

# AI 전환 시대, 인프라 효율의 새로운 기준

AI 인프라의 새로운 선택지 - MangoBoost DPU 솔루션

# WeZON DPU Solution

## 현장에서 반복되는 3가지 인프라 한계

성능은  
기대 이하

비싸게 구축했는데  
생각보다 느립니다

GPU, 고성능 서버를 도입했지만  
실제 운영에서는 기대만큼 성능이 나오지 않습니다.  
고가의 장비가 충분히 활용되지 못하고 있습니다.

AI 도입  
비용 부담

AI를 시작하려니  
인프라 비용이 너무 큼니다

LLM과 AI 서비스를 구축하려면  
서버, 네트워크, 스토리지까지 함께 증설해야 합니다.  
도입 규모가 커질수록 부담도 빠르게 증가합니다.

레거시  
인프라의 한계

기존 자산을 버리기엔 아깝고  
그대로 쓰기엔 부족합니다

현재 인프라를 유지하자니 AI 활용에 한계가 있고,  
새롭게 교체하자니 투자 부담이 큼니다.  
현실적인 전환 방안이 필요합니다.

성능, 비용, 인프라 전환 문제는 결국 같은 **구조적 병목**에서 시작됩니다.

# AI 인프라 병목의 근원 - 아키텍처 구조적 한계

"AI 애플리케이션의 성능 병목이 연산(Compute)이 아닌  
메모리·통신(Memory & Communication) 영역으로 이동하고 있다는 것이 업계의 공통된 분석입니다."

- UC Berkeley ICSI Amir Gholami 교수 등 다수 연구진 (AI and Memory Wall, 2024) -

## 문제는 GPU가 아니라 데이터 이동입니다

### CPU가 모든 데이터를 처리

네트워크 트래픽, 스토리지 I/O, 보안 처리까지 CPU가 담당  
→ AI 연산보다 데이터 이동에 더 많은 자원 소모

### GPU 활용률 저하

GPU 성능은 향상되었지만 데이터 공급 속도가 이를 따라가지 못함  
→ 고가의 GPU 자원이 유향 상태로 남음

### 인프라 비용 증가

병목 해소를 위해 서버를 추가 도입하지만 근본 원인은 해결되지 않음  
→ 운영 비용과 전력 비용만 증가

## 컴퓨터 성능

1976년 대비 6000배 향상

6000

## 네트워킹 성능

통신 병목으로 인해 30배 향상에 그침

30

\* 컴퓨팅·네트워킹 성능 성장을 비교는 업계에 통용되는 일반 통계를 인용함 (산업 자료 기준)

# DPU 기반 인프라 가속

## DPU란?

네트워크, 스토리지, 보안 처리와 같은

인프라 작업을 전담하는 데이터 처리 프로세서(DPU)

→ CPU와 GPU는 본연의 연산 업무에 집중하여 인프라 효율을 극대화



MangoBoost BoostX

데이터 이동 병목을 제거하는 차세대 인프라 가속 기술

## DPU 도입 전

- 네트워크·스토리지·보안 처리를 CPU가 담당
- 데이터 처리 병목으로 GPU 활용률 저하
- 인프라 증설에도 성능 개선 효과 제한
- 운영 비용 및 전력 사용량 지속 증가



## DPU 도입 후

- DPU가 데이터 처리 업무를 전담하여 CPU 부하 감소
- GPU는 AI 연산에 집중하여 자원 활용률 향상
- 기존 인프라 활용 기반의 성능 개선
- 성능 향상과 운영 비용 절감을 동시에 실현

