Hyosung AI Platform

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, 추가로한 정된 GPU 자원의 효율적 사용과 만들어진 AI 모델에 대한 효율적인 관리 유지 보수 또한 AI 플랫폼을 운영하는 고객의 핵심 해결 과제입니다.

효성 AI 플랫폼을 선택해야 하는 이유

기업의 AI 업무를 위한 효성인포메이션의 통합 AI 플랫폼은
성능 최적화 된 GPU 서버 인프라와 AI모델 운영관리 시스템을 결합한 제품 입니다.
성능, 자원 효율, 운영관리를 최우선으로 고려한 AI 플랫폼으로
고객의 요구사항과 현재 운영 환경을 고려하여 맞춤 제안을 하고 있습니다.

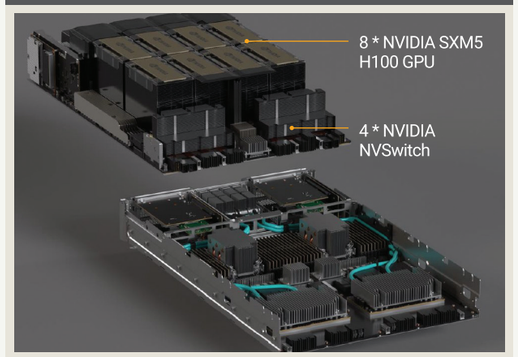
성능 최적화	자원 효율	AI모델 운영 관리
GPU Direct Storage 적용을 통한 GPU 연산성능 최적화	GPU 리소스의 효율적 사용을 위해 컨테이너 환경 기반 GPU 가상화	지속적으로 AI/ML 모델 개발 및 운영 적용 가능한 컨테이너 기반 AI 분석 플랫폼 구현
HCSF (초고성능 병렬 파일 저장 시스템) 연계 성능 최적화	NVIDIA MIG, vCS 기반 GPU 자원 분할 구현으로 다수의 사용자에게 GPU 자원 제공	사전 정의 분석 환경 템플릿의 버전별 제공으로 손쉬운 분석 환경 구현
NVIDIA HGX 플랫폼은 SXM5방식의 GPU 탑제와 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 NVIDIA DGX 및 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 DGX / HGX



- 4세대 NVIDIA NVLink는 900GB/s GPU 대역폭으로 PCIe Gen4 대비 14배 고성능
- NVIDIA H100 Tensor 코어 GPU의 고속 상호 연결 구현

NVLink와 PCIe 비교

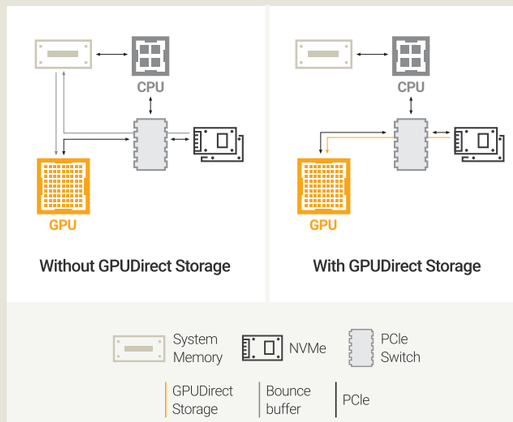
항목	서브 링크 속도	서브 링크 수	전체 속도	GPU 아키텍처
PCIe 3.x	16GB/s	1	16GB/s	Pascal, Volta, Turing
PCIe 4.0	32GB/s	1	64GB/s	Volta, Ampere
NVLink 1.0	20GB/s	4	160GB/s	Pascal
NVLink 2.0	25GB/s	6	300GB/s	Volta
NVLink 3.0	25GB/s	12	600GB/s	Ampere
NVLink 4.0	25GB/s	18	900GB/s	Hopper

