Ai 개발 및 운영환경 구축을 위한
Nvidia DGX Platform System

BN I&C TEN

㈜비엔아이엔씨는 AI인프라 토탈 솔루션 공급자로서,
최고의 기술력과 경험을 바탕으로 신뢰할 수 있는 IT전문 기업입니다.

㈜비엔아이엔씨는 다양한 하드웨어 및 소프트웨어 벤더들과 파트너십을 맺고
AI인프라 관련 토탈 솔루션을 제공하는 전문 기업입니다.
세계적인 기업인 NVIDIA, HPE, Supermicro Computer, Purestorage 등의 파트너로
2017년 출범하여 고객사와 함께 성장해왔습니다.
급변하는 IT 시장에서 ㈜비엔아이엔씨는 고객과 파트너 여러분께
언제나 감동과 만족을 드리기 위해 최선의 노력을 다하겠습니다.

BN I&C TEN

Compute DGX 기술을 획득한 업체로써

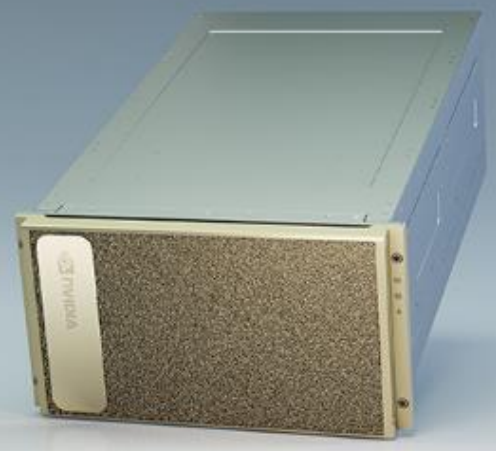
DGX SuperPOD 를 구축한 국내 유일의 회사

- **Naver SuperPOD _ 2020**
- **KT SuperPOD _ 2021**

BN I&C TEN

NVIDIA DGX A100

AI 인프라를 위한 유니버설 시스템



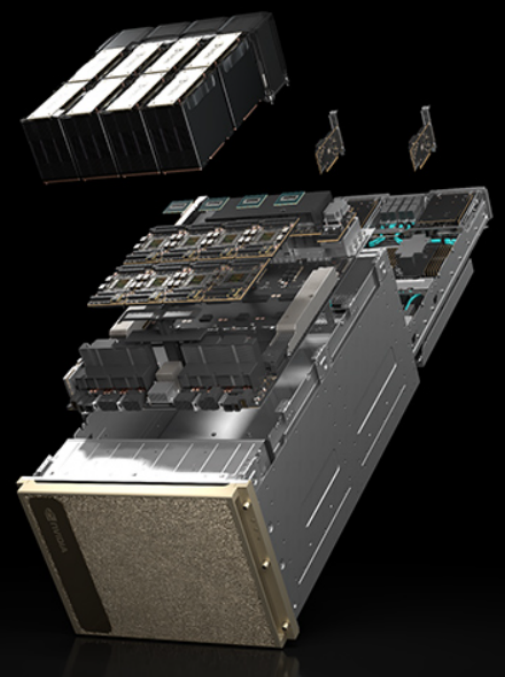
NVIDIA A100 기반의 세계 최초 AI 시스템

NVIDIA DGX™ A100 은 모든 AI 워크로드를 위한 유니버설 시스템으로, 세계 최초의 5페타플롭스 AI 시스템을 통해 유례없는 컴퓨팅 밀도, 성능, 유연성을 제공합니다. NVIDIA A100 Tensor 코어 GPU를 탑재한 DGX A100 은 기업이 NVIDIA AI 전문가의 직접적인 지원과 함께 훈련에서 추론, 분석에 이르기까지 배포하기 쉬운 통합 AI 인프라를 구축할 수 있게 합니다.