CPU 중심 구조에서 DPU 중심 구조로 전환하여 인프라 효율을 극대화

# 인프라 병목을 제거하는 5가지 가속 기술

| LLMBoost                                                                                                                                                                                                                                                   | GPUBoost                                                                                                                                                 | Storage Boost                                                                                                                                        | CPU Offload                                                                                                                                           | Open Standard                                                                                                                                                                    |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>AI 추론 및 학습 최적화</p> <hr/> <p>추론</p> <ul style="list-style-type: none"> <li>KV-Cache 및 서빙 성능 최적화</li> <li>커널·알고리즘 자동 선택</li> </ul> <p>훈련</p> <ul style="list-style-type: none"> <li>분산 병렬화 자동 구성</li> <li>통신 라이브러리 최적화</li> <li>코드 수정 없이 간편 배포</li> </ul> | <p>GPU 간 초고속 데이터 전송</p> <hr/> <ul style="list-style-type: none"> <li>GPU 간 직접 데이터 전송</li> <li>CPU 개입 없는 GPU 메모리 접근</li> <li>RoCEv2 기반 초저지연 통신</li> </ul> | <p>스토리지 병목 제거</p> <hr/> <ul style="list-style-type: none"> <li>NVMe-oF 기반 스토리지 가속</li> <li>DPU 직접 처리로 I/O 병목 제거</li> <li>대규모 데이터 처리 성능 향상</li> </ul> | <p>CPU 자원 최적화</p> <hr/> <ul style="list-style-type: none"> <li>네트워크·스토리지·보안 처리 오프로드</li> <li>CPU 부담 감소 및 자원 효율 향상</li> <li>AI 연산 자원 활용 극대화</li> </ul> | <p>오픈 아키텍처 기반 확장성</p> <hr/> <ul style="list-style-type: none"> <li>표준 Ethernet 기반 인프라 구성</li> <li>특정 벤더 종속성 최소화</li> <li>다양한 GPU 플랫폼 선택 가능</li> <li>인프라 구축 및 운영 비용 절감</li> </ul> |
| 추론 성능 최대 43.5배 향상                                                                                                                                                                                                                                          | 최대 400Gb/s 데이터 처리                                                                                                                                        | MLPerf Storage 세계 1위                                                                                                                                 | CPU 오버헤드 Near-zero                                                                                                                                    | NVIDIA 대비 비용 62% 절감<br>(전체 솔루션 스택 기준)                                                                                                                                            |

AI 인프라의 연산·통신·스토리지 병목을 통합적으로 해결합니다.

## DPU 적용으로 달라지는 성능과 비용

82%

### CapEx 절감

8PB 규모 스토리지 환경 기준  
컴퓨터·네트워크 인프라 규모 축소

3.9x

### AI 추론 효율 향상

MLPerf 공식 결과 기준  
동일 자원 대비 처리 효율 개선

90%

### 스토리지 인프라 최적화

동일 성능 달성 기준  
스토리지 노드 수 최대 90% 감소

\* MLPerf Storage v1.0 · MLPerf Inference 5.0 / 5.1 공식 결과 기준 (2025)

### 기술 검증

- MLPerf Storage 세계 1위
- MLPerf Inference 5.0/5.1 공식 제출

### 산업 검증

- AMD · Dell · Supermicro 공동 검증
- AMD MI300X 멀티노드 최초 제출

### 기술 역량

- MangoBoost 관련 특허 40건 이상 보유

# 기존 인프라를 활용한 단계적 성능 개선



## 기존 서버 활용 극대화

기존 CPU 서버의 성능 한계로  
신규 인프라 투자 필요성 증가

DPU 적용 후

DPU 추가를 통해 CPU 부하를 줄이고  
기존 서버의 활용 효율을 향상

vLLM 대비 최대 43.5배



## 스토리지 가속·대체

기존 스토리지 구조의 I/O 병목이  
전체 시스템 성능을 제한

DPU 적용 후

DPU 기반 NVMe-oF로  
고성능 데이터 경로 구현

스토리지 노드 최대 90% 감소



## 네트워크 인프라 고도화

네트워크 처리 부하로 인해  
CPU 자원 활용 효율 저하

DPU 적용 후

CPU는 연산 업무에 집중하여  
네트워크 처리 효율 향상

CPU 오버헤드 Near-zero

기존 인프라를 유지하면서도 DPU 기반 가속기술을 통해 **성능과 효율을 동시에 개선**할 수 있습니다.

# AI 인프라 최적화가 필요한 산업

대규모 데이터 처리, GPU 연산, AI 서비스 운영 환경이라면 DPU 기반 인프라 최적화 효과를 기대할 수 있습니다.



## 제조·자동차

- 대규모 CAE·CFD 시뮬레이션 환경
- GPU 클러스터 운영 비용 증가
- 설계·해석 워크로드 처리 시간 개선 필요

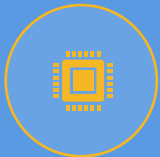
→ GPU 가속 + 기존 인프라 활용



## 금융·공공기관

- 보안과 안정성이 중요한 AI 서비스 환경
- 레거시 인프라 활용 필요
- AI 도입 시 비용 효율성 확보 필요

→ 보안 가속 + 기존 인프라 활용



## 연구기관·반도체

- HPC 및 대규모 연구 연산 환경
- 스토리지 병목 및 클러스터 확장 한계
- 추가 투자 없이 성능 개선 요구

→ 병렬 스토리지 가속 + 멀티노드 AI



## 통신사·IT 서비스

- AI 서비스 및 대규모 데이터 처리 환경
- 추론 인프라 운영 비용 증가
- 서비스 확장에 따른 TCO 절감 필요

→ 스토리지 가속 + AI 추론 최적화

AI 연산, 데이터 처리, HPC 워크로드를 운영하는 모든 환경이 적용 대상입니다.