- GPU-GPU, CPU-GPU간 전송 기술로 GPU 메모리 직접 통신
- NVLink 1세대에서 NVLink 4세대로 NVLink 방식 발전

GPUDirect Storage/RDMA

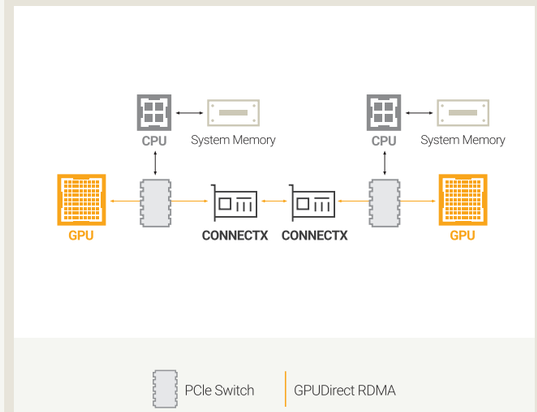
GPUDirect Storage/RDMA를 활용한 IO 성능 최적화로 AI 모델의 학습 시간을 단축할 수 있습니다.

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



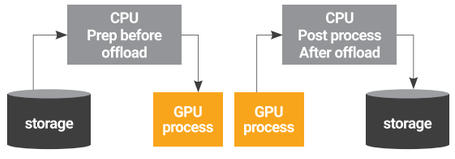
- RDMA를 통해 PCIe가 GPU 메모리에 직접 액세스
- 원격 시스템에서 NVIDIA GPU 간의 직접 통신을 제공
- CPU와 메모리 데이터 버퍼를 제거, 10배 성능 향상

GPUDirect Storage 성능

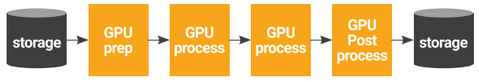
GPUDirect 기술을 적용하여, CPU 대비 보다 빠르고 많은 대역폭의 IO 처리 가능
NVIDIA GPU 서버와 HCSF 구성 시, 8 EDR LINKS & DGX 서버 1대에서 약 80GB/s 고성능 제공

I/O 흐름 : CPU vs GPU

CPU 프로세싱 처리 필요



빠른 IO로 성능향상



GPU서버 + HCSF 동시 구성 → 고성능 제공

WekaFS READ THROUGHPUT TO SINGLE CPU SERVER

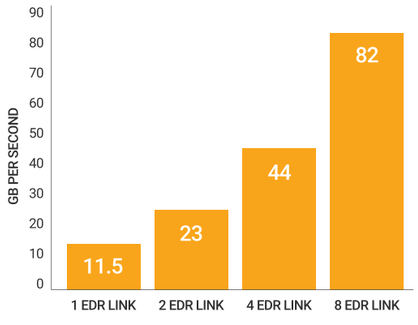
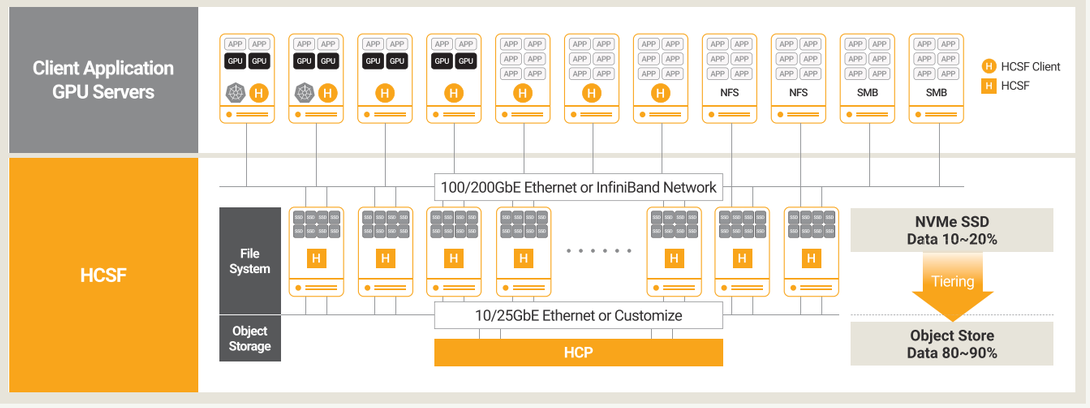


Figure 20: Performance Scaling Inside a Single NVIDIA DGX-2 GPU Server with GPUDirect Storage

Auto 티어링기반 초고성능 병렬파일 스토리지 (HCSF)

