DWcts

DIA

AI Nexus

AI Datacenter to AI Edge & On Device AI

About Us

DIA

대원씨티에스의 AI 비즈니스 브랜드인
DIA(Daewon CTS Innovative AI)의
심볼은 다이아몬드입니다.
한 줄기 빛을 여러 면을 통해 발산하는
다이아몬드처럼 DIA는 AI를 기업의 모든 비즈니스에
연결합니다.

AI Nexus

이런 뜻을 담은 슬로건이 'AI Nexus'입니다.
이 슬로건은 업무, 제품, 서비스 등 기업의 모든 것을
AI로 혁신할 수 있는 구심점을 제공하겠다는
대원씨티에스의 약속이자 다짐입니다.

DX to AX

AI Nexus는 기업 컴퓨팅 환경의 기능과 역할을 재정의합
니다.기업의 컴퓨팅 환경은 더 이상 업무 시스템이나
대외 서비스 지원을 위해 존재하지 않습니다.
데이터가 기업의 가장 큰 자산인 디지털 경제 시대,
AI 기술은 IT의 기능과 역할을 재정의 하고 있습니다.
이 혁신의 여정을 대원씨티에스가 함께 하겠습니다.

AI 전환(AX)이 디지털 전환(DX)의 완성인 시대

IT 투자 전략에서 AI는 이제 모든 조직의 우선순위 최상단에 위치한 과제입니다. 많은 조직이 AI 인프라 구축과 확장에 투자하고 있지만 나날이 커지고 정교해지는 모델 발전을 따라가기 위한 성능과 용량 요구를 충족하는 데 어려움을 겪고 있습니다. 이런 이유로 현재의 요구를 넘어 미래 수요까지 충족할 수 있는 차세대 AI 인프라 전략에 관심이 높아지고 있습니다. 실제로 업계 선도 기업들은 AI 전환(AX)이 디지털 전환(DX)의 완성이라고 보고 차세대 인프라 전략 수립에 나서고 있습니다. 현재 선도 기업들이 목표로 하는 차세대 AI 인프라의 핵심 요건은 다음과 같습니다.

01 **통합 설계**

AI 컴퓨팅 환경을 사일로가 아니라 현재부터 미래까지 일관성을 유지하도록 컴퓨팅, 스토리지, 네트워킹을 모두 고려하여 아키텍처 설계

02 **END TO END 인프라**

데이터센터부터 클라우드 그리고 엣지를 포괄하는 고성능 AI 인프라 구축과 운영

03 **분산 데이터 통합**

AI 워크로드 및 모델이 사내, 클라우드, 엣지 환경에 있는 모든 데이터에 성능 손실 없이 안전하게 접근할 수 있는 데이터 플랫폼과 스토리지 인프라 설계 및 구축

대원씨티에스의 차세대 AI 인프라 제안

대원씨티에스는 통합 설계 역량을 바탕으로 데이터센터에서 엣지에 이르는 포괄적인 AI 인프라 구축에 필요한 컴퓨팅, 네트워킹, 스토리지 및 데이터 플랫폼, AI 엣지를 위한 NPU 그리고 MLOps 플랫폼을 제공합니다. 이를 통해 워크로드 최적화, 통합 설계, 유연한 자원 사용, 데이터 통합, 경계를 넘어서는 데이터 통합에 대한 고객의 요구를 충족합니다.

