rebellions_

ATOM™-Max Server

High Performing Server
for Large-Scale AI Inference

ATOM™-Max Serverは、エネルギー効率の高い大規模なAI推論演算に最適なサーバです。このサーバには最大8枚のATOM™-Maxのカードを搭載でき、画像AI、LLM、マルチモーダルAI、そしてフィジカルAIなど数百のモデルを動かすことができます。vLLM、Triton、そしてKubernetesにも対応しており、現在お使いのGPUやその他ツールからシームレスに移行することが可能です。

Key Features

あらゆる規模のシステムにおいて高いパーフォーマンスを実現します。

高負荷環境においても、ATOM™-Max は高いスループットで安定したプロセスを実現します。1つのシステムで毎秒数千のトークンや画像を処理することが可能です。

サステイナブルなAIインフラを実現する不可欠な演算プラットフォーム

ATOM™-Maxは、比類ないエネルギー効率をもって最大のAI推論のパーフォーマンスを実現し、TCOを最小化しサステイナブルなAIインフラを実現させます。

多くのAIモデルに対応

vLLM、画像、マルチモーダル、フィジカルAIなど数百に及ぶAIモデルに対応しておりお客様の必要なサービスを構築することができます。

環境を変更せずに開発することが可能

お使いの開発環境をそのままお使いできます。PyTorch、TensorFlowなどお使いのワークフローをそのままお使い頂き、ステップバイステップのチュートリアルもご提供します。

フルスタックのソフトウェアのサポート

vLLM、Triton、Kubernetes、Prometheusなどに対応しており、高効率で柔軟なリソースマネジメントを実現し、シームレスなAIサービスを実現します。

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 Consumption	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

Rebellionsは、ソフトウェアの開発負荷を上げることなくカスタムAI半導体のパーフォーマンスと効率を実現します。PyTorch、vLLM、Tritonなどに対応しており、大規模なAI推論をGPUと同じワークフローで構築することができます。

Driver SDK
NPUの稼働の為のコアシステムと基本ツール

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

NPU SDK
モデル、サービスの開発のためのツール

- **Compiler, Runtime, Profiler**
- **Hugging Face integration**
- **Major Inference Servers Supported**
  (vLLM, TorchServe, Triton Inference Server etc.)

Model Zoo
RebellionsのNPUに対応した300以上の直ぐにご利用頂けるPyTorchやTensorFlowのモデル

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

Discover More
rebellions.ai