HS Hyosung AI Platform

HS 효성인포메이션시스템

GPU 서버 인프라와 AI 모델 운영 관리 시스템을 결합하여 성능, 자원 효율,
운영 관리 요건을 모두 충족하는 통합 AI 플랫폼

AI 플랫폼에 대한 고객의 고민

AI를 업무에 적용하여 비즈니스 성과를 도출하기 위한 기업의 고민은 IT 기술발전 속도만큼 커지고 있습니다.

도입한 인프라의 성능 최적화로 어떻게 AI 모델 학습시간을 단축할 것인지에 대한 기업의 고민은 어제오늘 일이 아니며, 추가로 한정된 GPU 자원의 효율적 사용과 만들어진 AI 모델에 대한 효율적인 관리 유지 보수 또한 AI 플랫폼을 운영하는 고객의 핵심 해결과제입니다.

HS효성 AI 플랫폼을 선택해야 하는 이유

기업의 AI 업무를 위한 HS효성인포메이션시스템의 통합 AI 플랫폼은
성능 최적화된 GPU 서버 인프라와 AI모델 운영 관리 시스템을 결합한 제품입니다.
성능, 자원 효율, 운영 관리를 최우선으로 고려한 AI 플랫폼으로
고객의 요구사항과 현재 운영 환경을 고려하여 맞춤 제안을 하고 있습니다.

성능 최적화	자원 효율	AI모델 운영 관리
GPU Direct Storage 적용을 통한 GPU 연산 성능 최적화	GPU 리소스의 효율적 사용을 위해 컨테이너 환경 기반 GPU 가상화	지속적으로 AI/ML 모델 개발 및 운영 적용 가능한 컨테이너 기반 AI 분석 플랫폼 구현
HCSF (초고성능 병렬 파일 저장 시스템) 연계 성능 최적화	NVIDIA MIG 기반 GPU 자원 분할 구현으로 다수의 사용자에게 GPU 자원 제공	사전 정의 분석 환경 템플릿의 버전별 제공으로 손쉬운 분석 환경 구현
NVIDIA HGX플랫폼은 고성능 GPU 탑재와 NVLink & NVSwitch 지원을 통해 최고 성능 제공	콜드데이터의 오브젝트 스토리지 Auto 티어링이 구현된 HCSF로 효율적인 저장 자원 확보 가능	직관적이고 사용자 친화적인 GUI 환경에서 컨테이너 운영 관리

01 성능 최적화

- NVLink 지원 슈퍼마이크로 HGX 플랫폼 제공으로 검증된 GPU 연산 성능 보장
- 고성능 병렬 파일 스토리지 연계 GPU Direct Storage/RDMA 구성 지원
- GPU 서버와 HCSF(초고성능 병렬파일 스토리지)를 같이 공급하여 연산 성능 최적화

GPU with NVLink

NVIDIA HGX

- NVIDIA NVLINK & NVSwitch 5세대로 PCIe 5.0 대비 14배 고성능
- NVIDIA B200/B300 Tensor 코어 GPU의 고속 상호 연결 구현

NVLink와 PCIe 비교

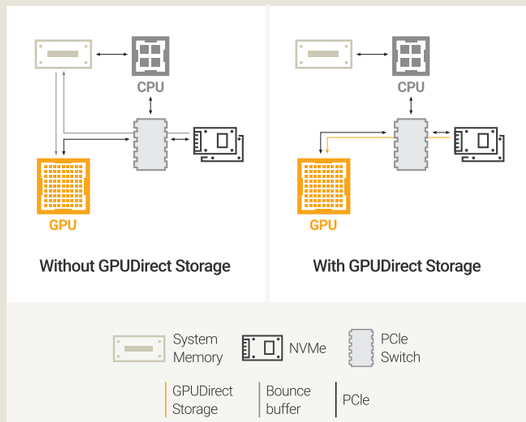
구분	PCIe Gen (규격)	NVLink Gen (NVIDIA)	GPU 아키텍처	서브링크 당 속도	총 서브링크 수(Link 당)	GPU당 총 양방향 속도
Pascal	PCIe 3.0	1.0	P100 (2016)	20 GB/s	4 links (8 lanes)	160 GB/s
Volta	PCIe 3.0/4.0	2.0	V100 (2017)	25 GB/s	6 links (12 lanes)	300 GB/s
Ampere	PCIe 4.0	3.0	A100 (2020)	50 GB/s	12 links (4 links X 3 group)	600 GB/s
Hopper	PCIe 5.0	4.0	H100 (2022)	50 GB/s (100GT/s)	18 links	900 GB/s
Black well	PCIe 5.0	5.0	B200/B300 (2024~)	100 GB/s (200GT/s)	18 links	1.8 TB/s