워크로드 최적화

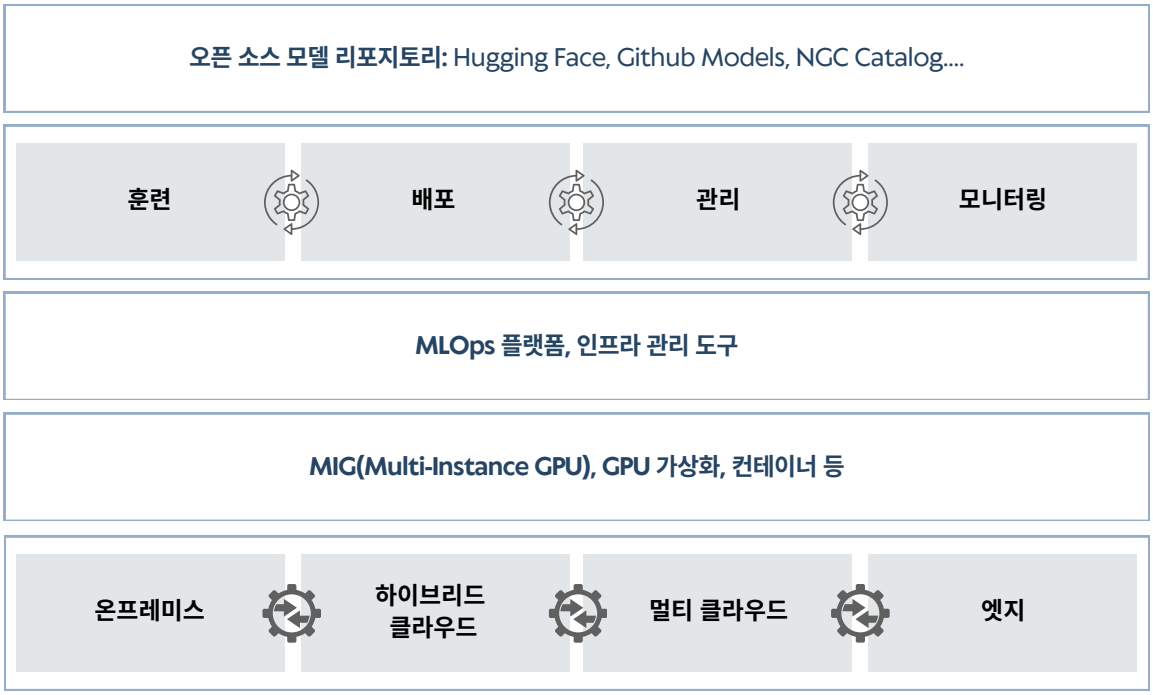
예측을 위한 AI/ML(Predictive AI), 거대 언어 모델(LLM) 기반의 생성형 AI(Generative AI), 엣지 장치와 사용자 장치를 위한 온디바이스 AI (On Device AI) 그리고 HPC까지 워크로드 유형에 따라 컴퓨팅, 네트워킹, 스토리지 요구가 다릅니다. 대원씨티에스는 폭넓은 솔루션 포트폴리오를 바탕으로 워크로드에 최적화된 인프라 설계와 이에 맞는 장비를 제안합니다.

통합 설계 및 AI 인프라 구축 역량

기업의 AI 인프라는 목적과 필요에 따라 규모가 다릅니다. 대원씨티에스는 수십 개의 CPU/GPU로 구성한 소규모 클러스터와 수천 개의 CPU/GPU로 구축한 AI 데이터센터 규모까지 컴퓨팅, 네트워킹, 스토리지를 하나의 관점에서 바라보고 인프라를 통합 설계합니다. 이를 위해 NVIDIA, KAYTUS, Supermicro, VAST Data 및 Arista 등 주요 AI 데이터센터 컴포넌트 기업의 레퍼런스 아키텍처를 분석해 고객의 요구에 맞게 최적화하여 적용합니다.

