[제 18회] 2023년 전망, 금융IT Innovation 컨퍼런스

AI시대, 23년 금융 플랫폼 혁신 방향

Add **AI** to your **Business**.

효성인포메이션시스템

김형섭 컨설턴트

시작된 AI 비즈니스 시대

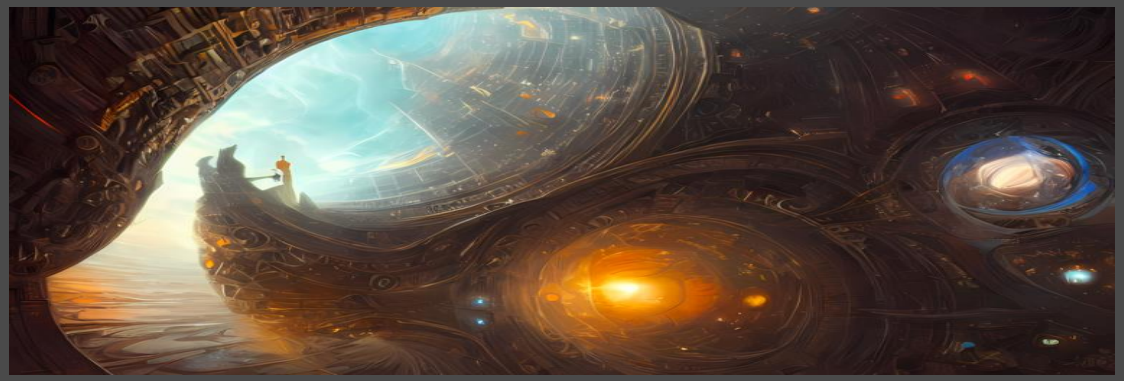
HYOSUNG
효성인포메이션시스템(주)

휴머노이드 로봇 (T사 3년내 대량생산, Auto Self Service)

완전자율주행 FSD (16만명 사용 중, 훈련 모델 7만개)

핀테크 (대출 심사 및 리스크 탐지, '30년 40조원 시장성장)

예술 창작 (미술대회 우승, 스테이블디퓨전 오픈소스)



기업의 AI 운영전략과 미래비전

HYOSUNG
효성인포메이션시스템(주)

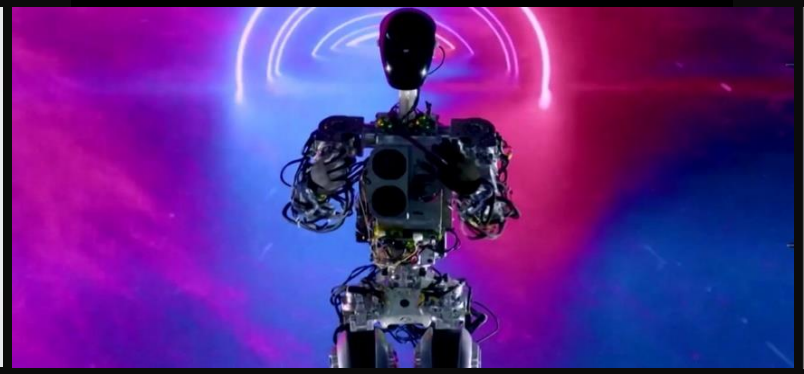
가장 많이 전기자동차를 판매하는 기업

300만대 판매, FSD 16만명 구독
쌓이는 주행 데이터

TESLA

가장 강력한 수퍼 컴퓨터를 갖춘 기업

7대 엑사 파드 건설계획
1대당 67.8 PFLOPS * 7 = 474.6 PFLOPS (FP32) 세계2위



	System	Cores	Rmax (Pflop/s)	Power (kW)
1	Frontier	8,730,112	**1,102**	21,100
	Exapod x 7	***9,912,000***	***474.6***	
2	Supercomputer Fugaku	7,630,848	**442**	29,899
3	LUMI	1,110,144	**151.9**	2,942
4	Summit	2,414,592	**148.6**	10,096
5	Sierra	1,572,480	**94.64**	7,438

AI 플랫폼 기업으로의 미래 비전

Phase 1 **자율주행**

주행거리 배터리 경쟁이 아닌,
Auto Self Driving

Phase 2 **휴머노이드**

매카닉 기술이 아닌
Auto Self Service

Phase 3 **AI 플랫폼**

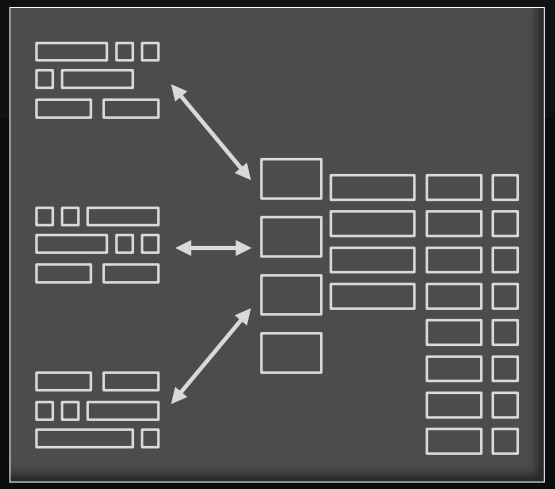
Not the Robot
but the Platform

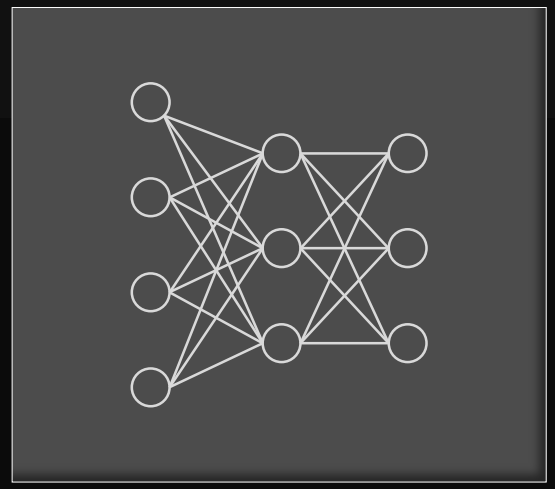
비즈니스에 AI 적용을 위한 요소

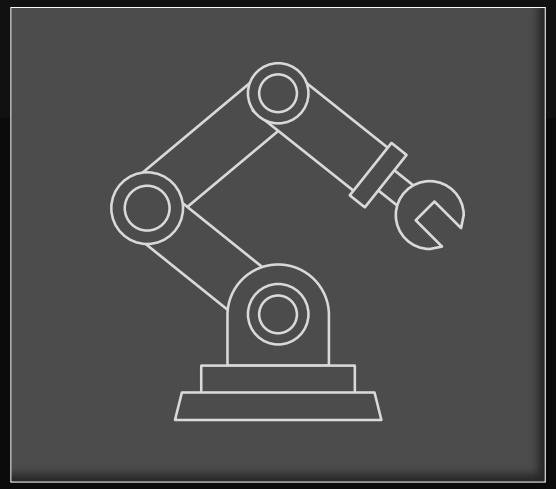
HYOSUNG
효성인포메이션시스템(주)

데이터 + AI 모델 + 연산 자원

데이터 준비 | AI모델 개발 | 학습/평가 | 활용







효성의 AI플랫폼

HYOSUNG
효성인포메이션시스템(주)

