rebellions_

ATOM™-Max: 대규모 추론을 위한 성능

May 02, 2025

AI 워크로드가 점점 복잡해지고 규모가 커짐에 따라, 기존의 GPU 기반 시스템은 효율성과 지속 가능성 모두에서 점점 더 큰 한계에 직면하고 있습니다. 과도한 전력 소모와 인프라 요구로 인해, 대규모 데이터센터 수준의 장기 운영에서 성능 및 비용 측면의 병목이 발생합니다.

ATOM™-Max는 이러한 한계를 해소하기 위해 대규모 추론에 특화된 아키텍처로 설계되었으며, 범용 가속기 대비 더 높은 효율성과 확장성을 제공합니다. 이를 통해 가장 까다로운 엔터프라이즈 및 데이터센터급 AI 워크로드에서도 높은 자원 활용률과 처리량을 보장하며, 전력 효율 최적화와 탄소 감축 요구에도 부합하는 지속 가능한 인프라를 구현합니다.

확장가능한 고효율 추론

ATOM™-Max는 대규모 AI 추론에 특화된 구조로, 탁월한 연산 성능과 고대역폭 메모리 접근성을 결합합니다.

- **128 TFLOPS (FP16), 최대 1024 TOPS (INT4)** 성능으로 LLM과 엔터프라이즈급 AI 워크로드를 안정적으로 처리합니다.
- **PCIe Gen5 ×16 기반의 카드 간 직접 통신(card-to-card communication)** 으로 낮은 지연과 빠른 데이터 교환을 구현하며, 노드 간 수평 확장을 손쉽게 지원합니다.

대규모 추론 환경을 위해 설계된 ATOM™-Max는 높은 활용률, 예측 가능한 지연 시간, 간편한 배포를 보장하며, 현대 데이터센터 인프라와 완벽히 호환됩니다.

총소유비용(TCO) 절감의 핵심

대규모 추론 환경에서는 **처리량, 전력 효율(Performance-per-Watt), 배포 유연성** 간의 균형이 필수적입니다. 기존 GPU 시스템은 높은 CapEx·OpEx를 요구해 지속적 확장이 어렵지만, **ATOM™-Max (RBLN-CA25)** 는 실제 환경에서 **Tokens-per-Second per Watt (TPS/W)** 기준으로 L40S를 상회하는 성능 효율을 제공합니다.

데이터·메모리 관리의 긴밀한 통합으로 **하드웨어 활용률을 극대화**하며, 유휴 오버헤드와 자원 낭비를 최소화합니다.이러한 **아키텍처 효율성은 TCO 절감 효과로 이어지며,** 배포 규모가 커질수록 그 이점은 더욱 확대됩니다.

ATOM™-Max: 스케일링 가능한 전력 효율

기존 인프라는 전력 효율, 메모리 병목, 배포 비용 등 다양한 제약에 직면해 있습니다. ATOM™-Max는 이러한 문제를 구조적으로 해결합니다.

더 많은 스케일, 더 큰 절감 — 고집적 연산 효율

350W의 전력 예산 내에서 128 TFLOPS (FP16), 1024 TOPS (INT4)의 성능을 제공하며, 랙당 처리량을 극대화합니다. 동일한 워크로드를 수행하기 위해 필요한 서버 수를 줄여 인프라 비용을 절감하고, 시스템 효율을 향상시킵니다.워크로드가 커질수록 규모의 경제(economies of scale) 효과가 커져, ATOM™-Max는 대규모 AI 인프라에 최적의 선택이 됩니다.

하드웨어-소프트웨어 공동 최적화

ATOM™-Max는 하드웨어와 소프트웨어의 정교한 공조(Co-optimization) 를 통해 고급 동기화 및 공유 메모리(SHM) 구조로 메모리 효율과 활용률을 극대화합니다. 컴파일러는 AI 모델을 ATOM™-Max 아키텍처에 최적화된 실행 명령으로 자동 변환하여, 지연 시간과 연산 오버헤드를 최소화한 상태에서 최고 성능을 실현합니다.

[ATOM™-Max 가속기를 탑재한 고집적 AI 서버]