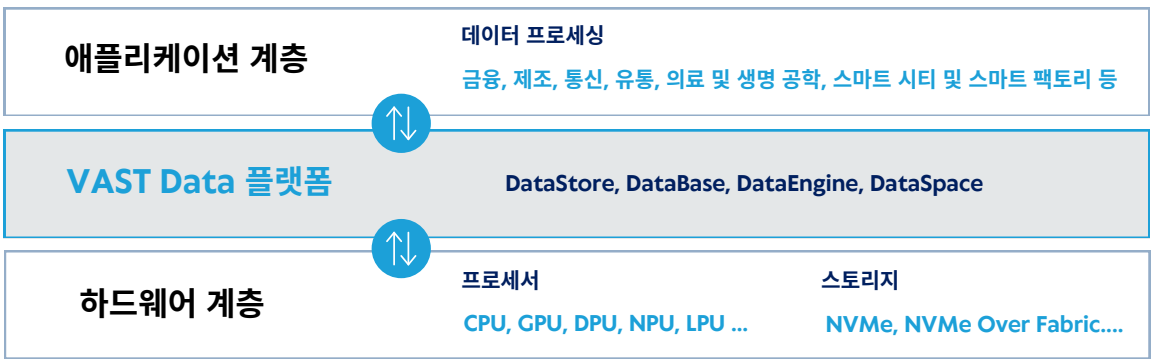
유연한 자원 사용

AI 워크로드를 위한 컴퓨팅과 스토리지 자원은 온프레미스부터 프라이빗/하이브리드 클라우드 그리고 멀티 클라우드 환경까지 여러 위치에 분산되어 있습니다. 대원씨티에스는 국내외 유명 MLOps 플랫폼 파트너들과 협력하여 최적화된 자원 풀을 개발자와 데이터 과학자가 온디맨드 방식으로 활용할 수 있는 방법을 안내합니다.



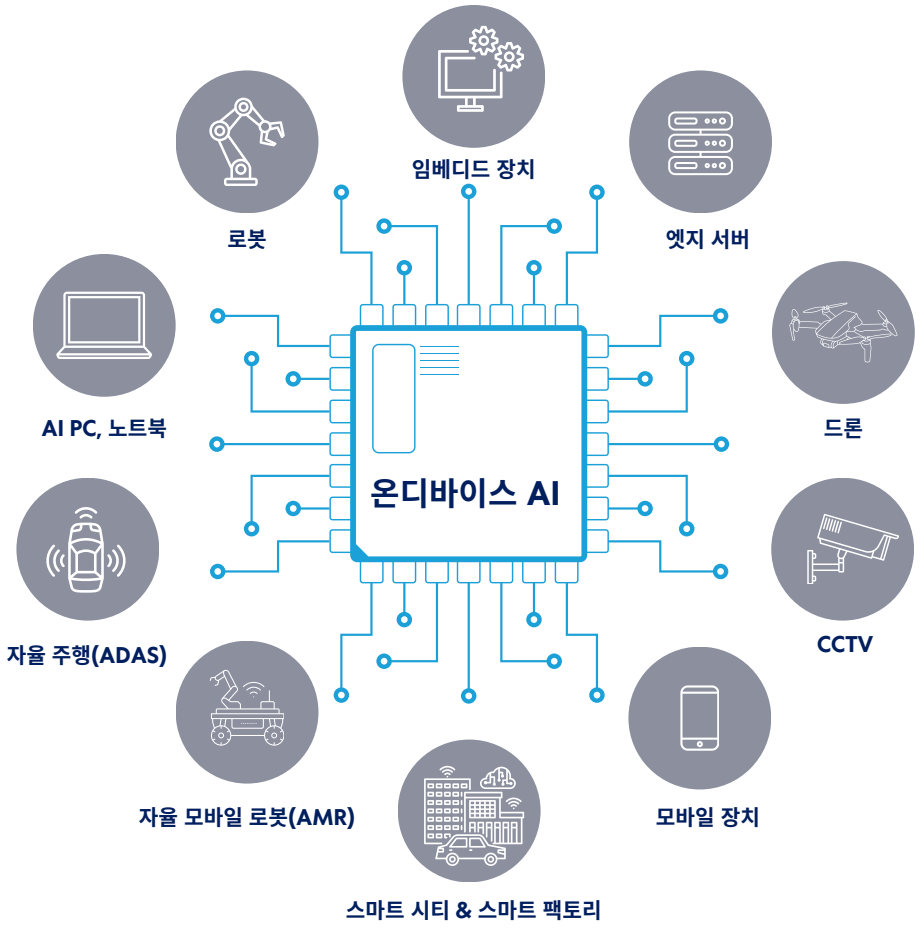
데이터 통합

AI 프로젝트의 성패는 데이터에 달려 있습니다. 모델 훈련, 튜닝, 추론에 필요한 양질의 데이터 세트를 확보하려면 사내, 클라우드, 엣지 등 다양한 환경에 분산된 데이터 통합이 필요합니다. 대원씨티에스는 VAST Data 플랫폼의 스토리지 레이어와 데이터 레이어 기반의 분산 공유 아키텍처를 제안합니다. 이 아키텍처는 다양한 환경에 분산된 데이터를 공유 방식으로 통합하고, 글로벌 네임스페이스로 데이터 접근성을 보장합니다.