- GPU-GPU, CPU-GPU 간 전송 기술로 GPU 메모리 직접 통신
- NVLink 1세대에서 NVLink 5세대로 NVLink 방식 발전

GPUDirect Storage/RDMA

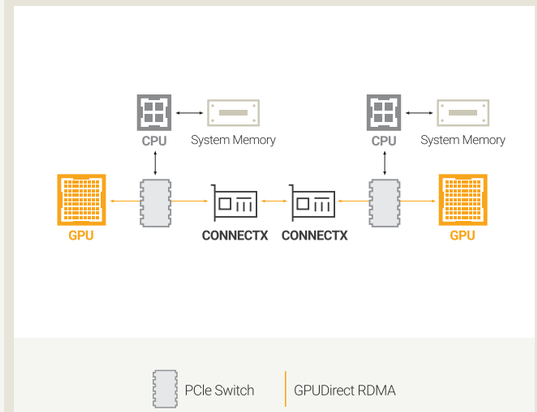
GPUDirect Storage/RDMA를 활용한 IO 성능 최적화로 AI 모델의 학습 시간을 단축할 수 있습니다.

스토리지 I/O GPUDirect Storage



- NVMe/NVMe-oF 스토리지와 GPU 메모리 간 직접 연결
- CPU 메모리의 바운스 버퍼를 제거
- 스토리지와 GPU 메모리의 데이터 로드 프로세스 IO 개선

네트워크 I/O GPUDirect RDMA



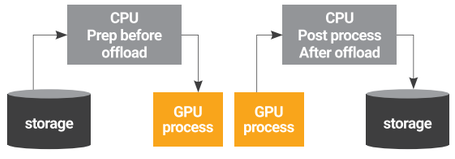
- RDMA를 통해 PCIe가 GPU 메모리에 직접 액세스
- 원격 시스템에서 NVIDIA GPU 간의 직접 통신을 제공
- CPU와 메모리 데이터 버퍼를 제거, 10배 성능 향상

GPUDirect Storage 성능

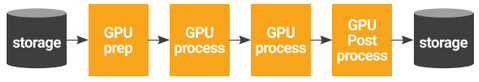
GPUDirect 기술을 적용하여, CPU 대비 보다 빠르고 많은 대역폭의 IO 처리 가능