미래형 AI 서버 아키텍처

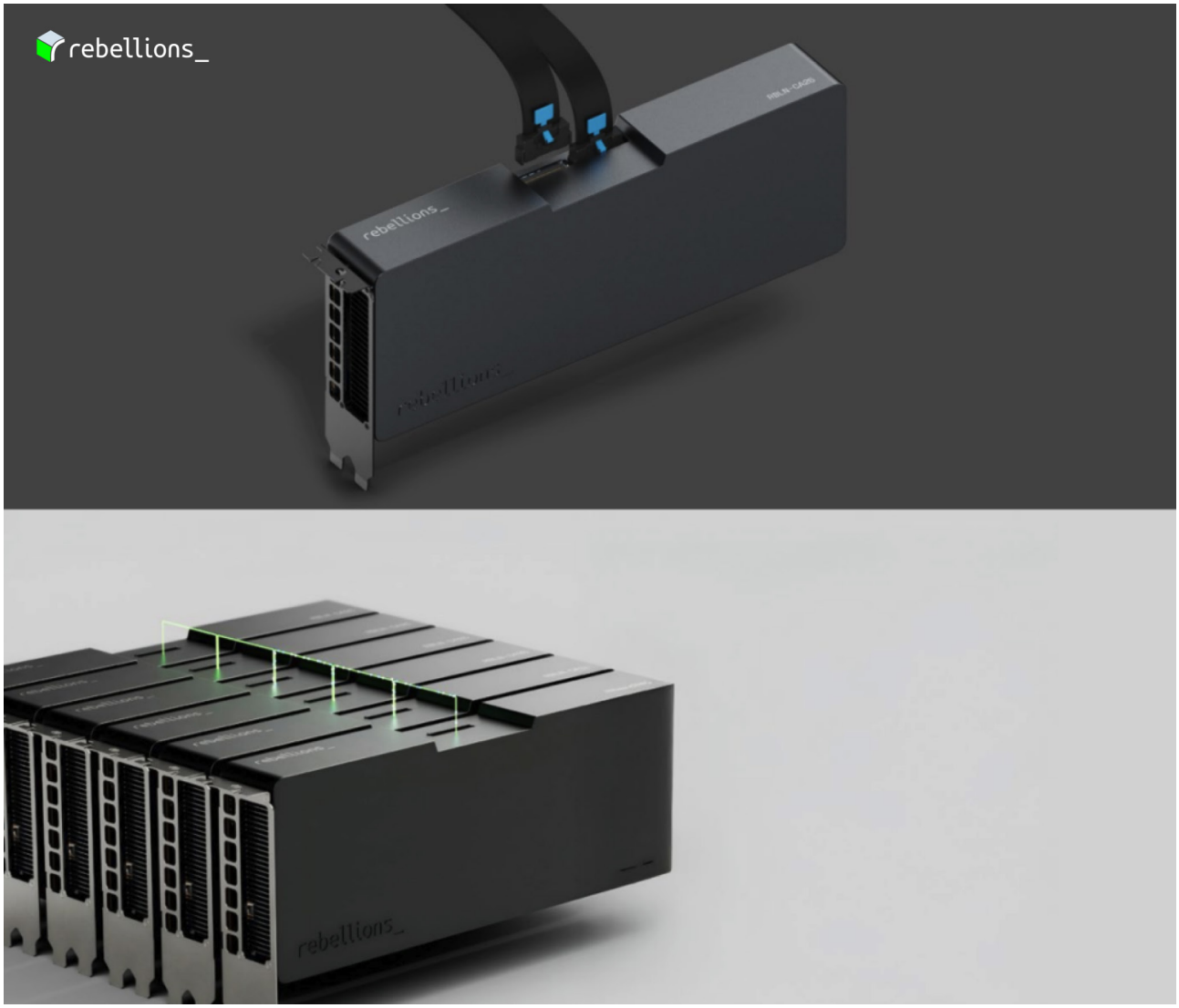
고대역폭 메모리

64GB 고대역폭 메모리(GDDR6)와 **1024 GB/s의 메모리 대역폭**을 제공하여, LLM 및 대규모 AI 추론

에 필요한 **데이터 처리 속도와 용량**을 동시에 확보합니다. 최적화된 데이터 플로우 설계로 병목을 제거하고, 처리량을 극대화합니다.

매끄러운 확장성

ATOM™-Max는 단일 카드 성능을 넘어, **최대 8장까지의 다중 카드 확장(Multi-card Scaling)** 을 지원합니다. 전용 **고속 인터카드 커넥터(High-speed intercard connector)** 을 통해 카드 간 데이터 전송 지연을 최소화하고, 대규모 AI 연산 환경에서도 뛰어난 확장 성능을 제공합니다.



[멀티카드 확장을 위한 고속 인터카드 커넥터]

결론

AI 추론 수요가 학습(Training)을 넘어서는 시대에, 기존 GPU 중심의 인프라는 점점 더 비효율적이 되고

있습니다. 높은 전력 비용과 인프라 부담, 운영 제약이 그 성장을 가로막고 있습니다. ATOM™-Max는 이러한 구조적 한계를 해결하는 전용 AI 추론 가속기로, 다음과 같은 이점을 제공합니다.

- 더 높은 연산 밀도를 통한 효율적 대규모 추론
- 낮은 전력 소모로 인한 TCO 및 환경 부담 절감
- 차세대 AI 모델을 위한 매끄러운 확장성

전력 효율적이고 고집적 구조의 ATOM™-Max는 지속 가능하면서도 비용 효율적이며, 미래형 AI 인프라 확장의 핵심이 될 것입니다.