BN I&C TEN

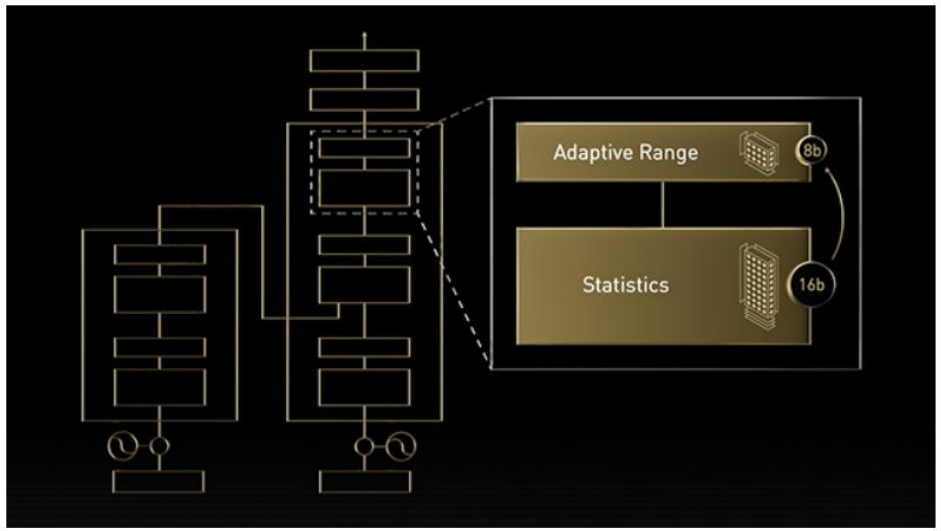
DGX H100 둘러보기

- 최대 640GB의 총 GPU 메모리를 탑재한 NVIDIA H100 GPU 8개
  GPU당 NVIDIA® NVLink® 12개, 900GB/s의 GPU 간 양방향 대역폭
- NVIDIA NVSWITCH™ 4개
  초당 7.2 테라바이트의 양방향 GPU 간 대역폭으로 이전 세대대비1.5배 향상
- NVIDIA CONNECTX®-7 8개 및 NVIDIA BLUEFIELD® DPU 400Gb/s 네트워크 인터페이스 2개
  1TB/s의 최대 양방향 네트워크 대역폭
- 듀얼 x86 CPU 및 2TB 시스템 메모리
  초고도 AI 작업을 위한 강력한 CPU
- 30TB NVME SSD
  최고의 성능을 위한 고속 스토리지



기술 혁신

최첨단 TSMC 4N 프로세스를 사용하여 800억 개 이상의 트랜지스터로 구축된 Hopper는 NVIDIA H100 Tensor 코어 GPU를 지원하는 획기적인 5가지 혁신을 제공합니다.



트랜스포머 엔진

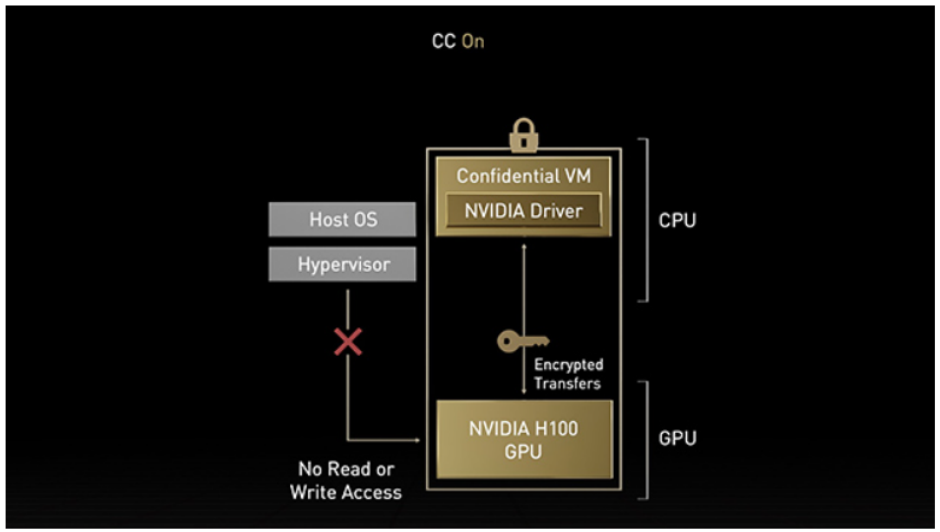
NVIDIA Hopper 아키텍처는 AI 모델의 훈련을 가속하도록 설계된 트랜스포머 엔진을 통해 Tensor 코어 기술을 발전시킵니다. Hopper Tensor 코어는 FP8과 FP16 정밀도를 혼합 적용하여 트랜스포머의 AI 계산을 획기적으로 가속할 수 있습니다. 또한 Hopper에서는 TF32, FP64, FP16, INT8 정밀도의 부동 소수점 연산(FLOPS)을 이전 세대의 3배로 늘렸습니다. 트랜스포머 엔진 및 4세대 NVIDIA® NVLink®와 결합된 Hopper Tensor 코어는 HPC 및 AI 워크로드에 최고 수준의 속도 향상을 제공합니다.