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

GPUDirect Storage/RDMA 구성



02 자원효율

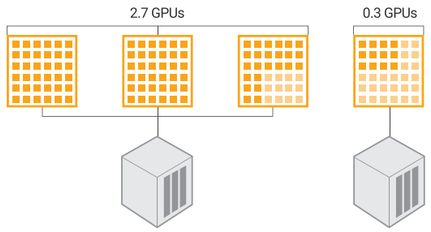
- GPU리소스의 효율적 사용을 위해 컨테이너 환경 기반 GPU 가상화
- NVIDIA MIG, vCS 기반 GPU자원 분할 구현으로 다수의 사용자에게 GPU자원 제공
- 다수의 GPU 서버 리소스를 GPU 종류별 그룹으로 묶어서 사용자 및 프로젝트 그룹에 할당하여 리소스 성능 및 자원 효율에 최적화된 구성 제공

컨테이너 기반 GPU 분할 가상화

Lablup Backend.AI의 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
  [Lablup Backend.AI 대한민국, 미국, 일본 등록 특허]

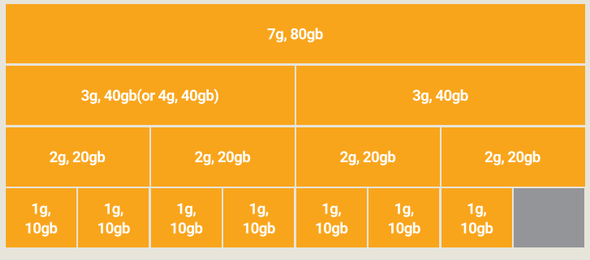


NVIDIA MIG기반 GPU분할

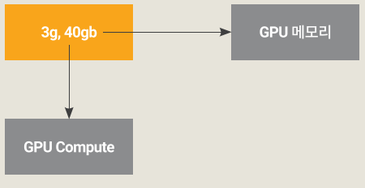
MIG 지원 GPU 리스트

제품명	아키텍처	메모리	GPU 최대 분할
H100-SXM5	NVIDIA Hopper	80GB	7
H100-PCIe	NVIDIA Hopper	80GB	7
A100-SXM4	NVIDIA Ampere	40GB	7
A100-SXM4	NVIDIA Ampere	80GB	7
A100-PCIe	NVIDIA Ampere	40GB	7
A100-PCIe	NVIDIA Ampere	80GB	7
A30	NVIDIA Ampere	24GB	4

