

GRAFHCORE

BOW POD₁₆

탐색 | 빌드 | 확장

Bow Pod 시스템은 대규모 머신 인텔리전스 배포를 위한 높은 성능과 효율성을 제공합니다. 이 시스템은 오늘날의 복잡한 대규모 모델을 가속화하고 혁신가들이 미래의 솔루션을 탐색 및 발명할 수 있는 플랫폼을 제공할 수 있도록 설계되었습니다.

Bow Pod₁₆ 시스템은 탐색에 최적화되었습니다. 이 시스템은 프로토타이핑 단계를 단축하여 신속하게 프로덕션으로 진행하기 위한 강력한 전력, 성능과 유연성을 제공합니다. Bow Pod₁₆은 언어 및 비전, GNN 및 LSTM 탐색 또는 완전히 새로운 개발 등 모든 머신 러닝 분야에서 IPU에서 더욱 향상된 혁신적인 AI 솔루션을 빌드하기 위한 최적의 시작점으로 기능합니다.

최신 IPU

Bow Pod₁₆ 시스템은 Bow-2000 머신 4대를 포함합니다. 각 머신에는 그래픽코어의 선도적인 Bow IPU 프로세서 4대가 탑재되어 있습니다. 이 혁신적인 IPU는 세계 최초로 웨이퍼 온 웨이퍼(WoW) 기술을 사용하여 제조된 프로세서로, 이미 입증된 IPU 기술의 이점을 새로운 차원으로 끌어올렸습니다.

시스템 사양

프로세서	Bow IPU 16개
1U 블레이드 유닛	Bow-2000 머신 4대
분리형 코어	23,552개
스레드	141,312개
성능	5.6페타플롭스 FP16.16 1.4페타플롭스 FP32
메모리	14.4GB의 인프로세서 메모리™ 최대 1TB의 스트리밍 메모리™
소프트웨어	Poplar® SDK

성능 및 효율

Bow Pod₁₆은 혁신적인 실리콘 기술 사용, 효율성과 스케일아웃 중심의 연산 및 메모리 아키텍처, 그리고 소프트웨어 및 애플리케이션에 최적화된 솔루션 배포 접근법에 힘입어 최대 5.6페타플롭스의 AI 연산과 업계 최고의 효율을 제공합니다.

원활한 배포 및 짧은 시장 출시 기간

하드웨어와 소프트웨어를 포함한 시스템 전체가 함께 설계되었습니다. Bow Pod₆₄는 모든 표준 프레임워크 및 프로토콜을 지원하므로, 기존의 데이터 센터 환경뿐 아니라 프라이빗 클라우드 및 퍼블릭 클라우드와의 간편한 통합이 가능합니다.

여러 선도적인 서버 플랫폼과 AI용으로 설계된 고성능 스토리지 장치에 대한 테스트 및 검증은 완료했으며, 시스템 수집자의 구성 및 배포 시간도 단축했습니다.

혁신가들은 익숙한 도구와 프레임워크를 활용하는 동시에 최첨단 성능과 효율성의 혜택을 누리며 대규모로 AI 워크로드를 배포할 수 있습니다.

호스트 링크	100 GE RoCEv2
시스템 중량	66 kg + 호스트 서버 및 스위치
시스템 디멘션	4U + 호스트 서버 및 스위치
호스트 서버	그래프코어® 파트너가 엄선한 승인된 시스템
스토리지	그래프코어® 파트너가 엄선한 승인된 시스템
냉각 방식	공냉식

BOW POD₁₆

맞춤형 연산을 위한 분할

머신 인텔리전스 워크로드의 연산 수요는 제각기 다릅니다. 프로덕션 배포의 경우 AI와 호스트 연산의 비율을 최적화함으로써 성능을 극대화하고 총 소유 비용을 절감할 수 있습니다. Bow Pod 시스템을 사용하면 서버 및 스위치 수를 필요한 Bow-2000 머신 수에 맞춰 유연하게 매핑할 수 있으며, 이를 통해 프로덕션 AI 워크로드에 더욱 최적화된 배포가 가능합니다. Bow Pod₁₆은 다중 서버 구성을 지원합니다.