I/O 흐름 : CPU vs GPU

CPU 프로세싱 처리 필요

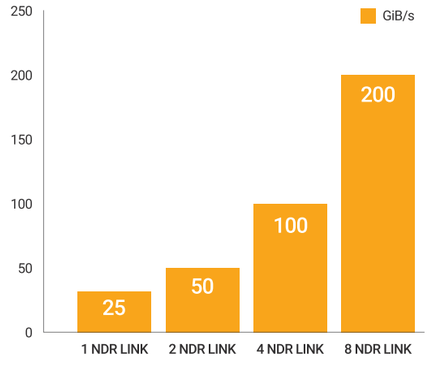


빠른 IO로 성능향상



GPU서버 + GDS 지원가능 스토리지 동시 구성 → 고성능 제공

WekaFS READ THROUGHPUT TO SINGLE CPU SERVER

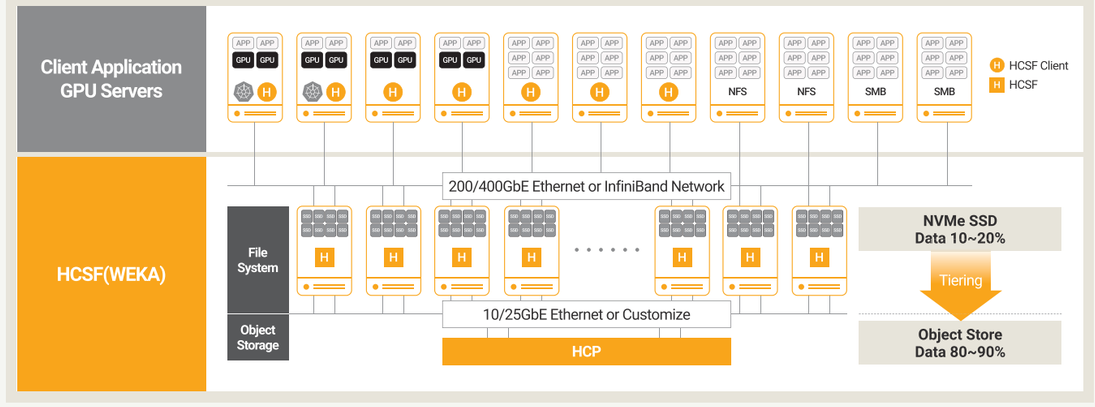


Auto 티어링기반 초고성능 병렬파일 스토리지 (HCSF)

정책 기반 오브젝트 스토리지 Auto 티어링으로 비용 대비 고성능, 대용량 제공

초고성능 병렬 파일시스템 + 대용량 Object Storage = 고성능 Scale-Out Storage

GPUDirect Storage/RDMA 구성



02 자원효율

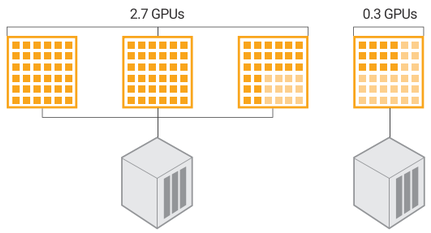
- GPU 리소스의 효율적 사용을 위해 컨테이너 환경 기반 GPU 가상화
- NVIDIA MIG 기반 GPU 자원 분할 구현으로 다수의 사용자에게 GPU 자원 제공
- 다수의 GPU 서버 리소스를 GPU 종류별 그룹으로 묶어서 사용자 및 프로젝트 그룹에 할당하여 리소스 성능 및 자원 효율에 최적화된 구성 제공

컨테이너 기반 GPU 분할 가상화

GPU 분할 가상화

컨테이너 기반 GPU 스케일링

- 교육 및 추론 워크로드를 위한 단일 GPU 공유
- 모델 훈련 등 대규모 워크로드를 위한 다중 GPU 할당
- 자체 개발한 CUDA 가상화 계층으로 구현

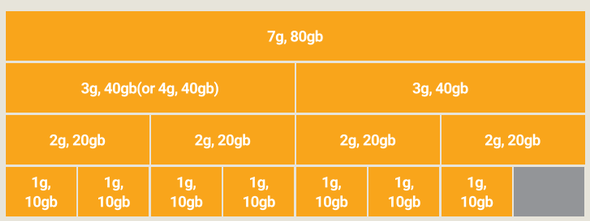


NVIDIA MIG기반 GPU 분할

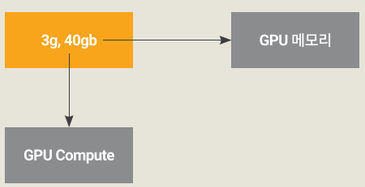
MIG 지원 GPU 리스트

제품명	아키텍처	인스턴스 유형	GPU 최대 분할
B200	Blackwell	Up to 7x 23GB Up to 4x 45GB Up to 2x 90GB Up to 1x 180GB	7
RTX PRO 6000	Blackwell	Up to 4x 24GB Up to 2x 48GB Up to 1x 96GB	4
H200	Hopper	Up to 7x 18GB Up to 4x 35GB Up to 2x 71GB Up to 1x 141GB	7
H100	Hopper	Up to 7x 10GB Up to 4x 20GB Up to 2x 40GB Up to 1x 80GB	7