구성 예



※선택 블록이 수직으로 겹치지 않는 구성 가능



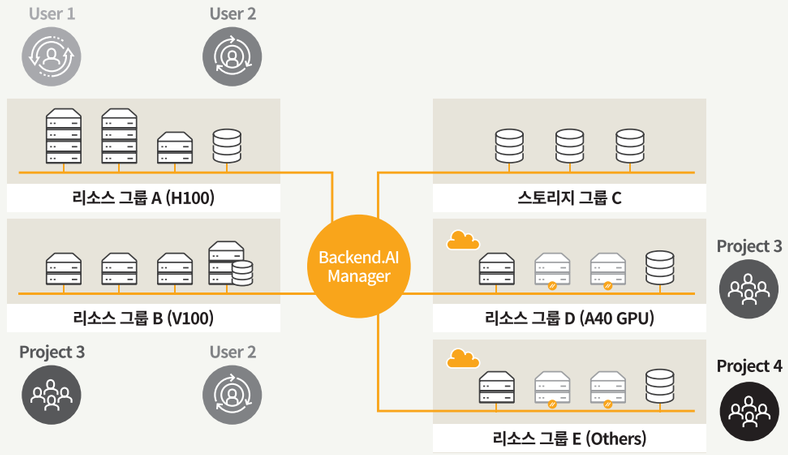
GPU 리소스 그룹별 사용자 할당

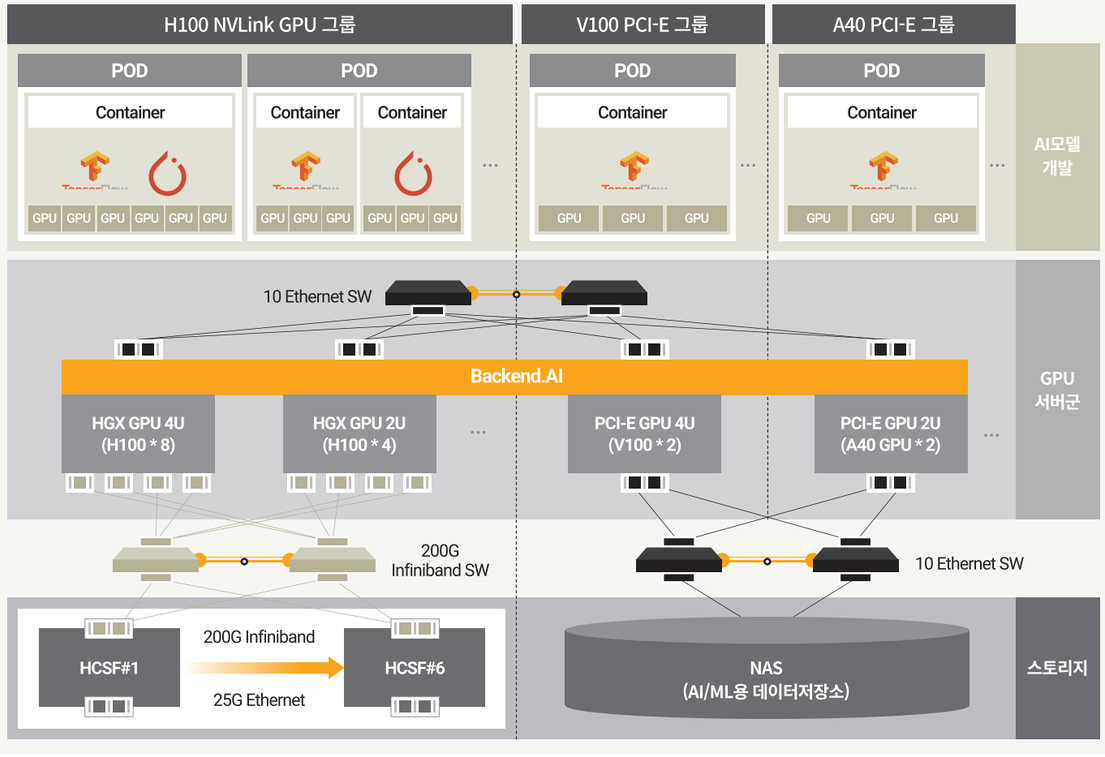
리소스 그룹 A
NVIDIA H100 GPU 그룹

리소스 그룹 B
NVIDIA H100 GPU 그룹

사용자 별로 사용 자원을 분리
User 1은 H100만을 사용,
User 2는 H100 및 V100을
모두 사용

프로젝트에 따라 사용 자원을 분리
프로젝트 3은 V100 그룹 및
A40 GPU그룹을
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

Lablup Backend.AI는 아태지역 최초 NVIDIA DGX-Ready Software로 검증된 AI 연구&개발 플랫폼으로
GPU 분할 가상화를 제공하여, 데이터 과학자 및 AI 플랫폼 담당자 들이 효율적으로 연산 자원을 사용할 수 있습니다.

Lablup

GPU 활용 극대화

- 컨테이너 수준 GPU분할 가상화
- NVIDIA GPU MIG 지원

직관적인 관리 및 사용자 경험

- GUI 기반 컨테이너 운영 관리
- 웹UI와 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

효성의 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. 효성인포메이션시스템은 이를 해결하기 위해 단순한 고속 인프라 장비 공급에 그치지 않고 실질적인 AI 플랫폼 구축을 위해 비용, 업무효율, 관리 측면에서 최적의 솔루션과 서비스를 제공하고 있습니다.

자문/컨설팅 | 계획/설계 | 구축&수행

통합 AI 플랫폼

인프라 최적화 (GPUDirect Storage, GPU가상화)

AI 운영 시스템 (컨테이너: Lablup Backend.AI 가상머신: VMware)

AI 인프라

- 연산 자원 (NVIDIA DGX/Supermicro HGX)
- 저장 자원 (초고성능 병렬 파일 스토리지)
- 네트워크 (Cisco&Mellanox)

AI플랫폼을 위한 GPU 서버 인프라

효성은 Supermicro의 GPU 서버와, NVIDIA DGX 서버를 기반으로한 통합 AI 플랫폼을 제공합니다.