1. 데이터
운영

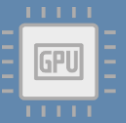
초고성능 스토리지
HCSF

AI를 위한 성능 수준 유지와
대용량 데이터 저장 효율

2. AI 모델
서비스

모델 개발환경
Lablup Backend.AI

컨테이너 기반 쉬운 AI모델
개발 및 운영 환경



3. 연산자원
성능

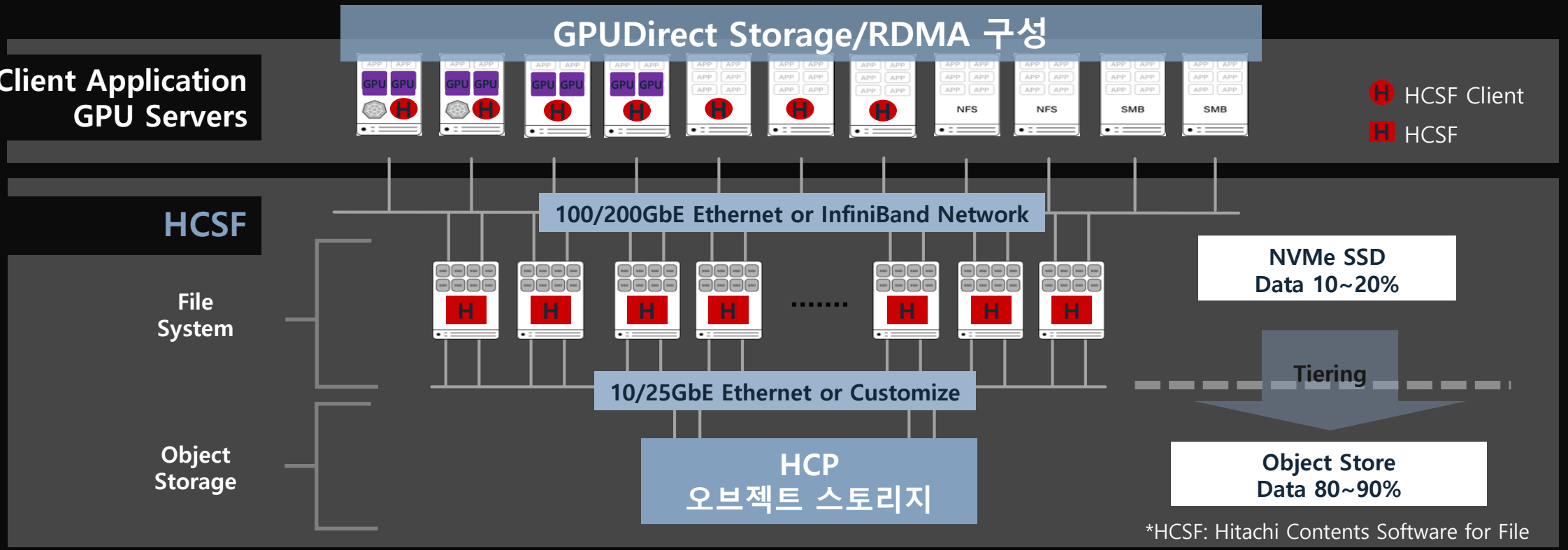
GPU시스템
Nvidia DGX & HGX

기존 전통 인프라 낮은 연산
성능 및 I/O 성능 개선

1. 데이터 운영 (HCSF)

HYOSUNG
효성인포메이션시스템(주)

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

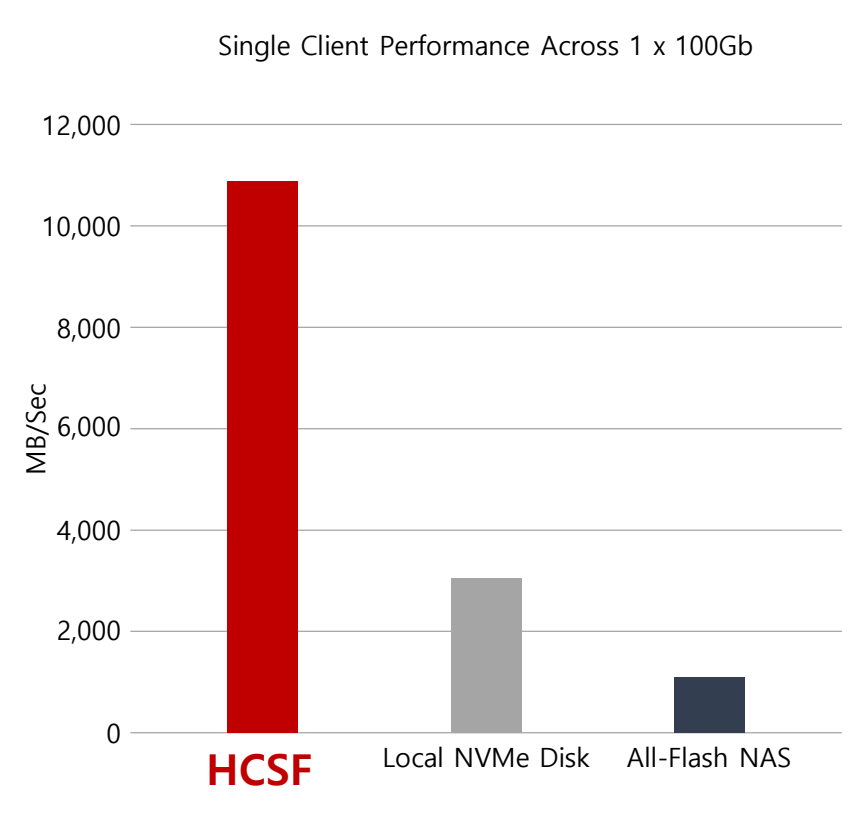


*HCSF: Hitachi Contents Software for File

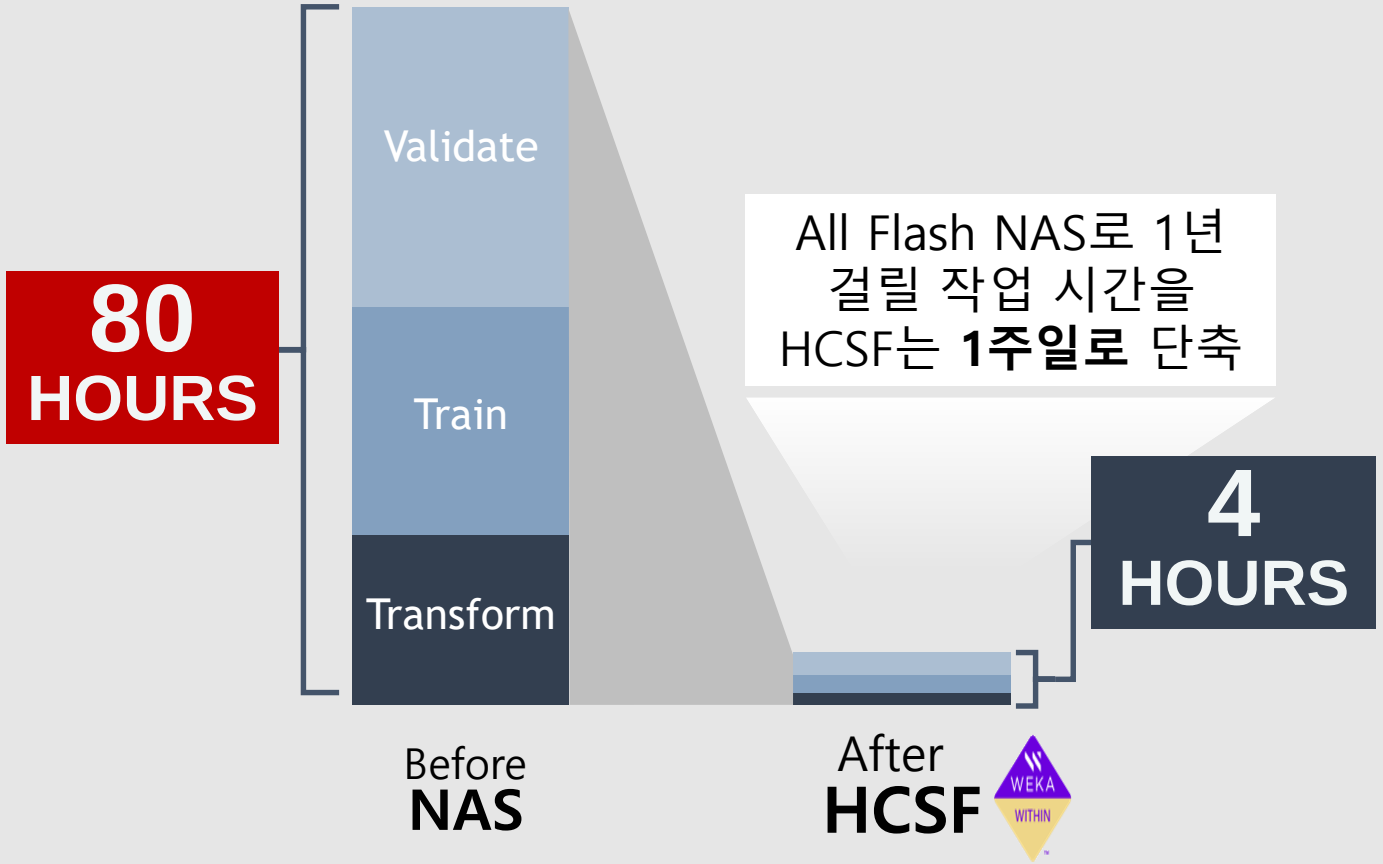
1. 데이터 운영 (HCSF)