구성 예



※선택 블록이 수직으로 겹치지 않는 구성 가능



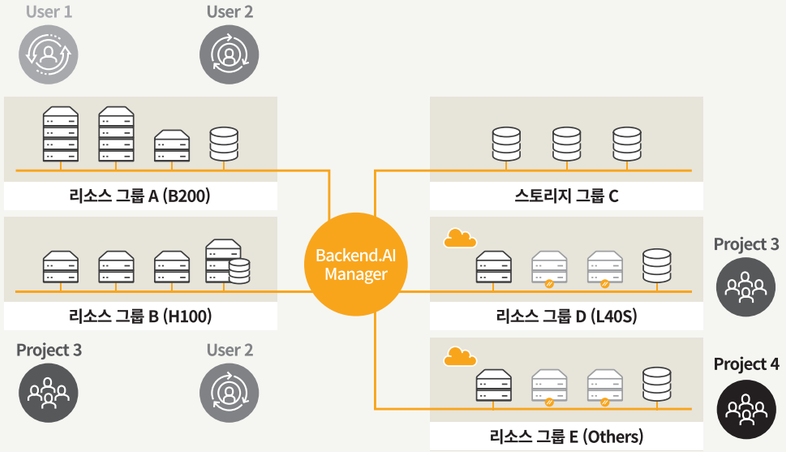
GPU 리소스 그룹별 사용자 할당

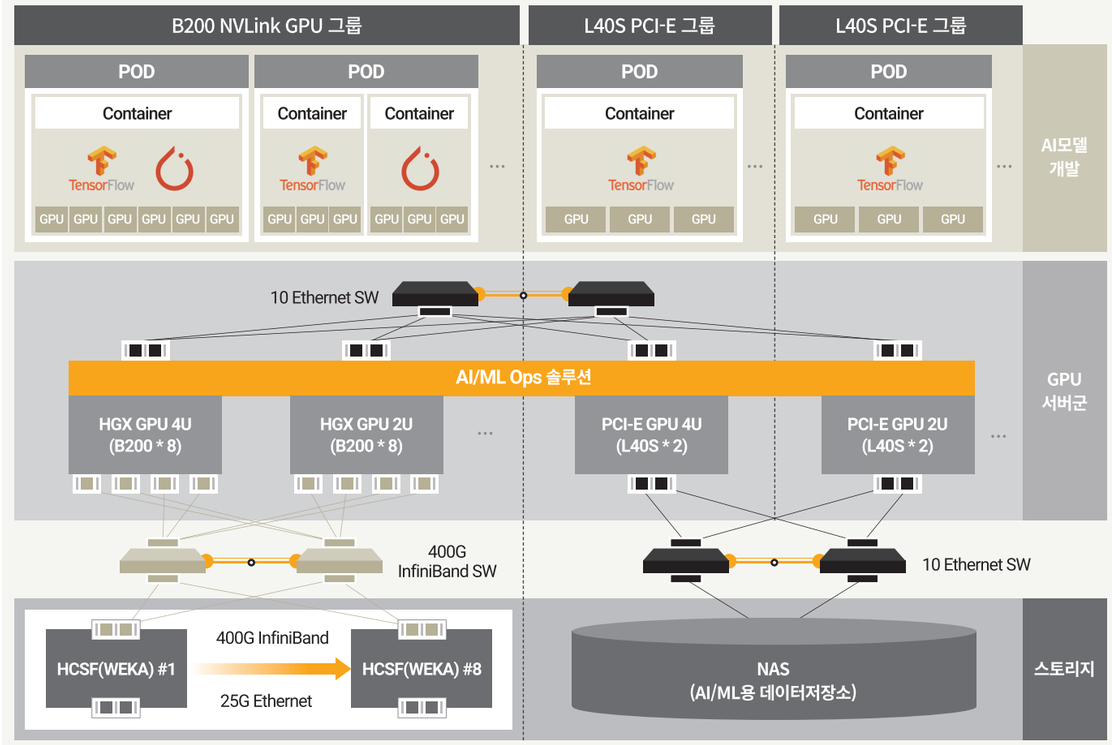
리소스 그룹 A
NVIDIA B200 GPU 그룹

리소스 그룹 B
NVIDIA H100 GPU 그룹

사용자 별로 사용 자원을 분리
User 1은 B200만을 사용,
User 2는 B200 및 H100을
모두 사용

프로젝트에 따라 사용 자원을 분리
프로젝트 3은 H100그룹 및
L40S GPU 그룹을
사용할 수 있음.
프로젝트 4는 Others만 사용





03 AI모델 운영관리

- 지속적으로 AI/ML 모델 개발 및 운영 적용 가능한 컨테이너 기반 AI 분석 플랫폼 구현
- 사전 정의 분석 환경 템플릿의 버전별 제공으로 손쉬운 분석 환경 구현
- 직관적이고 사용자 친화적인 GUI 환경에서 컨테이너 운영 관리

컨테이너 기반 사전 정의 AI 플랫폼

AI, ML, HPC를 R&D부터 Business Service, AI Service 추론 및 제공까지 하나의 일관된 플랫폼을 통해 효과적으로 관리