Supermicro GPU 서버

H100 GPU 서버 – HGX

SYS-821GE-TNHR

- Deep Learning Data Analytics and HPC
- Data Center Infrastructure

1. **Processor Support**
   Dual Socket E (LGA-4677)
   4th Gen Intel® Xeon® Scalable processors
2. **Memory Capacity**
   최대 8TB (32DIMMs)
3. **GPU**
   HGX H100 8-GPU SXM5
   Multi-GPU Board
4. **PCI-E Expansion Slots**
   8 PCIe 5.0 x16 LP slot(s)
   2 PCIe 5.0 x16 FHFL slot(s)
5. **I/O ports**
   1x VGA, 2x USB 3.0,
   and 1x Dedicated IPMI
6. **Power Supply**
   6x or 8x 3000W redundant
   Titanium level power supplies

AS -8125GS-TNHR

- High Performance Computing
- AI/Deep Learning Training

1. **Processor Support**
   Dual Socket SP5 AMD EPYC™
   9004 Series Processors
2. **Memory Capacity**
   최대 6TB (24DIMMs)
3. **GPU**
   HGX H100 8-GPU SXM5
   Multi-GPU Board
4. **PCI-E Expansion Slots**
   8 PCIe 5.0 x16 LP slot(s)
   2 PCIe 5.0 x16 FHFL slot(s)
5. **I/O ports**
   1x VGA, 2x USB3.0,
   1x Dedicated IPMI RJ45 port
6. **Power Supply**
   6x or 8x 3000W redundant
   Titanium level power supplies

SYS-421GU-TNXR

- Perfect Platform for HPC applications
- Data Center Infrastructure

1. **Processor Support**
   Dual Socket E (LGA-4677)
   4th Gen Intel® Xeon® Scalable processors
2. **Memory Capacity**
   최대 8TB (32DIMMs)
3. **GPU**
   HGX H100 4-GPU SXM5
   Multi-GPU Board
4. **PCI-E Expansion Slots**
   8 PCIe 5.0 x16 LP slot(s)
5. **I/O ports**
   1x VGA, 1x COM Header, 2x USB 3.0,
   and 1x Dedicated IPMI, 2x 10 GBE LAN
6. **Power Supply**
   4x 3000W redundant
   Titanium level power supplies

H100 GPU 서버 – PCI-E

SYS-421GE-TNRT

- AI Compute/Model Training/ Deep Learning
- High performance simulation of complex 3D
- Omniverse / Metaverse

1. **Processor Support**
   Dual Socket E (LGA-4677)
   4th Gen Intel® Xeon® Scalable processors
2. **Memory Capacity**
   최대 8TB (32DIMMs)
3. **GPU**
   10x GPU-NVH100-80,
   GPU-NVA100-80-NC
4. **PCI-E Expansion Slots**
   13 PCIe Gen 5.0 X16 FHFL
   Slots AIOM/OCP 3.0 Support
5. **I/O ports**
   1x VGA, 2x USB 3.0,
   and 2x RJ45 10GbE
   1x RJ45 1GbE IPMI
6. **Drive bays**
   24x 2.5" hot-swap
   NVMe/SATA/SAS drive bays
   2x M.2 NVMe
7. **Power Supply**
   4x 2700W Redundant Power
   Supplies, Titanium Level

SYS-521GE-TNRT

- AI Compute/Model Training/ Deep Learning
- High performance simulation of complex 3D
- Omniverse / Metaverse

1. **Processor Support**
   Dual Socket E (LGA-4677)
   4th Gen Intel® Xeon® Scalable processors
2. **Memory Capacity**
   최대 8TB (32DIMMs)
3. **GPU**
   10x GPU-NVH100-80,
   GPU-NVA100-80-NC
4. **PCI-E Expansion Slots**
   13 PCIe Gen 5.0 X16 FHFL
   Slots AIOM/OCP 3.0 Support
5. **I/O ports**
   1x VGA, 2x USB 3.0,
   and 2x RJ45 10GbE
   1x RJ45 1GbE IPMI
6. **Drive bays**
   24x 2.5" hot-swap
   NVMe/SATA/SAS
   drive bays2x M.2 NVMe
7. **Power Supply**
   4x 2700W Redundant Power
   Supplies, Titanium Level

