원합니다. 그래프코어는 이와 같이 포괄적인 배포 소프트웨어 및 인프라 제품군을 포함해 자체 솔루션을 확장한 업계 유일의 스타트업 기업입니다.

결론: 그래프코어, 대규모 운영 환경을 위한 구축

신경망 훈련 시, 개발자들은 자연어 처리 및 대화형 인터페이스 등 대형 모델에서 요구하는 대용량 계산 부하를 처리할 수 있는 빠르고 확장 가능한 액셀러레이터를 필요로 합니다. 보다 폭 넓게 보면, 신경망에서 일반적인 행렬 곱셈 연산 이상으로 고속 병렬 프로세서를 필요로 하는 계산 집약적인 애플리케이션들이 등장하고 있습니다.

그래프코어는 관련 요구사항을 충족하기 위해 이와 같은 다목적 고성능 하드웨어 및 소프트웨어 플랫폼을 개발한 것으로 보입니다.

다양한 영역에서 주목 받는 그래프코어의 소프트웨어 스택:

1. 포플러 그래프 컴파일러는 "멀티-IPU" 병렬 처리를 실행하고 효율적인 메모리 사용 및 데이터 이동성을 구현합니다.

2. 그래프 프레임워크(Graph Framework)는 전통적인 머신러닝 프레임워크의 범위를 벗어난 영역에서 등장한 새로운 유형의 워크로드, 특히 새로운 알고리즘(그래프)을 구현합니다.

3. 새로운 레이어 및 커널을 구현할 수 있는 커스텀 애플릿으로 사용자들은 오픈소스 머신러닝 프레임워크의 전체 지원 및 최적화 기능을 확장 가능합니다.

4. 그래프코어의 오픈소스 전략은 다양한 분야의 광범위한 연구원 커뮤니티 전반에서 기업의 영향력을 확장하는 데 도움을 줄 수 있습니다

그래프코어는 데이터센터들이 중요시하는 컨테이너화, 오케스트레이션, 보안 및 가상화를 제공하기 위해 관리 소프트웨어 분야에서 다음 단계에 착수했습니다. 이는 그래프코어의 하드웨어 설계 및 소프트웨어 스택과 더불어 점차 많은 애플리케이션들이 그래프코어 플랫폼에 배포됨에 따라 간편한 도입에 박차를 가하게 될 예정입니다.

그래프코어가 직면한 도전 과제 중 하나는 자사 플랫폼과 기타 대체 제품들 간의 비교입니다. 그래프코어가 목표로 두는 새로운 유형의 워크로드가 아직 업계 내 보편적인 도입이 이루어지지 않았고, 타업체 플랫폼이 매끄럽게 지원하지 한다는 점에서 이는 자연스러운 결과일 수 있습니다. 그래프코어는 IPU가 레스넷(Resnet)과 같은 비교적 '단순한' 모델에서 대략 2~4배 더 빠르게 실행될 수 있다고 보는 반면, 새로운 워크로드에서 보다 탁월한 성능을 보여주는 결과를 자체 웹 사이트에 게시했습니다. 1 그래프코어의 혁신은 여기에서 시작됩니다.

1 https://cdn2.hubspot.net/hubfs/729091/assets/pdf/Benchmarks_slides_May2020-comp.pdf

백서 관련 주요 정보

작성자

칼 프로인드(Karl Freund)/무어인사이트앤스트레티지 수석 애널리스트

발행인

패트릭 무어헤드(Patrick Moorhead)/무어인사이트앤스트레티지 설립자, 대표 겸 대표 애널리스트

문의사항

해당 보고서와 관련해 의견이 있으시면 언제든지 연락 주십시오. 무어인사이트앤스트레티지((Moor Insights & Strategy)에서 빠른 시간 내에 회신 드리겠습니다.

인용

본 백서를 언론 및 애널리스트들이 인용할 수 있지만, 반드시 본문의 문장 안에 내주로 표시하고 작성자의 이름, 직책 및 "무어인사이트앤스트레티지(Moor Insights & Strategy)"를 명시해야 합니다. 언론 및 애널리스트 이외에는 모든 인용에 대해 무어인사이트앤스트레티지의 사전 서면 허가를 받아야 합니다.

라이선스

정보 공개

본 백서는 그래프코어의 의뢰로 작성되었습니다. 무어인사이트앤스트레티지는 본 백서에서 언급된 많은 하이테크 업체들에게 연구, 분석, 자문 및 컨설팅 서비스를 제공하고 있습니다. 무어인사이트앤스트레티지의 직원 중 어느 누구도 본 백서에 언급된 그 어떤 기업과도 지분을 보유하고 있지 않습니다.

면책조항

본 자료에 나온 정보는 오직 정보 제공 목적으로만 제공되며 기술적 오류, 누락 및 인쇄상의 오류가 있을 수 있습니다. 무어인사이트앤스트레티지는 해당 정보의 정확성, 완전성 또는 적절성에 대한 모든 보증을 부인하며, 해당 정보의 오류, 누락 또는 부적절성에 대해 그 어떠한 책임도 지지 않습니다. 본 자료는 무어인사이트앤스트레티지의 의견으로 구성되며 사실의 진술로 해석해서는 안됩니다. 여기에 표현된 의견은 사전 통지 없이 변경될 수 있습니다.

무어인사이트앤스트레티지는 미래 사건에 대한 정확한 예측이 아닌 방향성 지표로서 예측 및 미래 전망 진술을 제공합니다. 자사의 예측 및 미래 전망 진술은 미래에 대한 현재의 판단을 나타내지만, 실제 결과들이 실질적으로 달라질 수 있는 위험과 불확실성의 영향을 받습니다. 독자는 본 자료의 간행일 현재를 기준으로 자사의 의견을 반영한 이들 예측 및 미래 전망 진술을 지나치게 의존해서는 안됩니다.

자사는 새로운 정보 또는 향후 사건에 따라 이들 예측 및 미래 전망 진술에 대한 여하한 변경 결과를 수정하거나 공식 발표할 의무가 없다는 것을 유념해 주십시오.