GPU 활용 극대화

컨테이너 수준 GPU 분할 가상화

NVIDIA GPU MIG 지원

직관적인 관리 및 사용자 경험

GUI 기반 컨테이너 운영 관리

웹 UI와 데스크탑 앱 지원

사전 정의 AI개발환경 제공

Tensorflow, Pytorch 등 사전 정의 이미지 제공

연구환경 선택 즉시 생성

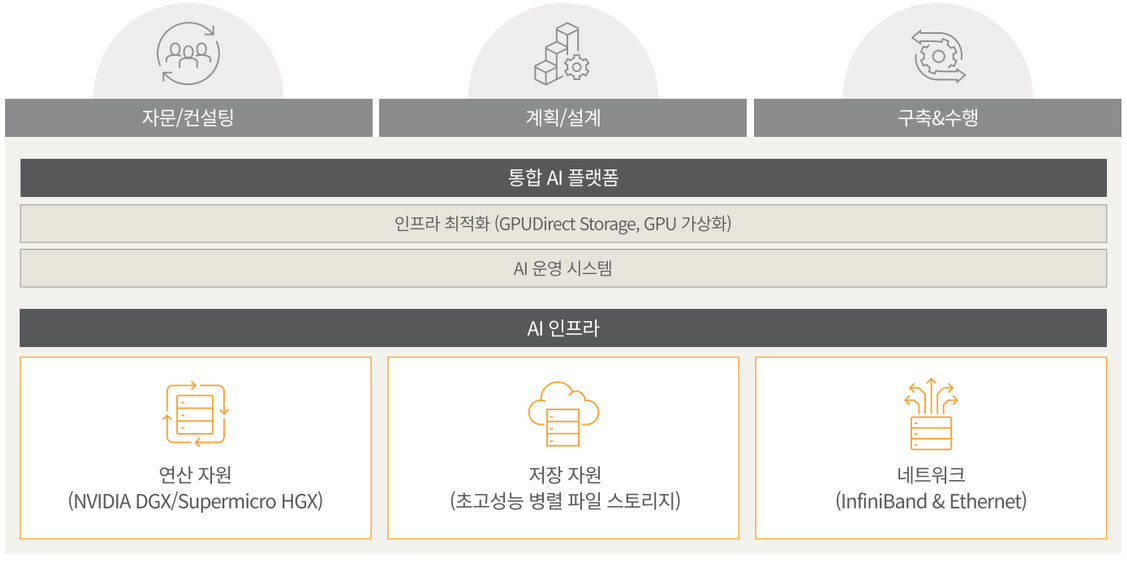
AI 및 HPC 성능 최적화

독자적 엔진으로 최적의 GPU 연산 자원 배치 구현

다중 노드 워크로드 및 데이터 I/O 병렬화 지원

HS효성의 통합 AI플랫폼 역량

AI 활용을 통한 기업혁신을 위해서는 클라우드, 컨테이너, 가상머신, GPU, 빅데이터, AI Ops 솔루션, 고성능 스토리지, 네트워크 등 많은 구성요소들이 필요하며 이는 비용, 업무 효율, 관리 측면에서 큰 변화가 이루어져야 합니다. 변화하고자 하는 기업들은 많은 어려움에 직면해 있습니다. 기업 내 한정된 GPU 연산 자원을 어떻게 효율적으로 사용할 것 인지, 분석을 위해 필요한 대량의 데이터를 어떻게 저장하는 것이 가장 효율적인지 고민이 깊어져 가고 있습니다. 또한 기존 전통적 인프라의 낮은 성능으로는 AI 모델 개발 및 운영을 할 수 없으며 단순하게 GPU 서버만 도입한다고 성능이 개선되지 않습니다. HS효성인포메이션시스템은 이를 해결하기 위해 단순한 고속 인프라 장비 공급에 그치지 않고 실질적인 AI 플랫폼 구축을 위해 비용, 업무 효율, 관리 측면에서 최적의 솔루션과 서비스를 제공하고 있습니다.



AI플랫폼을 위한 GPU 서버 인프라

HS효성은 Supermicro의 최신 GPU(NVIDIA/AMD)서버를 기반으로 한 통합 AI플랫폼을 제공합니다.
NVIDIA의 최신GPU 서버와(Blackwell, Hopper) AMD의 최신 MI350X 시스템을 공급합니다.

Supermicro GPU 서버

NVIDIA HGX 서버 - SXM

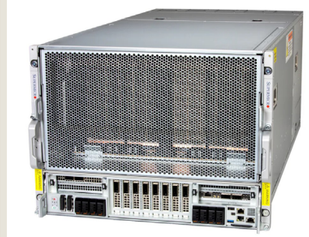
SYS-822GS-NB3RT

- AI/Deep Learning Training
- Conversational AI / Scientific Research

1. **Chassis**
   8U 랙마운트 타입
2. **Processor Support**
   Dual Socket E2 (LGA-4710)
   Intel® Xeon® 6700 series processors with P-cores
3. **Memory Capacity**