경계를 넘어서는 데이터 접근, 엣지로의 확장

데이터가 생성되는 모든 위치에서 AI의 가능성을 실현할 수 있습니다. AI 알고리즘의 진화, 모델 배포와 관리 방식이 발전하면서 IoT나 센서를 장착한 작은 장치 그리고 현장에 배치한 AI 서버는 이제 데이터가 생성되는 최일선에서 AI 기능을 실행하는 기기가 되었습니다. 예를 들어 공공 영역에서는 곳곳에 위치한 CCTV를 AI 엣지 장치로 활용하는 스마트 시티 시나리오가 실현되고 있습니다. 산업계에서는 제조, 로봇, 의료, 안전 등 다양한 현장에서 NPU를 장착한 장치나 서버가 AI 엣지로 기능하고 있습니다. 대원씨티에스는 엔터프라이즈의 AI 전략을 엣지까지 확장하는 '온디바이스 AI' 시대의 청사진을 DEEPX와 협력을 통해 제시합니다.



GPU SERVER PORTPOLIOO

대원씨티에스의 AI 서버 포트폴리오

Kaytus

Kaytus는 클라우드, 데이터센터, AI, 엣지 환경을 위한 인프라 제품군을 공급하는 글로벌 기업입니다.
Kaytus GPU 서버는 고성능 연산, AI 혁신, LLM, HPC 등 AI를 위한 최적의 선택입니다.
미래를 주도하는 기술, Kaytus가 제공합니다.

GPU 서버

KR6288V2

GPU
Intel : Up to 1x NVIDIA HGX-Hopper-8GPU module, TDP up to 700W per GPU, AMD : Up to 1x NVIDIA HGX-Hopper-8GPU module, TDP up to 700W per GPU

Processor Support
Intel : 2x 4th or 5th Gen Intel® Xeon® Scalable Processors, AMD : 2x AMD EPYCTM 9004 Series Processors

Memory
Intel : Up to 32x DDR5 DIMMs(5,600 MT/s), AMD : Up to 24x DDR5 DIMMs(4,800 MT/s)

Interface
Intel : Up to 10x PCIe Gen5 x16 slots. One PCIe Gen5 x16 slot can be replaced with two x16 slots (PCIe Gen5 x8 rate), AMD : Up to 10x PCIe Gen5 x16 slots. One PCIe Gen5 x16 slot can be replaced with two x16 slots (PCIe Gen5 x8 rate)

Storage
24x 2.5’ SSD, up to 16x NVMe U.2

Power Supply
Intel : 2x 12V 3200W and 6x 54V 2700W, Titanium CRPS PSU with N+N redundancy, AMD : 2x 12V 3200W and 6x 54V 2700W, Titanium CRPS PSU with N+N redundancy

KR4268V2

GPU
Intel : Up to 8x Dual-slot FHFL PCIe interface GPU cards, and supports ≥ 4 PCIe 5.0 x16 slots, AMD : Up to 10x full-height full-length double-width PCIe interface GPU cards

Processor Support
Intel : 2x 4th Gen Intel® Xeon® Scalable Processors, AMD : 2x AMD 4th Generation EPYC™ processors

Memory
Intel : Up to 32x DDR5 DIMMs(4,800 MT/s), AMD : Up to 24x DDR5 DIMMs(4,800 MT/s)

Interface
Intel : Up to 8x Dual-slot FHFL PCIe interface GPU cards, and supports ≥ 4 PCIe 5.0 x16 slots, AMD : Up to 10x full-height full-length double-width PCIe interface GPU cards

Storage
24x 2.5 or 12x 3.5-inch SAS/SATA drive bays in the front, supports up to 16x NVME or E3.S

Power Supply
Intel : 4x 1600W/2000W/2200W/3000W 80Plus Platinum/Titanium PSUs, support N+N redundancy, AMD : 4x 1600W/2000W/2200W/3000W 80Plus Platinum/Titanium PSUs, supports N+N redundancy