HYOSUNG
효성인포메이션시스템(주)

저장자원 성능 비교 (단일 GPU)



Global T사 사례

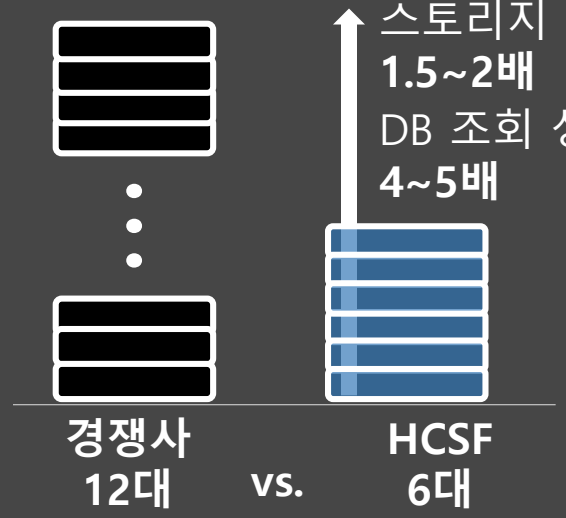


1. 데이터 운영 (HCSF)

HYOSUNG
효성인포메이션시스템(주)

BMT 성능

BMT 로 검증된 성능



GPU 환경 최적화

슈퍼/고성능 컴퓨팅에 최적화

GPU Application

AI, ML, BigData 등, 수십 PB 단위
대용량 데이터 분석에 최적화

사례

TSMC 등 국내외 다수 사례

해외

HCSF 100 노드 클러스터,
파운드리 데이터 적재 및
고속 분석 지원

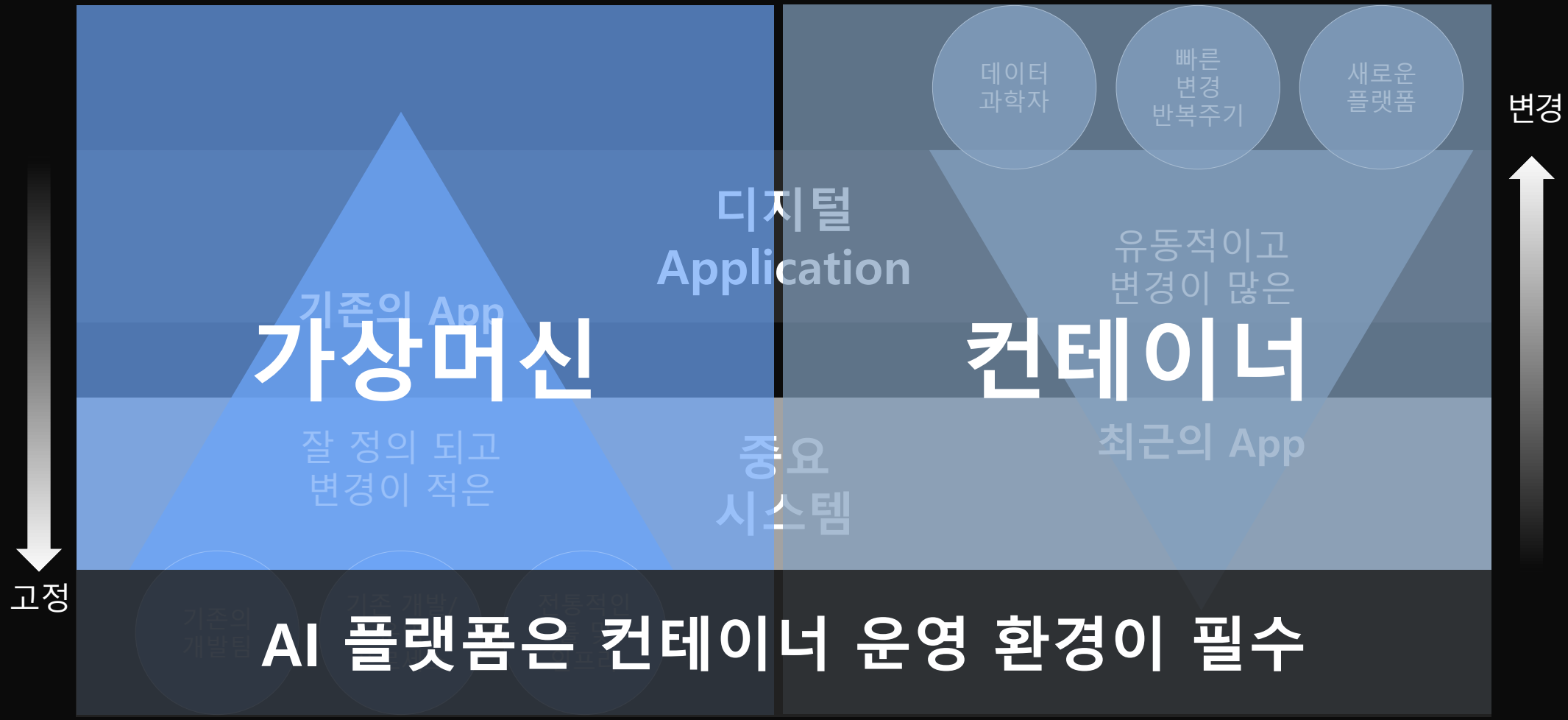
국내

- 국내 Display제조
- 국내 리서치
- 교육/개발
- AI서비스 기업

국내외 사례 및 BMT를 통해 데이터 운영 효율 검증

2. AI모델 서비스

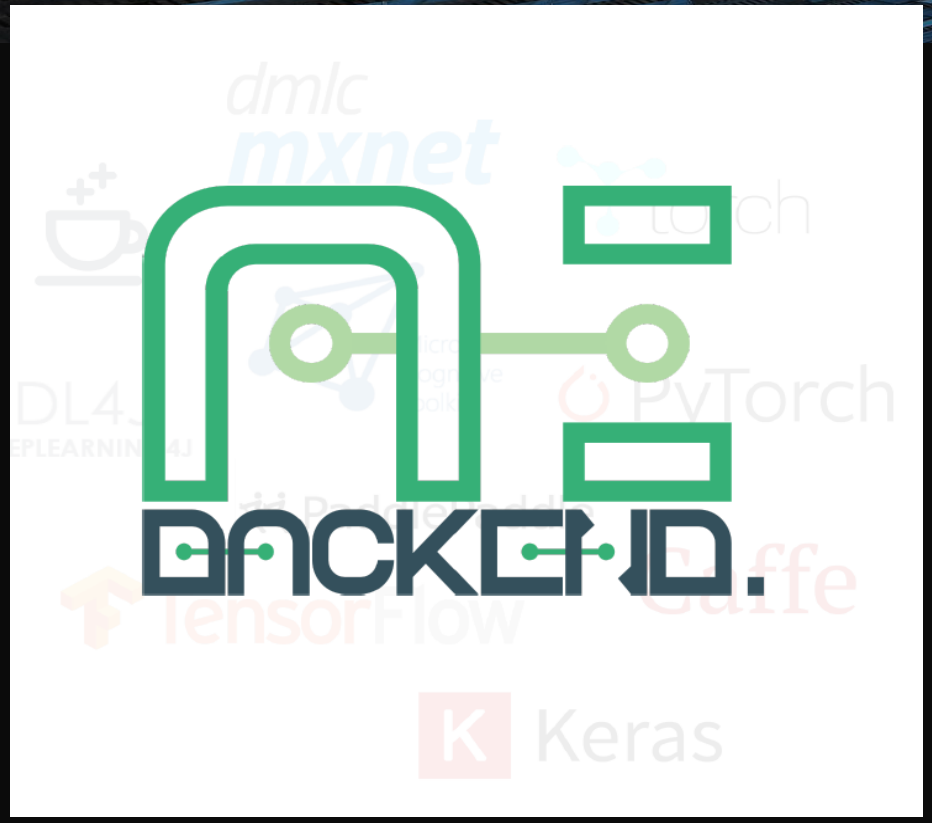