NVLink 스위치 시스템

비즈니스의 속도로 움직이려면 엑사스케일 HPC 및 매개 변수가 조 단위인 AI 모델이 규모별로 가속할 수 있도록 서버 클러스터의 모든 GPU 간에 빠르고 원활한 통신이 필요합니다.

4세대 NVLink는 스케일업 인터커넥트입니다. NVLink 스위치 시스템은 새로운 외부 NVLink 스위치와 결합하여 GPU 당 양방향 900GB/s 속도로 서버 간의 멀티 GPU 입력/출력(IO)을 가능하게 합니다. 이는 PCIe Gen5의 7배 이상의 대역폭이기도 합니다. NVLink 스위치 시스템은 최대 256개의 접속 H100 클러스터를 지원하며 Ampere의 InfiniBand HDR보다 9배 높은 대역폭을 제공합니다.

또한 NVLink는 이전에는 Infiniband에서만 지원되었던 네트워크 내 컴퓨팅인 SHARP를 지원하여 57.6TB/s의 All2All 대역폭을 제공하면서 FP8 희소성 AI 컴퓨팅의 1 엑사플롭(exaFLOP)이라는 놀라운 성능을 제공할 수 있게 되었습니다.



NVIDIA 컨피덴셜 컴퓨팅

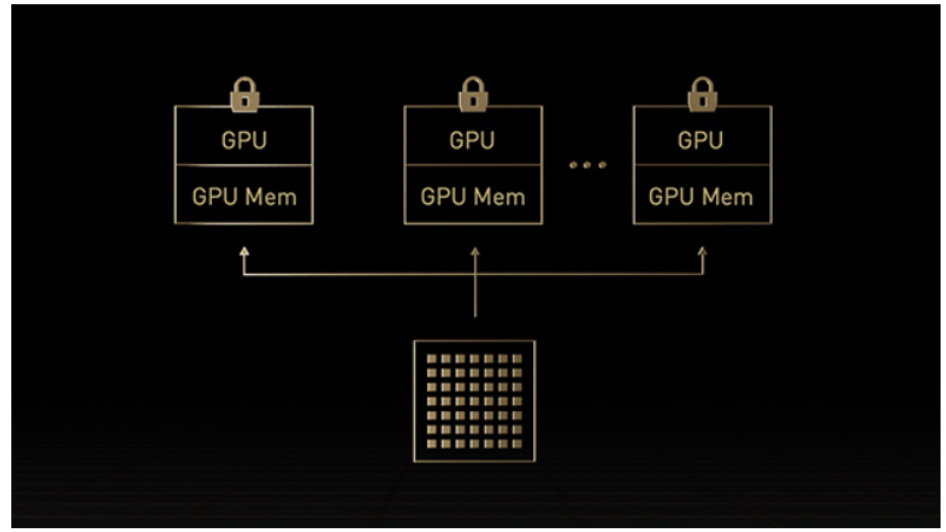
데이터는 저장 공간에 저장되고 네트워크를 통해 전송되는 동안에는 암호화되지만, 처리되는 동안에는 보호되지 않습니다. NVIDIA 컨피덴셜 컴퓨팅은 사용 중인 데이터와 애플리케이션을 보호하여 이러한 격차를 해소합니다. NVIDIA Hopper 아키텍처는 세계 최초로 NVIDIA 컨피덴셜 컴퓨팅 기능을 탑재한 가속 컴퓨팅 플랫폼을 도입했습니다.

강력한 하드웨어 기반 보안을 통해 사용자는 온프레미스, 클라우드, 엣지에서 애플리케이션을 실행할 수 있으며, 사용 중인 애플리케이션 코드와 데이터를 허가되지 않은 엔티티가 보거나 수정할 수 없다고 확신할 수 있습니다. 이를 통해 데이터와 애플리케이션의 기밀성과 무결성을 보호하는 동시에 AI 훈련, AI 추론, HPC 워크로드를 위한 H100 GPU의 전례 없는 가속에 액세스할 수 있습니다.