AS -4125GS-TNRT

- High Performance Computing
- AI/Deep Learning Training

1. **Processor Support**
   Dual Socket SP5
   AMD EPYC™
   9004 Series Processors
2. **Memory Capacity**
   최대 6TB (24DIMMs)
3. **GPU**
   Up to 8 Double-Width/
   Single-Width Cards
   (Full Height Full Length)
4. **PCI-E Expansion Slots**
   9 PCIe Gen 5.0 X16 FHFL
   Slots AIOM/OCP 3.0 Support
5. **I/O ports**
   1x VGA, 2x USB 3.0,
   and 1x Dedicated IPMI
6. **Drive bays**
   24x 2.5" hot-swap NVMe/
   SATA/SAS drive bays
   1x M.2 NVMe or M.2 SATA3
7. **Power Supply**
   4x 2200W Redundant Power
   Supplies, Titanium Level

SYS-741GE-TNRT

- AI Compute/Model Training/ Deep Learning, HPCReal time high quality multi GPU ray tracing
- High performance simulation of complex 3D

1. **Processor Support**
   Dual Socket E (LGA-4677)
   4th Gen Intel® Xeon® Scalable processors
2. **Memory Capacity**
   최대 4TB (16DIMMs)
3. **GPU**
   4x GPU-NVH100-80,
   GPU-NVA100-80-NC,
   GPU-NVA30-NC,
   GPU-NVQRTX-A6000
4. **PCI-E Expansion Slots**
   7 PCIe 5.0 x16
   FHFL slot(s)
5. **I/O ports**
   1x VGA port, 7x USB 3.0
   ports,Dual 10GbE ports,
   1 BMC LAN port
6. **Drive bays**
   8x 3.5" hot-swap
   NVMe/SATA/SAS
   drive bays 2x M.2 NVMe
7. **Power Supply**
   2x 2000W Redundant Power
   Supplies, Titanium Level

A100 GPU 서버 – HGX

AS - 4124GO-NART+

- High Performance Computing(HPC)
- AI/Deep Learning

❶ **Processor Support**	Dual AMD EPYC™ 7002, 7003 Series
❷ **Memory Capacity**	32 DIMM slots, up to 8TB DDR4 memory 3200 MHz DIMMs
❸ **GPU**	Supports 8x A100 80GB SXM4 GPUs with NVLink
❹ **PCI-E Expansion Slots**	8 PCIe 4.0 x16 LP, 1 PCIe 4.0 x16 AIOM
❺ **I/O ports**	RJ45 1GbE for IPMI, AIOM for selectable network options
❻ **Drive bays**	6x 2.5" drive bays
❼ **Power Supply**	4 3000W Redundant Power Supplies Titanium Level (96%+)

AS -2124GQ-NART+

- AI/ML, Deep Learning Training and Inference
- High Performance Computing

❶ **Processor Support**	Dual AMD EPYC™ 7002, 7003 Series Processors
❷ **Memory Capacity**	32 DIMM slots, up to 8TB DDR4 memory 3200 MHz DIMMs
❸ **GPU**	Supports 4x A100 80GB SXM4 GPUs with NVLink
❹ **PCI-E Expansion Slots**	4 PCIe 4.0 x16 LP, 1 PCIe 4.0 x8 LP
❺ **I/O ports**	Dual RJ45 10GbE LAN, RJ45 1GbE IPMI
❻ **Drive bays**	4x 2.5" drive bays
❼ **Power Supply**	2 3000W Redundant Power Supplies Titanium Level (96%+)

A100 GPU 서버 – PCI-E

AS -4124GS-TNR

- AI/Deep Learning
- High Performance Computing(HPC)

❶ **Processor Support**
Dual AMD EPYC™ 7002, 7003 Series Processors

❷ **Memory Capacity**
32 DIMM slots, up to 8TB DDR4 memory 3200 MHz DIMMs

❸ **GPU**
Supports up to 10 A100 PCIe GPUs(Default with up to 8 GPUs)

❹ **PCI-E Expansion Slots**
9 PCIe 4.0 x16(Option: 10 PCIe 4.0 x16 slots without NVMe devices)

❺ **I/O ports**
1 BMC LAN port, 1VGA port, 2 USB 3.0 ports

❻ **Drive bays**
24x 2.5" drive bays (Default with 4 SATA and 4 NVMe drives)