HYOSUNG
효성인포메이션시스템(주)



2. AI모델 서비스

HYOSUNG
효성인포메이션시스템(주)

Lablup Backend.AI는 아태지역 최초의 NVIDIA DGX-Ready Software 검증된 AI 플랫폼으로
국내 선도 대기업 대상 다수의 사례 보유



1 GPU 활용 극대화

✓컨테이너수준 GPU분할 가상화
✓NVIDIA GPU MIG 지원

2 직관적인 관리 및 사용자 경험

✓GUI 기반 컨테이너 운영관리
✓웹UI와 데스크탑 앱 지원

3 사전정의 AI개발환경 제공

✓Tensorflow, Pytorch 등 사전정의 이미지 제공
✓연구환경 선택 즉시 생성

4 AI 및 HPC 성능 최적화

✓독자적 엔진으로 최적의 GPU 연산 자원 배치 구현
✓다중노드 워크로드 및 데이터 I/O 병렬화 지원

2. AI모델 서비스

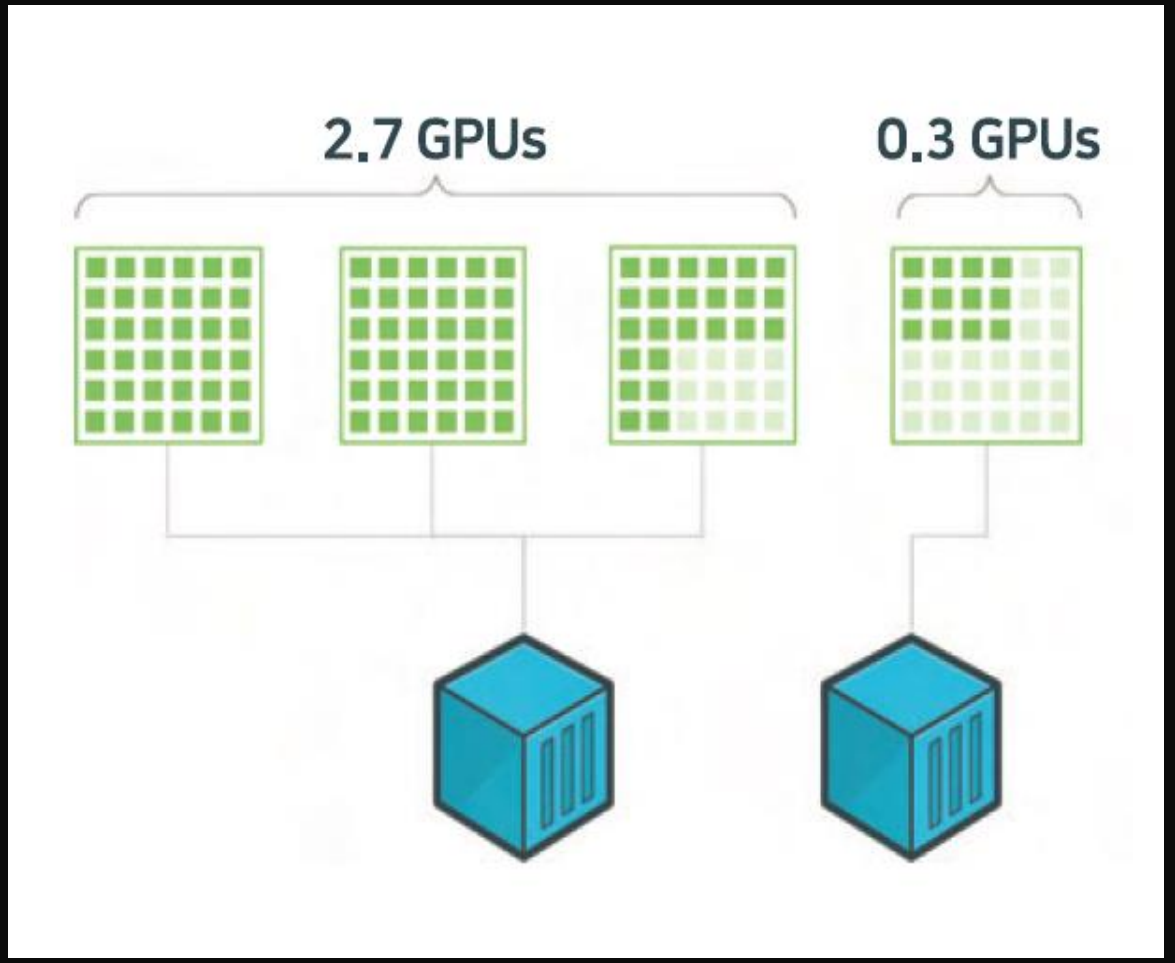
HYOSUNG
효성인포메이션시스템(주)

