ATOM™-Max Pod

Rack-Scale AI Infrastructure with RDMA-Based High-Speed Networking

Built for large-scale AI inference, ATOM™-Max Pod is a rack-scale infrastructure designed for distributed workloads. It combines Rebellions' AI accelerators with RDMA-based high-speed networking and a familiar software stack—all delivered as a turnkey solution. Starting from an 8-server Mini Pod, the system scales flexibly to meet enterprise-level AI demands.

Key Features

Limitless Scale Out Architecture

Start with an 8-server Mini Pod and expand to dozens of servers, all connected into a single cluster through RSD. Scale resources as workloads grow and achieve linear performance gains.

Ultra-Low Latency RDMA Fabric

Each server in the Pod is linked through a 400 GB/s RDMA network. Purpose-built for distributed processing, it delivers the throughput required for the most demanding models without latency bottlenecks.

All-in-One Turnkey Infrastructure

From AI accelerators to RDMA switches and node-to-node networking, the Pod delivers a fully integrated system. With a field-tested hardware and software stack, it is ready to move into production immediately, removing complexity and maximizing operational efficiency.

Ready-to-Deploy Rebellions Enterprise AI Solution

The ATOM™-Max Pod can be equipped with Rebellions' Enterprise AI Solution, optimized for enterprise environments. It supports the full lifecycle of AI serving in a cost-efficient way, offering a production-ready solution you can adopt today.

All information is provided "as is" without any warranties or guarantees by Rebellions, Inc.("Rebellions"). Rebellion assumes no liability for use of or reliance upon this information.

Spec

Chassis	42U
Server	8 servers
AI Accelerator	64x ATOM™-Max Cards
Management Network	1G UTP Switch
Storage Network	10G Optic Switch
RDMA Network	800G Data Switch
Power	4x redundant PDUs (2N redundancy)
Thermal	Air-Cooled

RBLN SDK

We deliver a full-stack inference platform that combines the familiar usability of GPUs with architecture built for next-generation AI workloads. From PyTorch development to LLM serving and deployment, every stage is designed for enterprise environments.

Driver SDK

Core system software
and tools for running NPUs

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

NPU SDK

Development toolkit for models
and services

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

Model Zoo

300+ ready-to-run PyTorch
and TensorFlow models
on Rebellions NPUs

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

Cloud SDK

Software suite for managing
NPU resources in the cloud

- K8s Device Plugin
- Metric-Exporter
- Node Feature Discovery
- Device Installer
- VFIO Manager
- K8s Operator

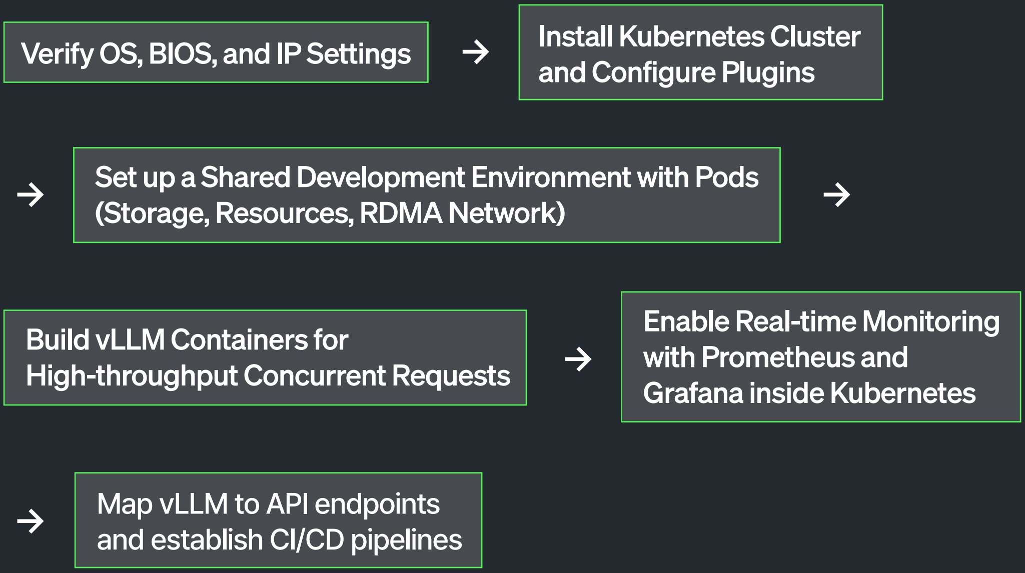
All information is provided "as is" without any warranties or guarantees by Rebellions, Inc.("Rebellions"). Rebellion assumes no liability for use of or reliance upon this information.

Enterprise AI Solution

Full-Lifecycle Solution for Enterprise AI Serving

On the ATOM™-Max Pod, you can run Rebellions AI Serving Solution, supporting the entire lifecycle of enterprise AI services. It provides development toolkits for node-level distributed serving, automated infrastructure management tools, and independent development environments for multiple developers.

Day 1 Build and Deploy



Verify OS, BIOS, and IP Settings → Install Kubernetes Cluster and Configure Plugins

→ Set up a Shared Development Environment with Pods (Storage, Resources, RDMA Network) →

Build vLLM Containers for High-throughput Concurrent Requests → Enable Real-time Monitoring with Prometheus and Grafana inside Kubernetes

→ Map vLLM to API endpoints and establish CI/CD pipelines

All information is provided "as is" without any warranties or guarantees by Rebellions, Inc.("Rebellions"). Rebellion assumes no liability for use of or reliance upon this information.