2세대 MIG

MIG(Multi-Instance GPU)를 사용하면 GPU를 자체 메모리, 캐시, 컴퓨팅 코어가 있는 더 작고 완전히 격리된 여러 인스턴스로 분할할 수 있습니다. Hopper 아키텍처는 최대 7개의 GPU 인스턴스에 걸쳐 가상화된 환경에서 멀티 테넌트 및 멀티 사용자 구성을 지원하여 MIG를 더욱 향상하고, 하드웨어 및 하이퍼바이저 수준에서 기밀 컴퓨팅으로 각 인스턴스를 안전하게 격리합니다. 각 MIG 인스턴스에 대한 전용 비디오 디코더는 공유 인프라에서 안전하고 처리량이 높은 지능형 영상 분석(IVA)을 제공합니다. 또한 관리자는 Hopper의 동시 MIG 프로파일링을 통해 적합한 크기의 GPU 가속을 모니터링하고 사용자를 위한 리소스 할당을 최적화할 수 있습니다.

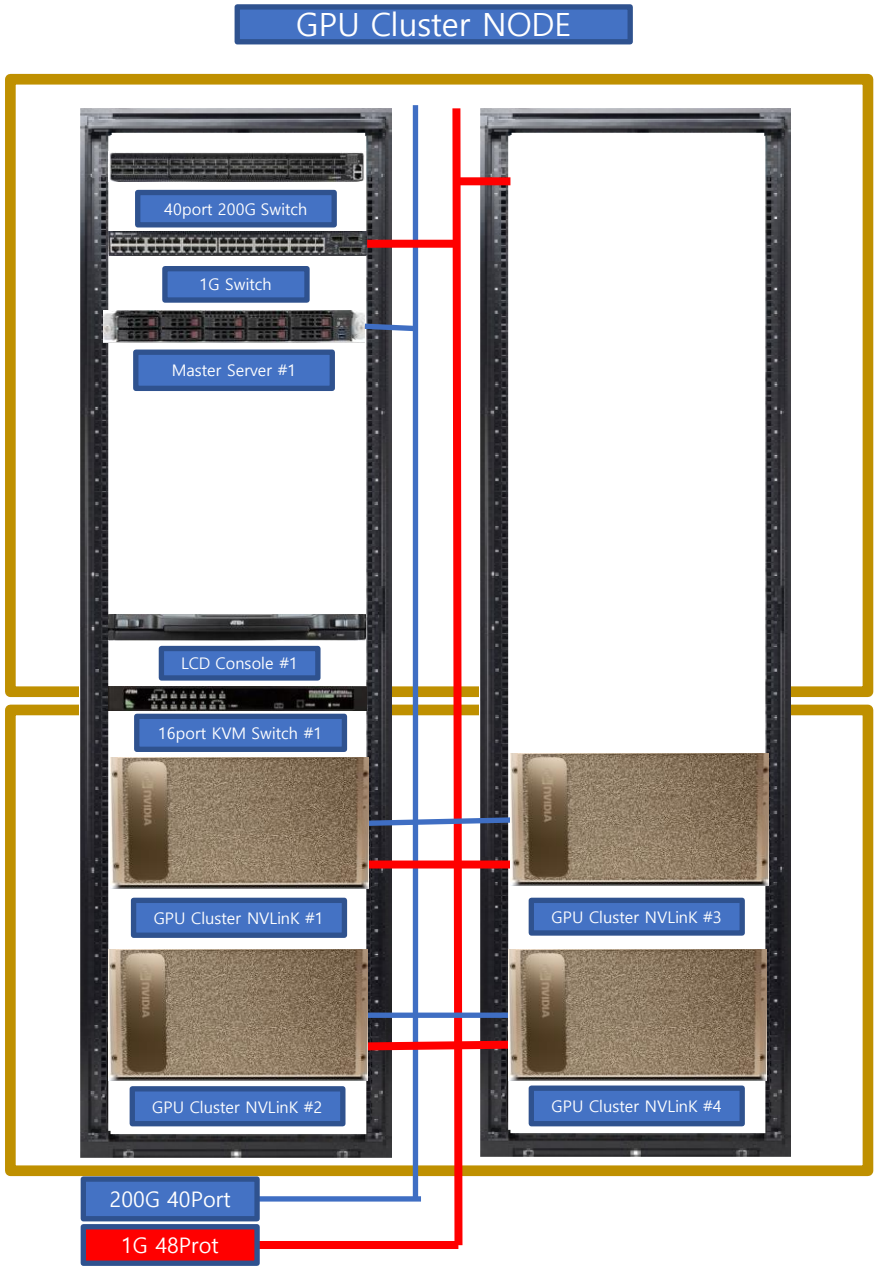
워크로드가 적은 연구원의 경우 전체 CSP 인스턴스를 대여하기보다는 MIG를 사용하여 GPU의 일부를 안전하게 격리하는 동시에 데이터를 저장, 전송, 컴퓨팅 시 안전하게 보호할 수 있습니다.