Lablup Backend.AI 의 GPU 분할 가상화

<u>컨테이너 기반</u> GPU 스케일링

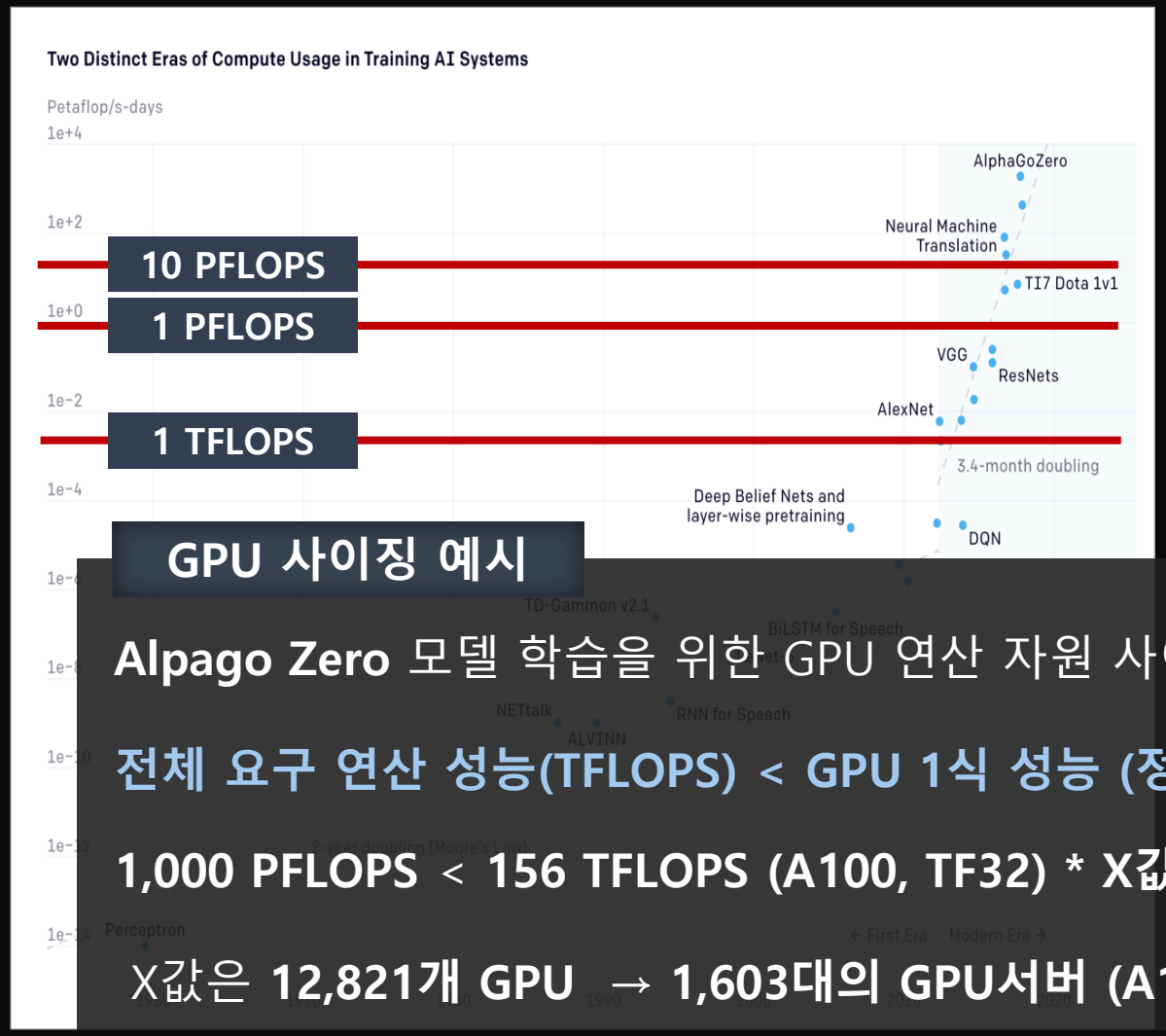
- ✓ 교육 및 추론 워크로드를 위한 단일 GPU 공유
- ✓ 모델 학습 등 대규모 워크로드를 위한 다중 GPU 할당
- ✓ 자체 개발한 CUDA 가상화 계층으로 구현

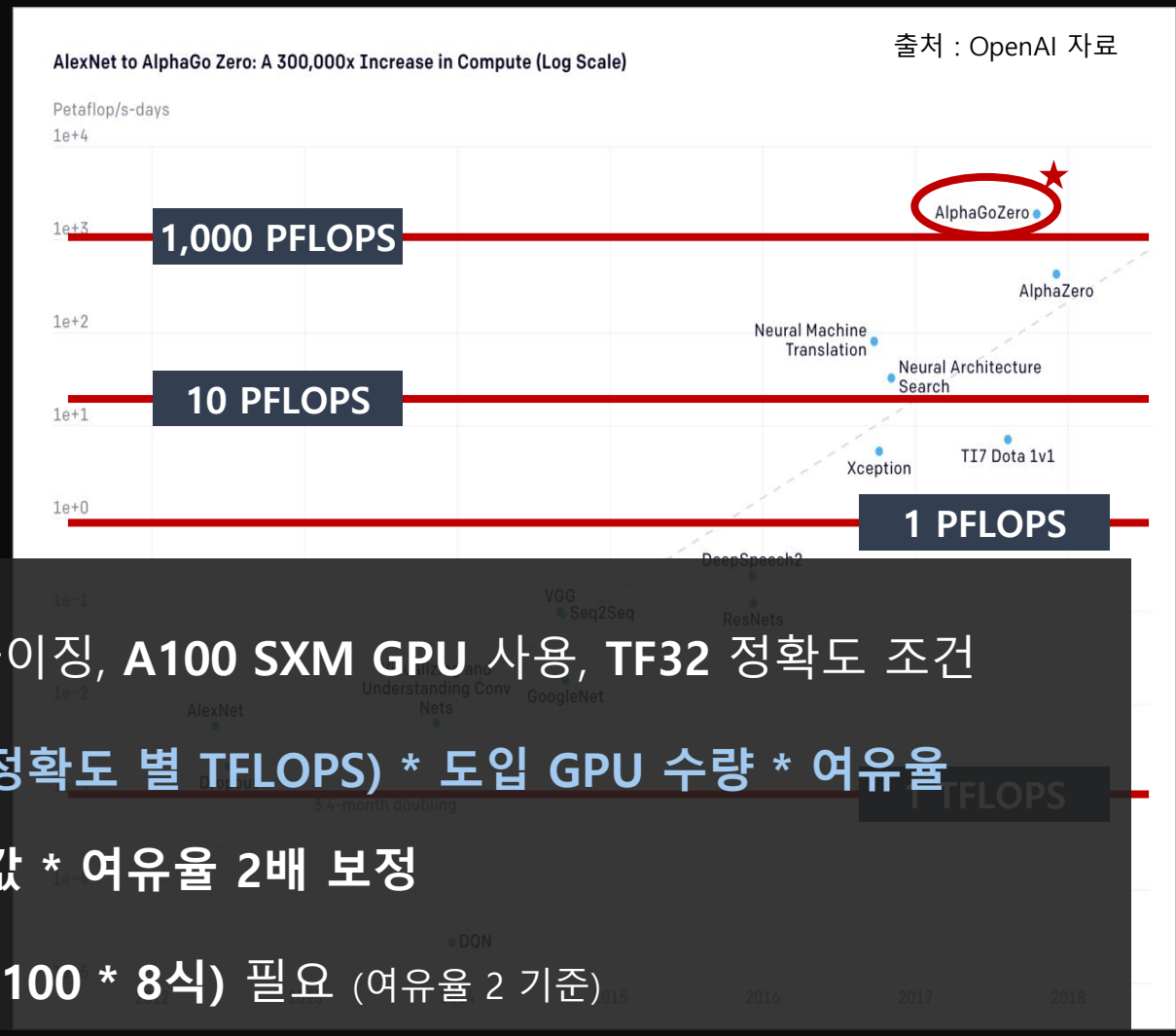
[Lablup Backend.AI 대한민국, 미국, 일본 등록 특허]



3. 연산자원 성능

HYOSUNG
효성인포메이션시스템(주)





출처 : OpenAI 자료

GPU 사이징 예시

Alpago Zero 모델 학습을 위한 GPU 연산 자원 사이징, **A100 SXM GPU** 사용, **TF32** 정확도 조건

전체 요구 연산 성능(TFLOPS) < GPU 1식 성능 (정확도 별 TFLOPS) * 도입 GPU 수량 * 여유율

1,000 PFLOPS < 156 TFLOPS (A100, TF32) * X값 * 여유율 2배 보정

X값은 **12,821개 GPU → 1,603대의 GPU서버 (A100 * 8식)** 필요 (여유율 2 기준)