# 왜 지금 DPU 기반 AI 인프라인가

01

## AI 인프라 수요 확대

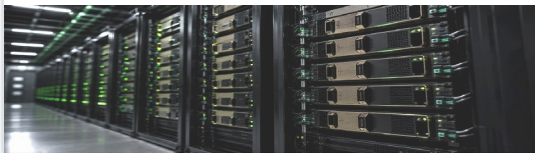
AI 도입이 본격화되면서  
고성능·고효율 인프라에 대한 수요가  
빠르게 증가하고 있습니다.  
인프라 최적화가 경쟁력의 핵심 요소가 되고  
있습니다.



02

## 오픈 아키텍처에 대한 관심 증가

고비용·벤더 종속 구조에 대한 부담이 커지면서  
표준 기반 인프라와 다양한 GPU 선택지가  
주목받고 있습니다.



03

## AI 투자 기준의 변화

PoC 중심의 검토 단계를 넘어  
실제 운영 환경에 대한 투자가 확대되고 있습니다.  
성능뿐 아니라 ROI와 운영 효율성이 중요한 평가  
기준이 되고 있습니다.



AI 인프라 시장은 성능 중심에서 **효율 중심으로 전환**되고 있으며,  
DPU 기반 아키텍처는 이러한 변화에 대응하는 현실적인 대안입니다.

# WeZON DPU Solution

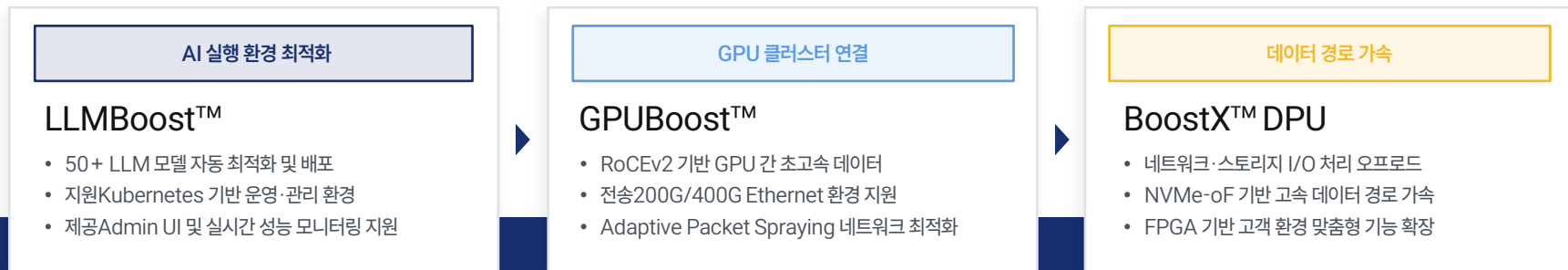
MangoBoost HW/SW 솔루션 — 성능과 비용 절감을 구현하는 핵심 기술 구조

---

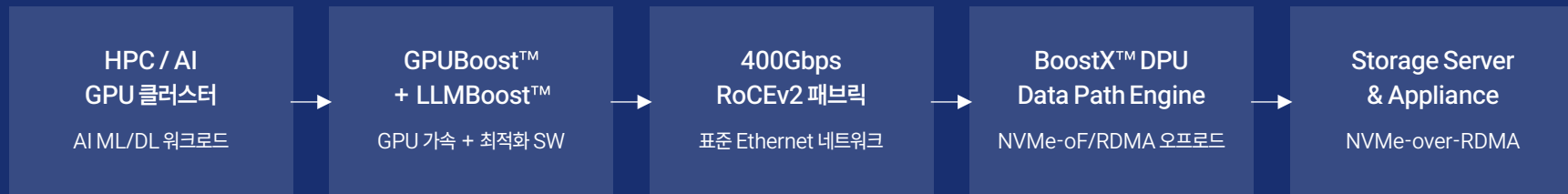
- A. 통합 솔루션 아키텍처
- B. LLMBoost 기술 상세
- C. GPUBoost NIC 기술 상세
- D. BoostX DPU 기술 상세
- E. 성능 검증 데이터
- F. BoostX DPU 제품 소개
- G. GPUBoost & Storage Server 소개