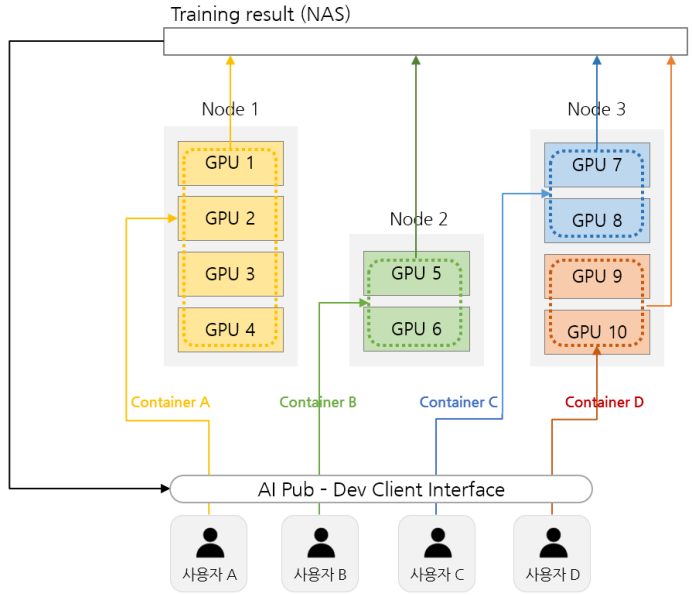
DPX 명령

동적 프로그래밍은 복잡한 재귀 문제를 더 단순한 하위 문제로 세분화하여 해결하는 알고리즘 기술입니다. 나중에 다시 계산할 필요가 없도록 하위 문제의 결과를 저장함으로써 기하급수적인 문제 해결의 시간과 복잡성을 줄입니다. 동적 프로그래밍은 일반적으로 광범위한 사용 사례에서 사용됩니다. 예를 들어 플로이드-워셜은 운송 및 배송 차량의 최단 경로를 매핑하는 데 사용할 수 있는 경로 최적화 알고리즘입니다. 스미스-워터맨 알고리즘은 DNA 서열 정렬 및 단백질 접힘 애플리케이션에 사용됩니다.

Hopper는 동적 프로그래밍 알고리즘을 CPU 대비 40배, NVIDIA Ampere 아키텍처 GPU에 비해 7배 가속하는 DPX 명령을 도입했습니다. 이를 통해 질병 진단, 실시간 라우팅 최적화, 그래프 분석 시간을 획기적으로 단축할 수 있습니다.



GPU 인프라를 병합하여 관리하므로, 사용자가 원하는 GPU 개수만큼 컨테이너 환경을 할당 받아 인공지능 학습을 할 수 있습니다.