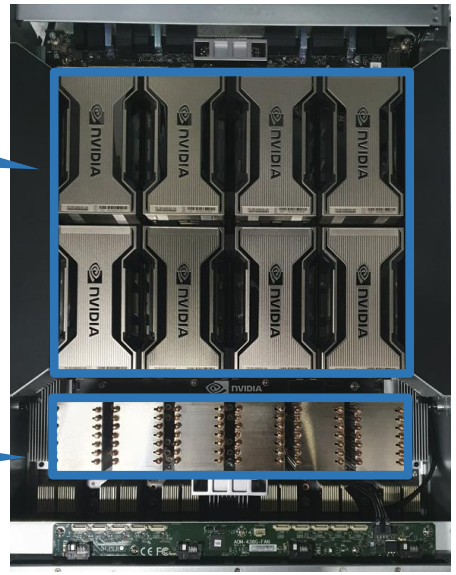
3. 연산자원 성능

HYOSUNG
효성인포메이션시스템(주)

NVIDIA DGX / HGX

8 * Nvidia HGX2 SXM4 A100 GPU

6 * NVIDIA NVSwitch



- 4세대 NVIDIA NVLink는 900GB/s GPU 대역폭으로 PCIe Gen4 대비 14배 고성능
- NVIDIA A100 Tensor 코어 GPU의 고속 상호 연결 구현

NVLink와 PCIe 비교

항목	서브 링크 속도	서브 링크 수	전체 속도	GPU 아키텍쳐
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**	A
NVLink 4.0	25GB/s	18	**900GB/s**	Hopper

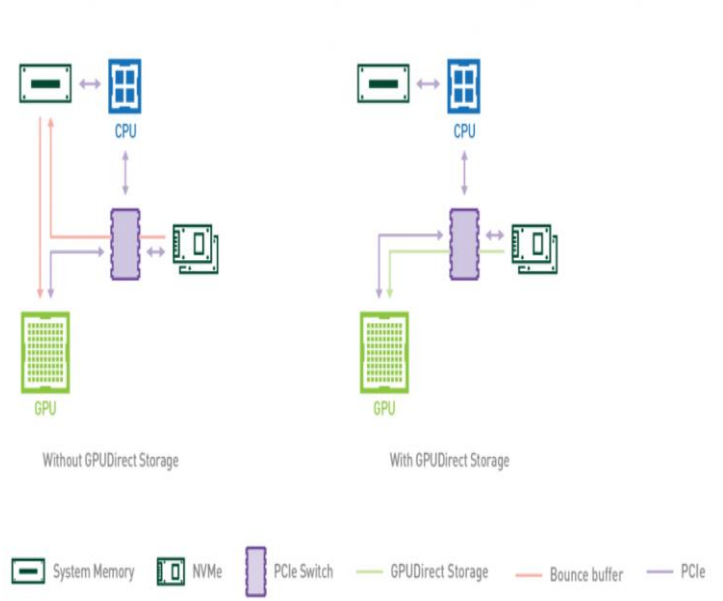
14배 고성능

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

3. 연산자원 성능 (성능 최적화)

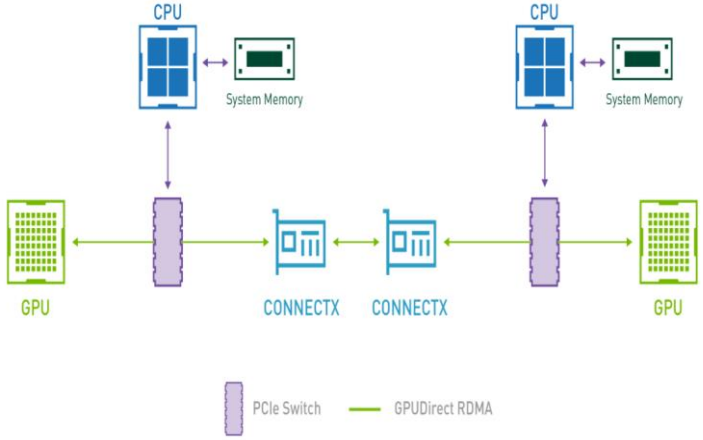
HYOSUNG
효성인포메이션시스템(주)

1. 스토리지 I/O GPUDirect Storage



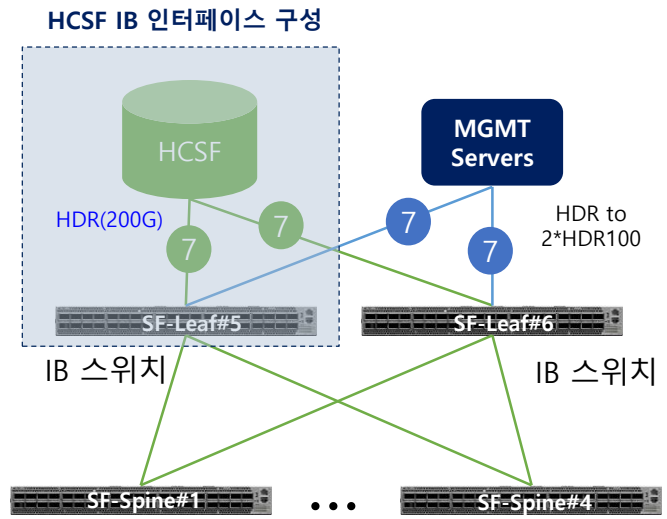
- NVMe /NVMe-oF 스토리지와 GPU 메모리 간 직접 연결
- CPU 메모리의 바운스 버퍼를 제거
- 스토리지와 GPU 메모리의 데이터 로드 프로세스 IO 개선

2. 네트워크 I/O GPUDirect RDMA



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

3. 고속 네트워크 Infiniband 구성



- 장치간의 100G/200G 고속 네트워크 구성이 GPU연산 성능 및 고성능 저장자원의 성능을 위해 필수
- IB NIC, IB 스위치의 설계 및 구성 매우 중요

3. 연산자원 성능 (성능 최적화)

HYOSUNG
효성인포메이션시스템(주)

GPUDirect Storage 성능 테스트 결과

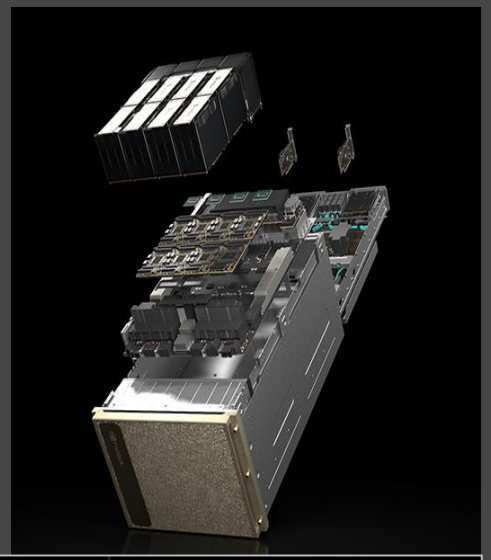
No.	성능 테스트 옵션			성능 결과		
	디스크 구분	GDS 구성	할당 GPU 수	Write Speed (GiB/s)	Read Speed (GiB/s)	비고
1	HCSF	On	1	21.66	21.09	GDS ON 구성으로 1.5배 Throughput 향상 (Read)
2	HCSF	Off	1	19.20	13.97	
3	HCSF	On	2	36.50	42.26	GDS ON 구성으로 3.1배 Throughput 향상 (Read/Write)
4	HCSF	Off	2	20.50	13.60	
5	HCSF	On	6	70.6	100.0	GDS ON 구성으로 3.0배 Throughput 향상 (Read)
6	HCSF	Off	6	33.7	33.1	
7	로컬 NVMe	On	2	3.22	6.19	변화 없음
8	로컬 NVMe	Off	2	3.77	6.12	

3. 연산자원 성능 (성능 최적화)

HYOSUNG
효성인포메이션시스템(주)