# MangoBoost 통합 솔루션 아키텍처

AI 연산부터 데이터 저장까지 하나의 데이터 경로로 연결하는 DPU 기반 통합 아키텍처

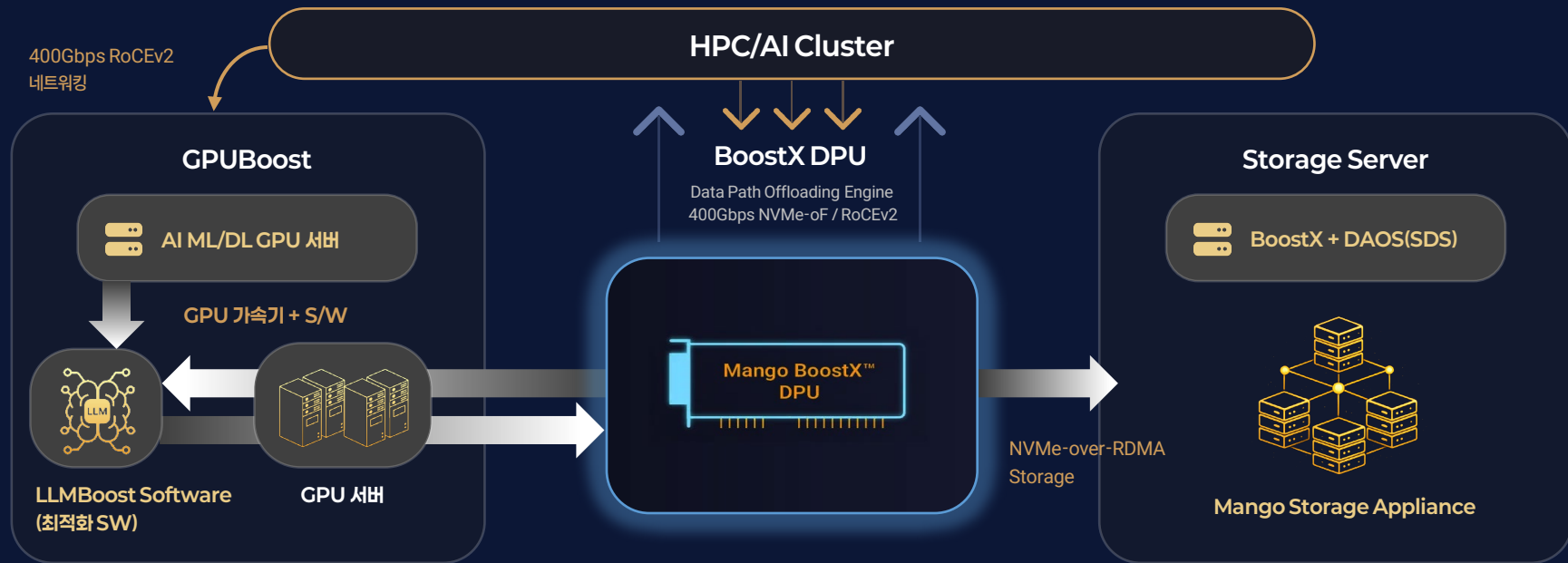


## 시스템 아키텍처 - Compute · Network · Storage 통합 데이터 경로



위존은 GPU 연산, 네트워크 전송, 스토리지 I/O를 하나의 데이터 경로로 최적화하여 전체 인프라 효율을 극대화합니다.

# MangoBoost 통합 솔루션 아키텍처



Compute-Storage 간 데이터 병목을 제거하는 DPU 중심 인프라 구조

# LLMBoost - AI 추론·훈련 가속 SW

AI 모델의 추론·학습 환경을 자동 최적화하여 성능을 높이고 운영 복잡도를 줄이는 엔터프라이즈 AI 플랫폼

## Auto-Tuning 메커니즘

### 추론 (Inference)

- 서빙 환경 자동 최적화 (Serving Optimization)
- 컨텍스트 기반 KV-Cache 최적화
- 다양한 데이터 타입 지원
- 커널 및 알고리즘 자동 선택

### 훈련 (Training)

- 멀티노드 학습 환경 최적화
- 분산 병렬화 자동 구성 (DP/PP/TP/FSDP/CP/EP)
- 통신 라이브러리 최적화 (vs RCCL)
- End-to-end Auto-Tuning 지원