KR2280V2

GPU
Up to 4 Double-width or
Up to 8 Single-width GPU

Processor Support
2x 4/5th Gen Intel® Xeon® Scalable Processors, AMD : 2x 4th Gen AMD EPYC™ processors

Memory
Intel : Up to 32 x DDR5 DIMMs(5,600 MT/s), AMD : Up to 24 x DDR5 DIMMs(4,800 MT/s)

Interface
Intel : Up to 16 PCIe expansion slots, including 1 RAID Mezz slot and 2 hot-swappable OCP 3.0 slots, AMD : Up to 8x PCIe Gen5 slots and 2x hot-swappable OCP3.0 slots or Up to 4x dual-width GPUs or 8x single-width GPUs Up to 4x dual-width GPUs

Storage
12 × 3.5-inch SAS/SATA/NVMe drives or 24 × 2.5-inch SAS/SATA/NVMe drives or 25 × 2.5-inch SAS/SATA drives(up to 4 × NVMe)

Power Supply
Intel : 2x 800W/1,300W/1,600W/2,000W/2,700W CRPS standard PSUs with 1+1 redundancy, AMD : 2x 550W/800W/1300W/1600W/2000W CRPS standard PSUs with 1+1 redundancy

대원씨티에스의 AI 서버 포트폴리오

Supermicro

Supermicro는 고성능 컴퓨팅 및 데이터센터 인프라 솔루션을 제공하는 글로벌 기업으로 유연한 확장 및 재구성을 통해 필요한 환경을 자유롭게 구현할 수 있는 제품군을 보유하고 있습니다. Supermicro의 다양한 인프라 솔루션을 활용하면 AI/ML, LLM, HPC 등 다양한 요구에 유연하게 대응 할 수 있습니다.

GPU 서버

AS -8125GS-TNMR2

GPU
Up to 8 onboard GPU Instinct MI300X Accelerators

Processor Support
AMD:2x 4th Gen AMD EPYC™ processors

Memory
Up to 6TB 4800MT/s ECC DDR5 RDIMM

Interface
Up to 16 CPU-to-GPU Interconnect, NVIDIA® NVLink® with NVSwitch™

Storage
Total 18,
12 front hot-swap 2.5" NVMe drive bays,
x 2 front hot-swap 2.5" SATA drive bays,
x 4 front hot-swap 2.5" NVMe* drive bays

Power Supply
6x 3000W Redundant Titanium Level(96%) power supplies

SYS-421GE-TNRT

GPU
Up to 10 double-width or 10 single-width GPU

Processor Support
Intel: 2X 4/5th Gen Intel® Xeon® Scalable processors

Memory
Up to 32 x DDR5 DIMMs(5,600 MT/s)

Interface
Up to 16 PCIe Gen5 Switch Dual-Root, ,Intel® Xe Link Bridges,13 PCIe 5.0 x16 FHFL slot(s)

Storage
Total 16 bays 8 front hot-swap 2.5" NVMe drive bays x 2

Power Supply
4x 2700W Redundant Titanium Level (96%) power supplies

SYS-221H-TNR

GPU
Up to 4 double-width or 8 single-width GPU

Processor Support
Intel: 2X 4/5th Gen Intel® Xeon® Scalable processors

Memory
Up to 32 x DDR5 DIMMs(5,600 MT/s)

Interface
4 PCIe 5.0 x16 FH/10.5"L double-widthslot(s), 2 PCIe 5.0 x16 AIOM slot(s) (OCP3.0 compatible), 2 M.2 NVMe/SATA slot(s)