NVIDIA DGX H100

1. 최대 640GB GPU 메모리 H100 GPU 8개
   900GB/s의 GPU 간 양방향 대역폭
2. NVIDIA NVSwitch 4개
   초당 7.2 테라바이트의 양방향 GPU 대역폭, 이전 세대 대비 1.5배 향상
3. NVIDIA CONNECTX-7 8개 및 NVIDIA BLUFIELD DPU 400Gb/s 네트워크
4. 듀얼 x86 CPU 및 2TB 시스템 메모리
   초고도 AO 작업을 위한 강력한 CPU
5. 30 TB NVME SSD
   최고의 성능을 위한 고속 스토리지



NVIDIA H100			
	FP8	4,000 TFLOPS	6X
	FP16	2,000 TFLOPS	3X
	TF32	1,000 TFLOPS	3X
	FP64/FP32	60 TFLOPS	3X

슈퍼마이크로 HGX H100 (AS -8125GS-TNHR)

1. HGX H100 8-GPU SXM5 Multi GPU Board 장착
2. AMD EPYC™ 9004 Series Processors Dual Socket (Socket SP5) 지원
3. 24 DIMM slots
   최대 6TB 시스템 메모리
4. 3000W redundant Titanium 레벨 파워 서플라이 8개 장착
5. 8 PCI-E Gen 5.0 X16 LP, 2 PCI-E Gen 5.0 X16 FHFL Slots 지원

※ 효성인포메이션시스템은 슈퍼마이크로 공식 총판사로서 GPU 및 x86 서버를 공급 합니다.

AI를 도입하려는 고객 유형 및 고민

HYOSUNG
효성인포메이션시스템(주)

비즈니스 성과

1. 기존 도입 고객

- 기 도입 인프라와 AI솔루션 성능&운영 이슈
- 기 자원 재활용, 향후 체계적 도입 요건

2. 신규 도입 고객

- 사업 방향성 수립에 대한 고민
- 성과 창출 불확실성 우려

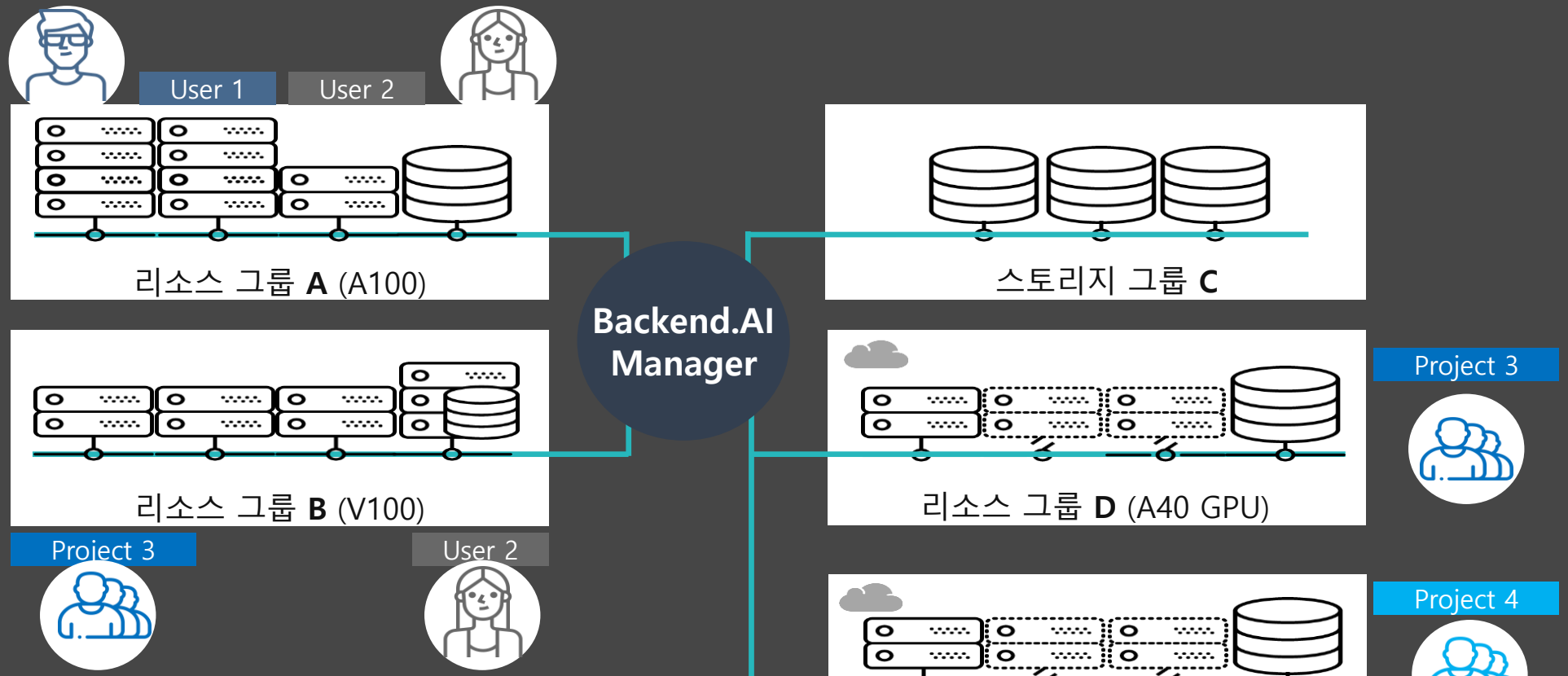
3. 대형사업추진 고객

- 대형 GPU Farm 구축을 통해, 본격적으로 AI를 기업내 적용
- 초 고성능 운영효율 중심

1. 기존 도입 고객 – 기 자원 활용 AI플랫폼 도입

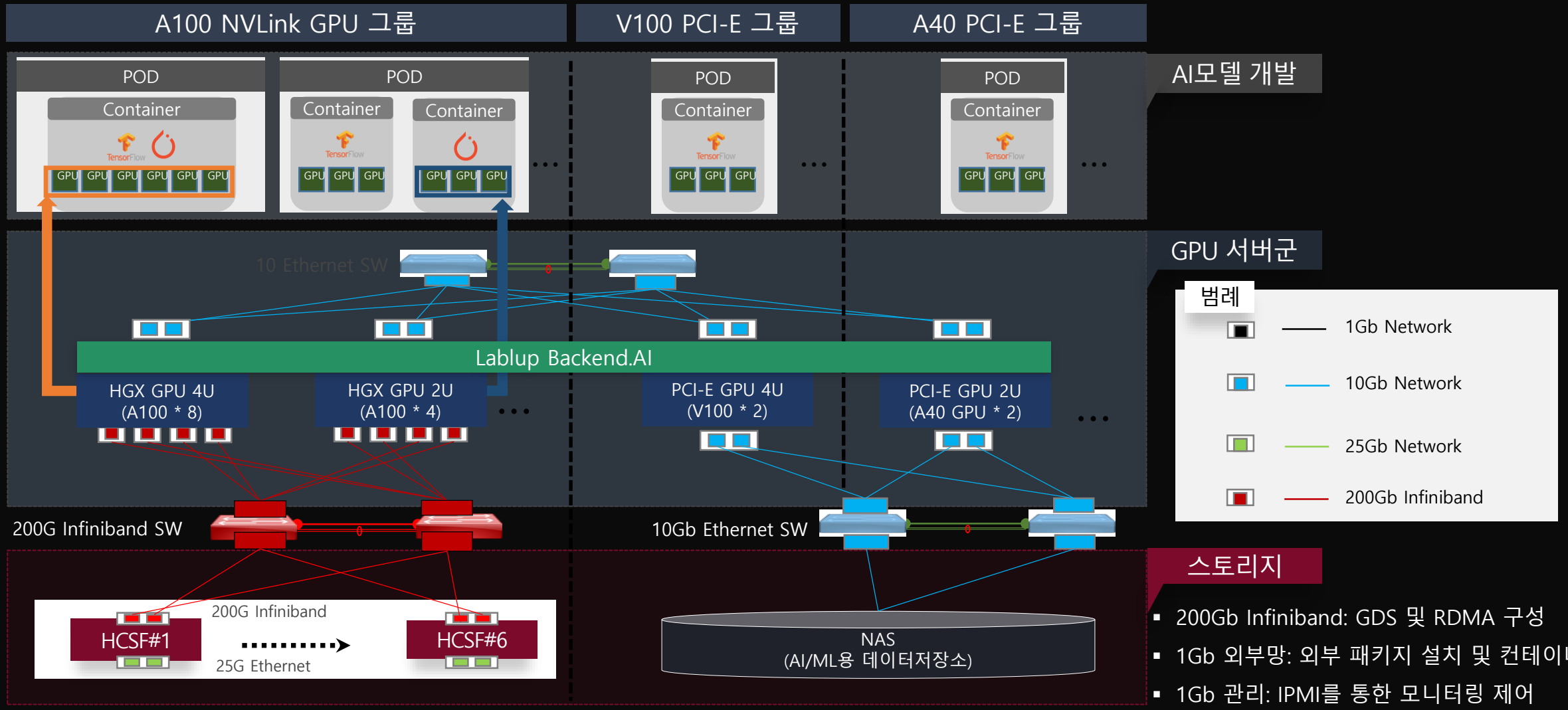
HYOSUNG
효성인포메이션시스템(주)

