rebellions_

ATOM™-Max Server

대규모 AI 추론을 위한 고성능 서버

ATOM™-Max 서버는 고효율 전력 설계를 기반으로, 단일 서버만으로도 대규모 AI 추론을 안정적으로 수행합니다. Vision AI, LLM, 멀티모달 AI, Physical AI 등 수백 종의 AI 모델과 vLLM, Triton, Kubernetes 같은 핵심 AI 서비스 운영 도구를 지원합니다. GPU 친화적인 개발 환경을 제공해 손쉽게 활용할 수 있습니다.

Key Features

Performance at Any Scale

사용자 요청이 급증해도 성능 저하 없이 대규모 AI 서비스를 지원합니다. 단일 서버로도 수천 개 토큰과 프레임을 실시간 처리합니다.

Sustainable AI Infrastructure

고효율 전력 설계로 AI 인프라의 총소유비용(TCO)를 절감하며, 지속가능한 AI 비즈니스 환경을 구축할 수 있습니다.

Variety of Models Applications

LLM, Vision AI, 멀티모달 AI, Physical AI 등 수백 종의 AI 모델을 즉시 활용해 맞춤형 AI 서비스를 구현할 수 있습니다.

Develop As You Always Have

익숙한 개발 환경과 워크플로우(Pytorch, TensorFlow 등)를 그대로 활용할 수 있어 튜토리얼을 통해 쉽게 개발을 이어갈 수 있습니다.

Full-Stack Software Support

vLLM, Triton Inference Server, K8s 등 오픈소스 생태계와 호환되며, 효율적인 서빙, 유연한 자원 운영 및 모니터링을 위한 다양한 AI 서비스 운영 도구로 end-to-end 서비스 구축이 가능합니다.

Spec

NPU	ATOM™-Max NPU Card *8
NPU Memory	512GB GDDR6, 8TB/s
Performance	1,024 TFLOPS (FP16) 4,096 TOPS (INT8)
Form Factor	4U
CPU	5th Gen. AMD EPYC Processor *2
Memory	1.5~2.3TB
Storage	1.92TB SATA * 2
Network	10G 2port * 2 400G 1port (Optional)
Max Power Consumptoin	Typical 3.4kW (Max ~4.3kW)
PCIe Slots	13x PCIe gen5 x16 [FHFL slots] [8x ATOM™-Max + 1x 400G 1-port NIC + 1x 10G 2-port NIC]
Compatible Software	- **OS**: Ubuntu, RHEL, AlmaLinux, RockyLinux - **Frameworks & Tools**: Hugging Face, PyTorch, TensorFlow, Triton - **Inference Serving**: vLLM, Triton Inference Server, TorchServe - **Orchestration**: Docker, OpenStack, Kubernetes, Ray

RBLN SDK

GPU의 익숙한 사용성을 제공하면서도, 차세대 AI 워크로드를 위해 설계된 Full-Stack Inference Platform을 제공합니다. PyTorch 개발부터 LLM 서빙과 배포까지, 모든 단계가 엔터프라이즈 환경에 맞춰 설계되었습니다.

Driver SDK
NPU 구동을 위한
기본 시스템 SW 및 도구 모음

- **Firmware**
- **Kernel Driver**
- **User Mode Driver**
- **System Management Tool**

NPU SDK
모델 및 서비스 개발을 위한 SW 도구 모음

- **Compiler, Runtime, Profiler**
- **Hugging Face 지원**
- **주요 추론 서버 지원**
  (vLLM, TorchServe, Triton Inference Server 등)

Model Zoo
리벨리온 NPU에서 곧바로 쓸 수 있는
300+ PyTorch와 TensorFlow 모델 제공

- **Natural Language Processing**
- **Generative AI**
- **Speech Processing**
- **Computer Vision**
- **Physical AI**