❼ **Power Supply**
4 2000W Redundant Power Supplies Titanium Level (96%)

SYS-220GP-TNR

- Scientific Virtualization
- High Performance Computing(HPC)
- VDI
- AI/Deep Learning Training

❶ **Processor Support**
Dual 3rd Gen Intel® Xeon® Scalable Processors upto 270W TDP

❷ **Memory Capacity**
16 DIMM slots, up to 4TB DDR4 memory 3200 MHz DIMMs

❸ **GPU**
6 PCIe GPUs Double Width FHFL

❹ **PCI-E Expansion Slots**
6 PCIe 4.0 x16 FHFL, 2 PCIe 4.0 x8 LP, AIOM support

❺ **I/O ports**
1 BMC LAN port,1VGA port, 2 USB 3.0 ports

❻ **Drive bays**
10x 2.5" drive bays, Up to 6 NVMe drives, 2x M.2

❼ **Power Supply**
Two 2600W High-efficiency (Titanium level) power supplies

SYS-740GP-TNRT

- Scientific Virtualization
- High Performance Computing
- Rendering
- AI/Deep Learning Training

❶ **Processor Support**
Dual 3 rd Gen Intel® Xeon® Scalable Processors up to 270W TDP

❷ **Memory Capacity**
16 DIMM slots, up to 4TB DDR4 memory 3200 MHz DIMMs

❸ **GPU**
4 PCIe GPUs Double Width FHFL

❹ **PCI-E Expansion Slots**
6 PCIe 4.0 x16 (4 FHFL & 2LP)
1 PCIe 4.0 x8 LP

❺ **I/O ports**
Dual 10GbE ports • 1 BMC LAN port
1 VGA port, 6 USB 3.0 ports

❻ **Drive bays**
8 Hot swap 3.5" drive bays Up to 8 NVMe drives (4 NVMe drives supported by default)

❼ **Power Supply**
2 2000W High efficiency (Titanium level power supply)

NVIDIA DGX H100

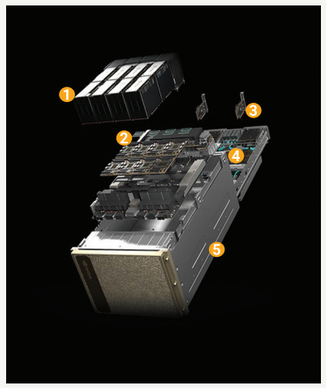
❶ **최대 640GB GPU 메모리 H100 GPU 8개**
900GB/s의 GPU 간 양방향 대역폭

❷ **NVIDIA NVSwitch 4개**
초당 7.2 테라바이트의 양방향 GPU 대역폭 1.5배 향상

❸ **NVIDIA CONNECTX-7 8개 및 NVIDIA BLUFIELD DPU**
400Gb/s 네트워크

❹ **듀얼 56 코어 4th Gen Intel Xeon CPU 및 2TB 시스템 메모리**

❺ **30 TB NVMe SSD**
최고의 성능을 위한 고속 스토리지



NVIDIA DGX A100

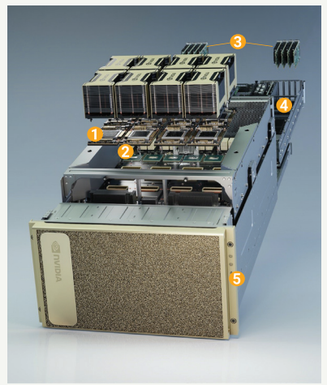
❶ **총 640GB의 GPU 메모리를 탑재한 NVIDIA A100 GPU 8개**
GPU당 NVLink 12개, GPU 간 양방향 대역폭 600GB/s

❷ **NVIDIA NVSwitch 6개**
양방향 대역폭 4.8TB/s, 이전 세대 NVSWITCH보다 2배 증가

❸ **NVIDIA CONNECTX-7 200GB/s 네트워크 인터페이스 10개**
양방향 대역폭 최대 500GB/s

❹ **듀얼 64코어 AMD CPU 및 2TB 시스템 메모리**
3.2배 더 많은 코어로 가장 집약적인 AI 작업 처리

❺ **30TB GEN4 NVMe SSD**
최대 50GB/s의 대역폭 Gen3 NVME SSD보다 2배 빠른 속도



2023년 2월
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