리소스 그룹의 응용 예



1. 기존 도입 고객 – 기 자원 활용 AI플랫폼 도입

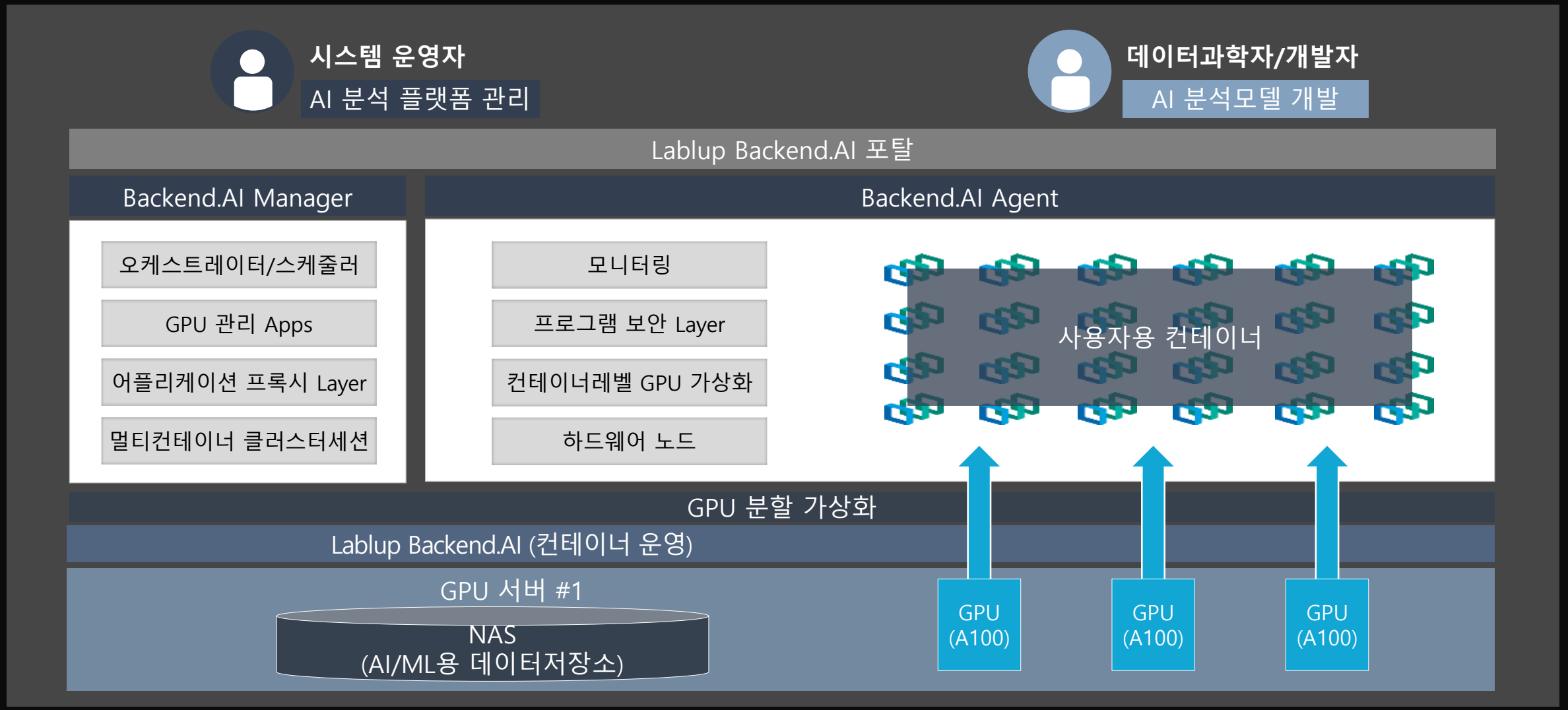
HYOSUNG
효성인포메이션시스템(주)



- 200Gb Infiniband: GDS 및 RDMA 구성
- 1Gb 외부망: 외부 패키지 설치 및 컨테이너
- 1Gb 관리: IPMI를 통한 모니터링 제어

2. 신규 도입 고객 – AI 테스트 베드 환경

HYOSUNG
효성인포메이션시스템(주)



3. 대규모 사업 추진 고객 – 효성 통합 AI솔루션

HYOSUNG
효성인포메이션시스템(주)

데이터
운영

효율적 데이터 운영을 위한 고성능 병렬 파일 스토리지 구성
(HCSF 기반 고성능 수준 유지, 오브젝트 스토리지 티어링 구성으로 비용 절감)

AI모델
서비스

AI/ML 모델 개발/운영 효율화
(AI/ML Ops 및 컨테이너 – Lablup Backend.AI)

연산자원
성능

GPU 연산 자원의 성능 최적화
(NVLink 지원 HGX GPU 서버, GPU Direct Storage/RDMA 성능 최적화)

AI를 직접 경험할 수 있는 DX센터

HYOSUNG
효성인포메이션시스템(주)

AI 업무 프로세스 체험

- Lablup Bakcend.AI 기반 AI모델 개발 환경 경험
- AI모델 학습 및 추론 시연

초고성능 스토리지 기술

- HCSF 기반 AI 학습 성능
- GPU Direct Storage 기술과 HCSF 연계 성능 최적화

GPU서버 성능

- DGX & HGX 서버 NVLink 연산성능
- Infiniband 200G 네트워크 구성

AI의 시작과 끝 효성이 함께 합니다.

HYOSUNG
효성인포메이션시스템(주)

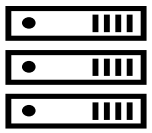
자문/컨설팅 | 계획/설계 | 구축&수행

효성 AI 플랫폼

인프라 최적화 (GPUDirect Storage, GPU가상화)

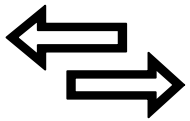
AI 운영시스템 (컨테이너: Lablup Backend.AI, 가상머신: VMware)

AI 인프라



연산 자원
(NVIDIA DGX/
Supermicro HGX)

저장 자원
(초고성능 병렬 파일
스토리지-HCSF)



네트워크
(Mellanox&Cisco)

AI의 시작과 끝 효성이 함께 합니다.

- 효성인포메이션시스템 -