Storage
8 front hot-swap 2.5" NVMe*/SAS*/SATA* drive bay

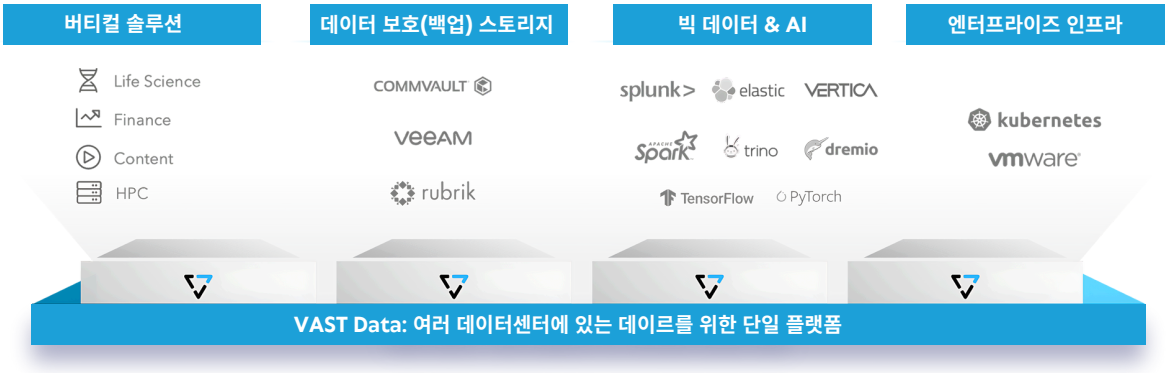
Power Supply
2x 1200W Redundant Titanium Level (96%) power supplies

대원씨티에스의 AI 스토리지 & 데이터 플랫폼

VAST Data

VAST Data는 AI/ML, 생성형 AI, HPC 워크로드를 위한 최적화된 데이터 솔루션을 제공합니다.
VAST Data의 플랫폼은 DASE(Disaggregated, Shared-Everything) 아키텍처를 기반으로 설계되어,
데이터 관리의 복잡성을 줄이고 높은 성능과 확장성을 동시에 제공합니다.

데이터 플랫폼



하드웨어 구성 요소

VAST QUAD SERVER / CL

CBOX(Controller)
4 x Stateless VAST Servers

I/O Connectivity
Up to 8 x 100GbE or InfiniBand

CPU
128 x Intel Xeon 2.1GHz Cascade Lake

Maximum Scale
Up to 10,000 VAST Servers

Power Supply
2 x 1600W

VAST MAVERICKS DF-5630

DBOX(Eclosure)
Meta Data, Write Buffer(12 x 1.5TB Storage Class Memory, U.2 NVMe), Data Store(44 x Hyperscale Flash 30.72TB, U.2 NVMe)

I/O Connectivity
4 x 100Gb Ethernet or EDR InfiniBand (Active/Active IO Modules)

Capacity
1,350TB / 1,195TB

Throughput
40GB/s

Power Supply
1,200W

VAST CERES DF-3060

DBOX(Eclosure)
Meta Data, Write Buffer(8 x 1.6TB Storage Class Memory, U.2 NVMe), Data Store(22 x Hyperscale Flash 61.44TB, E1.L NVMe)

I/O Connectivity
2 x Single NDR or Dual NDR200/HDR Ports (BF-3) or Ethernet

Capacity
1,350 / 1,194TB

Throughput
+50GB/s

Power Supply
500W

대원씨티에스의 AI 네트워크 플랫폼

Arista

Arista는 Meta나 Tesla 등 최근 AI 분야에서 상당한 실적을 올리고 있습니다. 또한, Arista는 최근 Etherlink AI 네트워킹 플랫폼을 공개하여 대규모 AI 클러스터를 위한 최적의 네트워크 성능을 제공합니다. 이 플랫폼은 수천에서 수십만 개의 XPU를 지원할 수 있는 효율적인 네트워크 토폴로지를 제공합니다.

네트워크 스위치

7800R3 Series

Switch Height
최대 32RU

Ports
최대 576 ports of 800G or
1,152 ports of 400G

Throughput
460 Tbps

Feature
- Non-blocking 460Tbps of system capacity
- High density 10/25/40/50/100/400G Ethernet
- Wirespeed TunnelSec encryption at 10G-400G
- MACsec, IPSsec and VXLANsec
- Scalable to 576 wire speed 400G ports
- ·Ultra deep buffers up to 24GB per line card
- Class leading latency of 3.5usec
- ·14.4Tbps of fabric capacity per line card
- Virtual Output Queues per port and CoS to eliminate head of line blocking