   최대 8TB (32DIMMs)
4. **GPU**
   NVIDIA SXM :
   HGX B300 8GPU (288GB)
5. **PCIe Expansion Slots**
   2 PCIe5.0 x16 FHHL slots
6. **Power Supply**
   6x 6600W Redundant (3+3) Titanium Level (96%) power supplies

SYS-822GS-NBRT

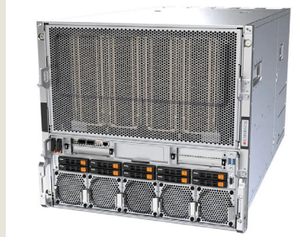
- Large Language Model and Generative AI
- AI/Deep Learning Training and Inference



1. **Chassis**
   8U 랙마운트 타입
2. **Processor Support**
   Dual Socket E2 (LGA-4710)
   Intel® Xeon® 6700/6500 series processors with P-cores
3. **Memory Capacity**
   최대 8TB (32DIMMs)
4. **GPU**
   NVIDIA SXM :
   HGX B200 8GPU (180GB)
5. **PCIe Expansion Slots**
   8 PCIe5.0 x16 LP slots
   2 PCIe5.0 x16 FHHL slots
6. **Power Supply**
   6x 6600W Redundant (3+3) Titanium Level (96%) power supplies

SYS-A22GA-NBRT

- High Performance Computing (HPC)
- AI/Deep Learning Training and Inference



1. **Chassis**
   10U 랙마운트 타입
2. **Processor Support**
   Dual Socket BR (LGA-7529)
   Intel® Xeon® 6900 series processors with P-cores
3. **Memory Capacity**
   최대 6TB (24DIMMs)
4. **GPU**
   NVIDIA SXM :
   HGX B200 8GPU (180GB)
5. **PCIe Expansion Slots**
   10 PCIe5.0 x16 LP slots
   2 PCIe5.0 x16 FHHL slots
6. **Power Supply**
   6x 5250W Redundant (3+3) Titanium Level (96%) power supplies