규모 확대에 최적화된 통신 아키텍처

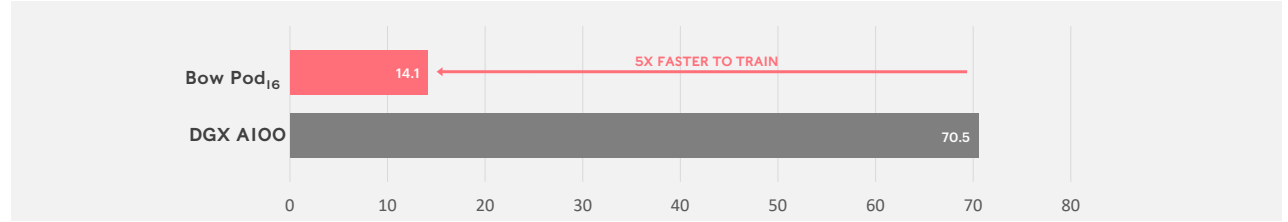
효율적인 데이터 액세스와 전송으로 더욱 탁월한 AI 성능을 실현할 수 있습니다. IPU Fabric은 시스템 전반에 걸친 데이터 전송을 위한 혁신적인 통신 아키텍처로, 개별 Bow IPU, 여러 대의 Bow-2000, Bow Pod 간 및 데이터 센터 전체로 고속 상호 연결을 확장합니다. IPU Fabric은 고성능 저지연 통신을 제공하여 AI 애플리케이션의 효율성을 극대화하며, 표준 데이터센터 통신 기술과 함께 사용 가능하게 구축되었습니다.

AI 개발자에게 최적화된 플랫폼

방대한 규모의 PopLibs™ 라이브러리와 함께 커뮤니티 중심의 협업과 혁신을 위한 TensorFlow, PyTorch, PaddlePaddle을 비롯하여 널리 사용되는 ML 프레임워크가 지원되며 오픈소스로 제공됩니다. 그래프코어 Poplar SDK는 성능을 극대화하려는 개발자를 위해 C++를 사용한 직접 IPU 프로그래밍을 지원합니다.

시작할 준비가 되셨나요?

info@graphcore.ai로 그래프코어 전문가에게 문의하여 AI 인프라 요구 사항 및 적합한 솔루션을 알아보세요.



EfficientNet-B4 Training Time To Train Performance
IPU-POD Platforms | Preliminary Results (Pre-SDK2.5) | G16-EfficientNet-B4 Training
DGX A100 (A100-SXM4-80GB) | TensorFlow | Mixed Precision | <https://developer.nvidia.com/deep-learning-performance-training-inference>

대규모 배포에 최적화된 설계

혁신가들은 Poplar SDK 도구, 프레임워크 이미지와 기본 제공 Docker 컨테이너를 활용하여 즉시 시작할 수 있습니다. Slurm, Kubernetes, Open-Stack 등 컨테이너 오케스트레이션, 플랫폼 시각화 및 프로비저닝에 가장 많이 사용되는 여러 프레임워크도 지원됩니다.

소프트웨어 중심

IPU에 최적화되고 완전히 통합된 Poplar 소프트웨어는 IPU의 고유한 특성을 활용하여 최상의 성능과 유연성을 제공하는 AI 애플리케이션을 구축합니다. Poplar를 사용하면 복잡한 개발 없이 모델을 IPU 수천 개 규모로 간단히 확장할 수 있어 혁신가들이 애플리케이션의 정확도와 성능을 향상하는 데 리소스를 집중 투자할 수 있습니다.

AI 전문가의 지원 활용

풍부한 경험을 갖춘 그래프코어 AI 전문가와 탁월한 파트너 네트워크가 설치, 프로덕션 및 애플리케이션 개발을 전 세계적으로 지원합니다.