7050CX3-32S

Switch Height
최대 18RU

Ports
최대 288 ports of 400G or
432 ports of 100G

Throughput
230 Tbps

Feature
- Non-blocking 230Tbps of system capacity
- High density 10/25/40/50/100/400G Ethernet
- 288 wire speed 400G ports
- Ultra deep buffers up to 16GB per line card
- Class leading latency of 3.5usec
- 9.6Tbps of fabric capacity per line card
- Virtual Output Queues per port and CoS to eliminate head of line blocking

720DT-24S

Switch Height
2RU

Ports
64 ports of 800G or
128 ports of 400G

Throughput
460 Tbps

Feature
- 2X performance for 7060X Series(51.2Tbps: 512 x 100G PAM4 Peregrine SerDes, Powerful new SerDes help in supporting LPO optics)
- Efficient and Scalable Architecture(IPv4/v6, VxLAN and advanced instrumentation)
- High Radix with flexible port speeds(Up to 320 front panel ports at 10G to 800G speeds)
- Cloud-optimized pipeline and unified packet buffer(165MB shared buffer, Absorbs bursts 10x better)

대원씨티에스의 온디바이스 AI 솔루션

DEEPX

DEEPX는 온디바이스 AI 컴퓨팅을 위한 AI 반도체를 개발하는 회사입니다. DEEPX의 NPU 라인업은 전력 효율성, 계산 효율성, 비용 효율성, AI 정확도 및 최신 AI 알고리즘 지원 등의 경쟁력을 바탕으로 공공과 다양한 산업 분야에서 온디바이스 AI 비전을 실현하고 있습니다.

NPU 프로세서

DX-M1 (Accelerator)

Performance
25TOPS

Type
Accelerator (M.2 Module)

Feature
PCIe Gen3, ARM CPU LPDDR5

Process
5nm

DX-H1 Quattro (Accelerator)

Performance
100TOPS

Type
Accelerator (PCIe Card)

Feature
PCIe Gen3, ARM CPU LPDDR5

Process
5nm

DX-V1 (SoC Type)

Performance
5TOPS

Type
Application Processor

Feature
RISC-V CPU, ISP, LPDDR4, Video Codec

Process
28nm

DX-V3 (SoC Type)