NVIDIA GPU 서버 – PCIE

SYS-322GA-NR

- Industrial Automation, Retail
- Conversational AI
- Business Intelligence & Analytics

1. **Chassis**
   3U 랙마운트 타입
2. **Processor Support**
   Dual Socket BR (LGA-7529)
   Intel® Xeon® 6900 series processors with P-cores
3. **Memory Capacity**
   최대 6TB (24DIMMs)
4. **GPU**
   NVIDIA PCIe : RTX PRO 6000 Blackwell Server edition, H200 NVL, L40S
5. **Drive Bays**
   6 front hot-swap E1.S NVMe drive bays
6. **PCIe Expansion Slots**
   8 PCIe5.0 x16 FHFL double-width slots
7. **Power Supply**
   3x 3200W Redundant (2+1) Titanium Level (96%) power supplies

AS-5126GS-TNRT2

- AI/Deep Learning Training
- Visualization / Simulation
- Multimedia/Digital Content creation

1. **Chassis**
   5U 랙마운트 타입
2. **Processor Support**
   Dual Dual processor(s)
   AMD EPYC™ 9005/9004 Series Processors
3. **Memory Capacity**
   최대 6TB (24DIMMs)
4. **GPU**
   NVIDIA PCIe : RTX PRO 6000 Blackwell Server edition, H200 NVL, L40S
5. **Drive Bays**
   2 front hot-swap 2.5" SATA drive bays
   8 front hot-swap 2.5" NVMe drive bays
6. **PCIe Expansion Slots**
   13 PCIe5.0 x16 FHFL slots
7. **Power Supply**
   6x 2700W Redundant (4+2) Titanium Level (96%) power supplies

SYS-521GE-TNRT

- High Performance Computing
- VDI / 3D Rendering
- Design & Visualization

1. **Chassis**
   5U 랙마운트 타입
2. **Processor Support**
   Dual Socket E (LGA-4677)
   Intel® Xeon® 5th Gen Scalable processors
3. **Memory Capacity**
   최대 8TB (32DIMMs)
4. **GPU**
   NVIDIA PCIe : RTX PRO 6000 Blackwell Server edition, H200 NVL, L40S
5. **Drive Bays**
   24 front hot-swap 2.5" NVMe/SATA drive bays
6. **PCIe Expansion Slots**
   13 PCIe5.0 x16 FHFL slots
7. **Power Supply**
   4x 2700W Redundant Titanium Level (96%) power supplies

Grace Hopper MGX 서버

ARS-221GL-NR

- High Performance Computing
- AI/Deep Learning Training
- Large Language Model (LLM) Natural Language Processing

❶ **Processor Support**	NVIDIA Grace CPU Superchip with 144 Cores
❷ **Memory Capacity**	최대 960GB : Onboard LPDDR5X DRAM 메모리타입 : 4800MHz ECC DDR5 DRAM (LPDDR5X)
❸ **GPU**	NVIDIA PCIe : H100 NVL, L40S
❹ **PCI-E Expansion Slots**	7 PCIe 5.0 x16 FHFL slot(s)
❺ **Storage I/F**	E1.S
❻ **Drive bays**	8x E1.S hot-swap NVMe drive slots
❼ **Power Supply**	2,000W(2+1) Redundant Titanium Level (96%+) Power Supplies

ARS-111GL-NHR

- High Performance Computing
- AI/Deep Learning Training and Inference
- Large Language Model (LLM) and Generative AI