## 엔터프라이즈 기능

- 컨테이너 기반 간편 배포 (Docker/K8s)
- Kubernetes 통합 관리
- Admin UI 제공
- 실시간 성능 및 자원 모니터링

## 지원 Foundation 모델

LLaMA 3

DeepSeek

Qwen

Gemma

Phi-3

Mistral

NVIDIA Nemotron

and 50+ more

AI 모델 최적화부터 운영·모니터링까지 하나의 플랫폼에서 통합 지원

# GPUBoost NIC - GPU 간 직접 고속 통신

CPU를 거치지 않는 GPU 간 직접 통신



## RoCEv2 최적화

- 표준 Ethernet 기반 RDMA 지원
- GPU 메모리 직접 접근(RDMA)
- CPU 개입 최소화
- 초저지연 데이터 전송 구현



## 네트워크 성능

- 최대 400Gbps 네트워크 대역폭 지원
- PCIe Gen5 기반 고속 인터페이스
- 양방향 90% 이상 라인레이트 성능
- AI 클러스터 확장 환경 최적화



## 혼잡 제어·최적화

- 프로그래머블 혼잡 제어 지원
- Adaptive Packet Spraying 기반 부하 분산
- Ultra Ethernet 아키텍처 지원
- PFC/ECN 기반 네트워크 안정성 강화



## 가상화·오프로드

- VirtIO 기반 가상화 성능 향상
- OVS Offloading 지원
- VXLAN·Geneve 터널링 지원
- NAT 및 Connection Tracking 지원

AI 클러스터의 통신 성능을 최적화하여 GPU 자원 활용률을 극대화합니다.

# BoostX DPU - FPGA 기반 스토리지·네트워킹 가속

## 하드웨어 스펙

|           |                                    |
|-----------|------------------------------------|
| 네트워크      | 2 × 200G QSFP56 / 1 × 400G QSFP-DD |
| 호스트 인터페이스 | PCIe Gen5 8-lane / Gen4 16-lane    |
| 온보드 CPU   | 2 × A72 on FPGA (Pro: +8 × Zen3+)  |
| 메모리       | LPDDR4 8GB (Pro: +DDR5 16GB)       |
| 전력        | 최대 75W (Pro: 최대 225W)              |
| 보안(Pro)   | MACsec · IPsec · TLS · AES-GCM     |
| 관리        | Integrated BMC · 1GbE OOB 포트       |
| 부팅        | Secure Boot (RSA) · PXE · UEFI     |

## FPGA 아키텍처의 강점

- 빠르게 진화하는 표준·프로토콜에 즉시 대응
- 고객 워크로드별 맞춤 기능 커스터마이징 가능
- 소프트웨어 업데이트로 기능 확장 (하드웨어 교체 불필요)
- 데이터 경로는 하드웨어 가속, 제어 경로는 CPU 활용

## 가속 기능 - 스토리지

|                                             |                                            |
|---------------------------------------------|--------------------------------------------|
| <b>NVMe/RoCEv2</b><br>Initiator / Target 모드 | <b>NVMe/TCP</b><br>Initiator / Target 모드   |
| <b>NVMe 가상화</b><br>멀티테넌트 스토리지 분리            | <b>NVMe 멀티패스</b><br>이중화·로드밸런싱              |
| <b>압축·중복제거</b><br>CapEx 절감 (Pro)            | <b>Erasure Coding</b><br>RAID 대체 고효율 (Pro) |

# 성능 검증 데이터

## MLPerf Inference 5.1

**43.5x** 처리 속도  
vLLM 대비 (Qwen 2.5 Vision)

**3.9x** 비용 효율  
vLLM 대비 추론 기준

**35%** ↑ 처리량  
차상위 결과 대비

**62%** 비용 절감  
NVIDIA 대비

World's First Heterogeneous GPU 결과  
(MI300X + MI325X + MI355X)