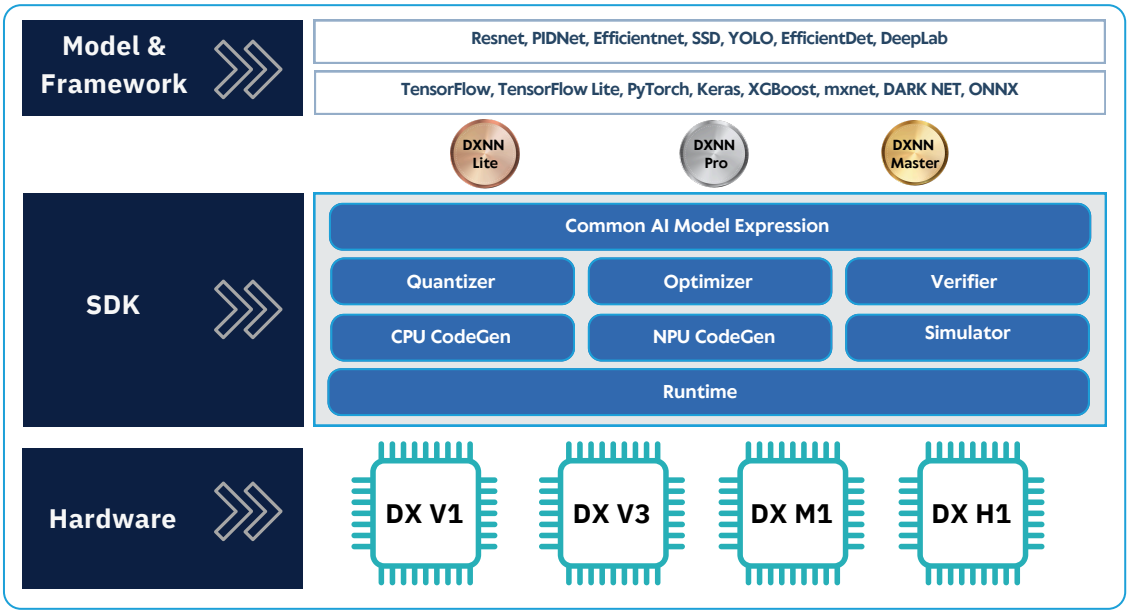
Performance
15TOPS

Type
Application Processor

Feature
ARM, ISP, DSP LPDDR5, Video Codec

Process
12nm

SDK

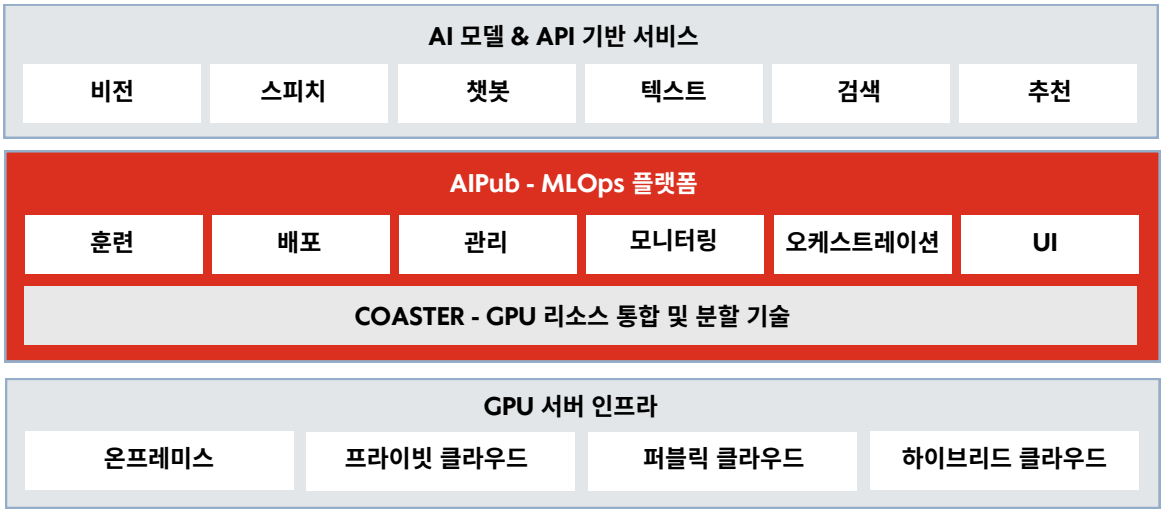


대원씨티에스의 MLOPS 플랫폼

TEN

TEN은 웹 UI 기반 MLOps 플랫폼 전문 기업입니다. TEN의 솔루션 라인업은 온프레미스, 클라우드 등 다양한 곳에 있는AI 컴퓨팅 자원을 GPU 파티셔닝 기술을 몰라도, 쿠버네티스 전문 인력이 없어도 누구나 쉽게 모델 개발, 훈련, 추론에 필요한 자원에 접근할 수 있는 편의성을 제공합니다.

MLOPS 플랫폼



도입 효과

1 Day

8.6개월이 소요되었던
환경 세팅 - AI 배포까지의 시간 단축

100% Util &
1/10 Cost

AI 모델 훈련 단계에는 최고의 가동률로 인프라를
사용하며, AI 운영 단계에는 최소로 인프라를 사용

Off-the-Shelf

다양한 분야의 기술 인력을 필요로 하는
복잡한 구축 과정을 생략한 준비된 플랫폼

Tailored

값 비싼 GPU 자원을 최고의 성능으로
사용할 수 있는 AI 전용 인프라 구성

고객의 비즈니스와 AI를 연결해
모든 가능성을 실현하는 구심점이 되겠습니다!

Website	www.dwcts.ai
Phone	02-2004-7700, 02-2004-7778(영업 문의)
E-mail	AI_sales@computer.co.kr
HQ address	본사: 서울시 용산구 청파로 109 나진전자월드빌딩 2층 판교 지사: 경기도 성남시 분당구 판교역로 240, 삼환하이펙스 A동 710-1, 711-1호