❶ **Processor Support**	NVIDIA GH200 Grace Hopper™ Superchip Up to 72C/144T
❷ **Memory Capacity**	최대 480GB : Onboard LPDDR5X (and additional 96GB HBM3 for GPU)
❸ **GPU**	NVIDIA GH200 Grace Hopper : H100 (최대 1)
❹ **PCI-E Expansion Slots**	3 PCIe 5.0 x16 FHFL slot(s)
❺ **Storage I/F**	E1.S
❻ **Drive bays**	8x E1.S hot-swap NVMe drive slots
❼ **Power Supply**	2x 2000W Redundant Titanium Level (96%+) Power Supplies

AMD GPU 서버 (MI350X/MI300X)

AS-8126GS-TNMR

- High Performance Computing
- AI/Deep Learning Training
- Machine Learning (ML)
- Data center

❶ **Processor Support**	Dual Processor(s) AMD EPYC™ 9004/9005 Series
❷ **Memory Capacity**	12-channel DDR5 memory support 24 DIMM slots for up to 6 TB ECC DDR5-6400 RDIMM
❸ **GPU**	AMD Instinct MI350X Accelerators, Instinct MI325X Accelerators
❹ **PCI-E Expansion Slots**	8 PCIe 5.0 x16 LP slots 2 PCIe 5.0 x16 FHHL slots
❺ **I/O ports**	1 RJ45 Dedicated IPMI LAN port 2 USB 3.0 Ports (rear) 1 VGA Connector
❻ **Drive bays**	2 front hot-swap 2.5" SATA drive bays 8 front hot-swap 2.5" NVMe drive bays
❼ **Power Supply**	6x 5250W Redundant (3+3) Titanium-Level Power Supplies

AS-8125GS-TNMR2

- High Performance Computing
- AI/Deep Learning Training
- Industrial Automation Retail
- Climate and Weather Modeling

❶ **Processor Support**	Dual Processor(s) AMD EPYC™ 9004/9005 Series
❷ **Memory Capacity**	12-channel DDR5 memory support 24 DIMM slots for up to 6 TB ECC DDR5-4800 RDIMM
❸ **GPU**	AMD Instinct MI300X Platform with 8 MI300x OAM GPUs
❹ **PCI-E Expansion Slots**	8 PCIe 5.0 x16 low-profile slots connected to GPU via PCIe switch 2 PCIe 5.0 x16 full-height full-length slots Optional 2 PCIe 5.0 x16 slots via expansion kit
❺ **I/O ports**	1 RJ45 Dedicated IPMI LAN port 2 USB 3.0 Ports (rear) 1 VGA Connector
❻ **Drive bays**	12 PCIe 5.0 x4 NVMe U.2 drives Optional 4 PCIe 5.0 x4 NVMe U.2 drives2 2x M.2 NVMe boot drive 2 hot-swap 2.5" SATA drives2
❼ **Power Supply**	6x or 8x 3000W N+N redundant Titanium-Level power supplies

2025년 12월
www.his21.co.kr

본 카탈로그에 수록된 솔루션 사양은 인쇄일을 기준으로
사전 고지 없이 변경될 수 있으며, 최신 사양은 당사 영업대표 또는
홈페이지를 통해 확인하시기 바랍니다.
솔루션 관련 문의는 홈페이지의 <제품문의>를 통해 연락 부탁드립니다.

본사 서울특별시 강남구 도산대로 524 청담빌딩 5층 **TEL** 02-510-0300 **FAX** 02-547-9998
부산사무소 부산광역시 해운대구 센텀서로 30 KNN 타워 1303호 **TEL** 051-784-7811, 7813 **FAX** 051-463-7805
대구사무소 대구광역시 동구 화랑로 47 (신천동, 전문건설회관 3층) **TEL** 053-426-9800 **FAX** 053-426-9830
서부사무소 대전광역시 서구 둔산서로 59 고운손빌딩 702호 **TEL** 042-485-4856 **FAX** 042-484-0366
광주사무소 광주광역시 서구 상무연하로 112 제갈량비즈타워 3층 **TEL** 062-385-2193 **FAX** 062-385-2194
수원사무소 경기도 수원시 영통구 삼성로 182–1 R7빌딩 3층 **TEL** 031-216-8717~8 **FAX** 031-216-8719