## MLPerf Training 5.0

**2x** 훈련 속도  
vLLM 대비 (Qwen 2.5 Vision)

**4x** 스케일링  
32 GPU 선형 스케일링

**~4.5x** 통신 가속  
vs 표준 RCCL (MoE 기준)

**멀티노드** AMD 최초  
AMD MI300X 멀티노드 최초 제출

LLMBoost + MI300X 공동 검증  
AMD 공식 Tech Blog 게재

## MLPerf Storage v1.0

**1위** 순위  
가속기 점수 1,400+ 달성

**82%** CapEx 절감  
경쟁사 대비 8PB 기준

**63%** OpEx 절감  
경쟁사 대비 8PB 기준

**74%** TCO 절감  
전체 총소유비용 기준

스토리지 노드 90% 감소  
컴퓨트·스위치 64% 감소

# BoostX™ DPU

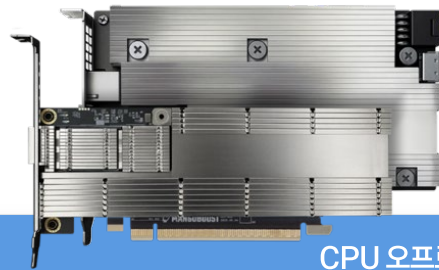
DPU는 네트워크·스토리지·보안 처리를 오프로드하는 인프라 가속 엔진입니다.  
CPU 부하를 제거하고 데이터 처리 성능을 근본적으로 향상시킵니다.

## 고성능 데이터 처리

- 2×200G QSFP56 또는 1×400G QSFP-DD 구성 지원
- 대규모 AI 워크로드 환경에서 고성능 데이터 처리

## 프로토콜 하드웨어 가속

- NVMe-oF, RoCEv2, TCP/IP 등 주요 데이터 경로를 하드웨어에서 직접 처리
- 소프트웨어 처리 지연을 줄이고 90% 이상의 라인레이트 성능을 제공



## CPU 오프로드

- 네트워크·스토리지·보안 처리 업무를 DPU가 전담
- CPU 자원이 AI 연산에 집중할 수 있도록 지원
- CPU 부하를 줄이고 시스템 효율을 향상 시킴

## FPGA 유연성

- 고객 워크로드에 맞춰 기능 확장과 최적화
- 표준 및 프로토콜 변화에 유연하게 대응
- 소프트웨어 업데이트 기반 기능 고도화 지원

# GPUBoost & Storage Server

## GPUBoost™

GPU 통신 병목 제거로 AI 처리 효율 극대화

### AI ML/DL GPU 서버 지원

다양한 GPU 구성에 대응하는 범용 가속 플랫폼

### LLMBoost 소프트웨어 통합

SW 자동 최적화와 HW 가속이 결합된 Full-stack 구조

### GPU Direct 기반 데이터 경로 최적화

GPU 메모리 ↔ 스토리지 직접 연결, CPU 개입 없음

### 표준 Ethernet 기반 멀티노드 확장

400Gbps RoCEv2로 멀티노드 클러스터 선형 확장 지원

## Storage Server

NVMe-over-RDMA 기반 고성능 스토리지 어플라이언스

### 즉시 구축 가능형 스토리지 어플라이언스

즉시 투입 가능한 완성형 스토리지 솔루션

### NVMe-over-RDMA 기반 초고속 데이터 경로

DPU가 직접 처리하는 초고속 스토리지 패브릭

### 최대 400Gbps 링크 대역폭 지원

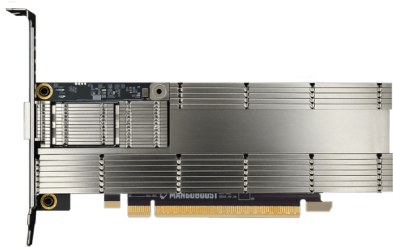
대규모 AI 클러스터 스토리지 I/O 병목 완전 해소

### AI 워크로드를 위한 초저지연 스토리지

GPU 연산과 스토리지 I/O 동시 처리로 대기시간 최소화

# GPUBoost & Storage Server

GPU Direct 기반 데이터 경로 최적화로 연산과 데이터 흐름을 직접 연결합니다.  
고속 스토리지와 결합하여 GPU 활용률과 처리 성능을 극대화합니다.



AI ML/DL GPU 서버 지원

최대 10배 성능 향상

LLMBoost 소프트웨어 통합

92% 추론 비용 절감

GPUBoost



어플라이언스 형태

NVMe-over-RDMA 기반

400Gbps NVMe-over-RDMA 대역폭

초저지연 스토리지 I/O

Storage Server

GPUBoost와 Storage Server의 **결합**으로 GPU 연산부터 데이터 저장까지 끊김 없이 연결되어,  
AI 워크로드 전체의 처리 효율이 극대화됩니다.



## Contact Us



Email  
[info@wezon.com](mailto:info@wezon.com)



Phone  
02-555-7033



Website  
[www.wezon.com](http://www.wezon.com)

HPC/AI 인프라 솔루션 전문기업 (주) 위존

Data Solution Division

A New Standard for GPU Acceleration

Where Cutting-Edge  
Technology  
Meets the Market.