AI 개발 및 운영 환경 구축을 위한
MLOps 솔루션, AI Pub

TEN

주식회사 텐

Better the world with AI

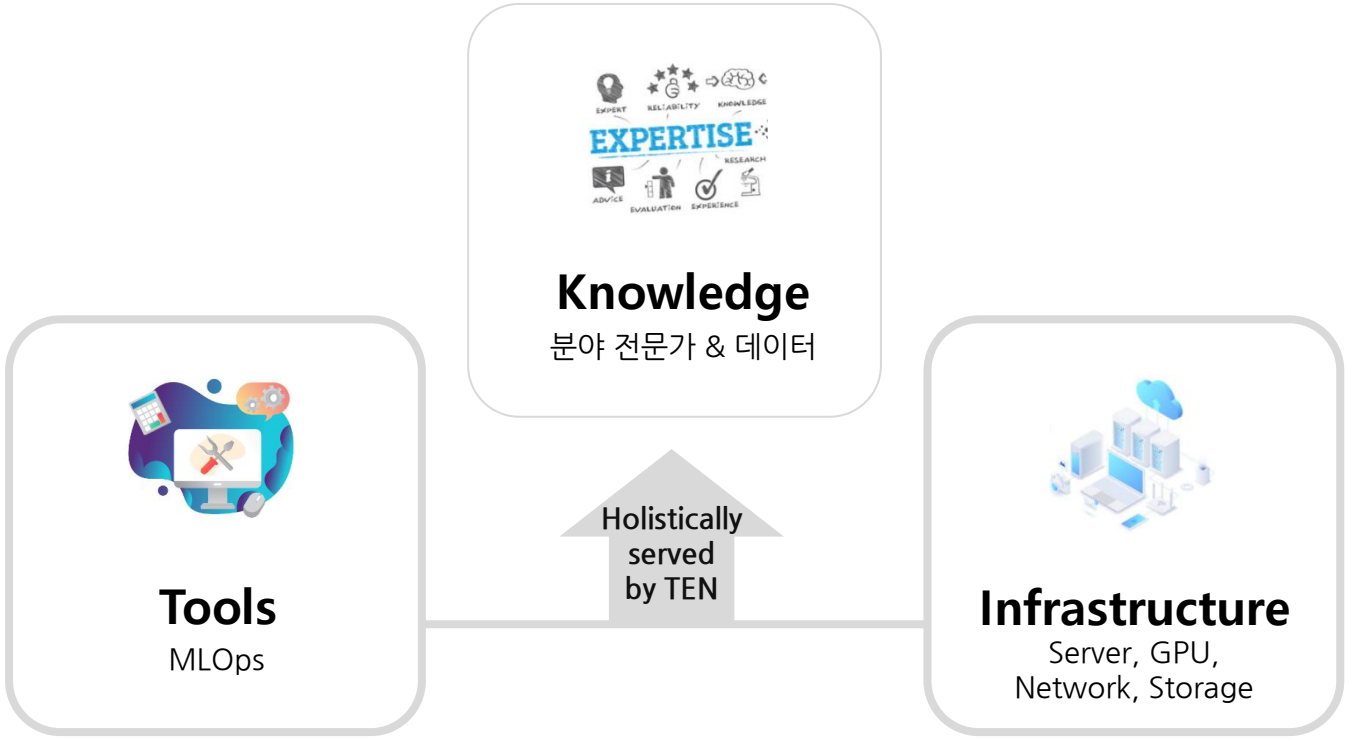
‘세상을 널리 AI롭게 하자’

/

AI를 통해 더 많은 기업과 개인이 가치를 창출하고, 공유하는 서비스를 만드는 기업

Our Goal

TEN은 AI를 잘 개발하고 운영할 수 있도록 솔루션(MLOps와 인프라)을 서비스하는 것에 주력



- 해당 분야의 지식과 노하우를 축적한 전문가들만이 비즈니스 맞춤형 인공지능을 더 잘 만들 수 있음
- 인공지능을 비즈니스에 도입하기 위해 공통으로 사용하는 도구와 인프라가 중요한 시장이 도래할 것이라는 확신

→ 이를 위해 TEN은 인공지능 개발과 운영을 위한 도구-MLOps 와 인프라를 서비스하는 기업으로 성장 중

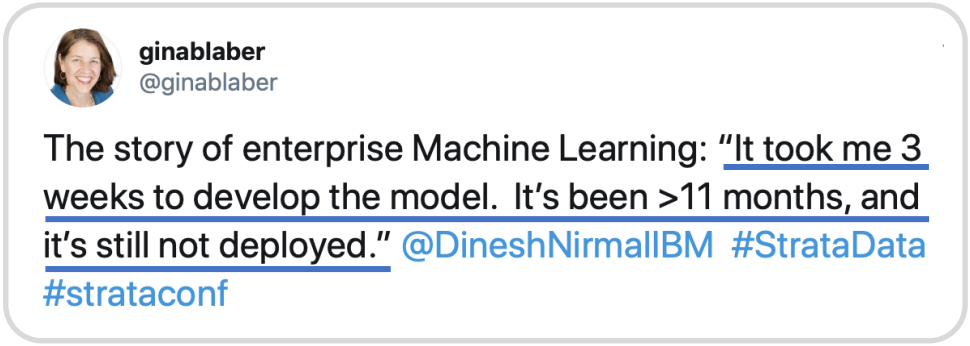
TEN

Challenges in AI Field

TEN

AI 모델 개발 완료 후 운영 배포까지 해결해야 할 기술적 과제로 인해 평균 8.6개월의 시간 소요

(그 조차도 운영 배포까지 도달하는 AI 프로젝트 비율이 50퍼센트 미만으로 조사됨)



8.6 Months

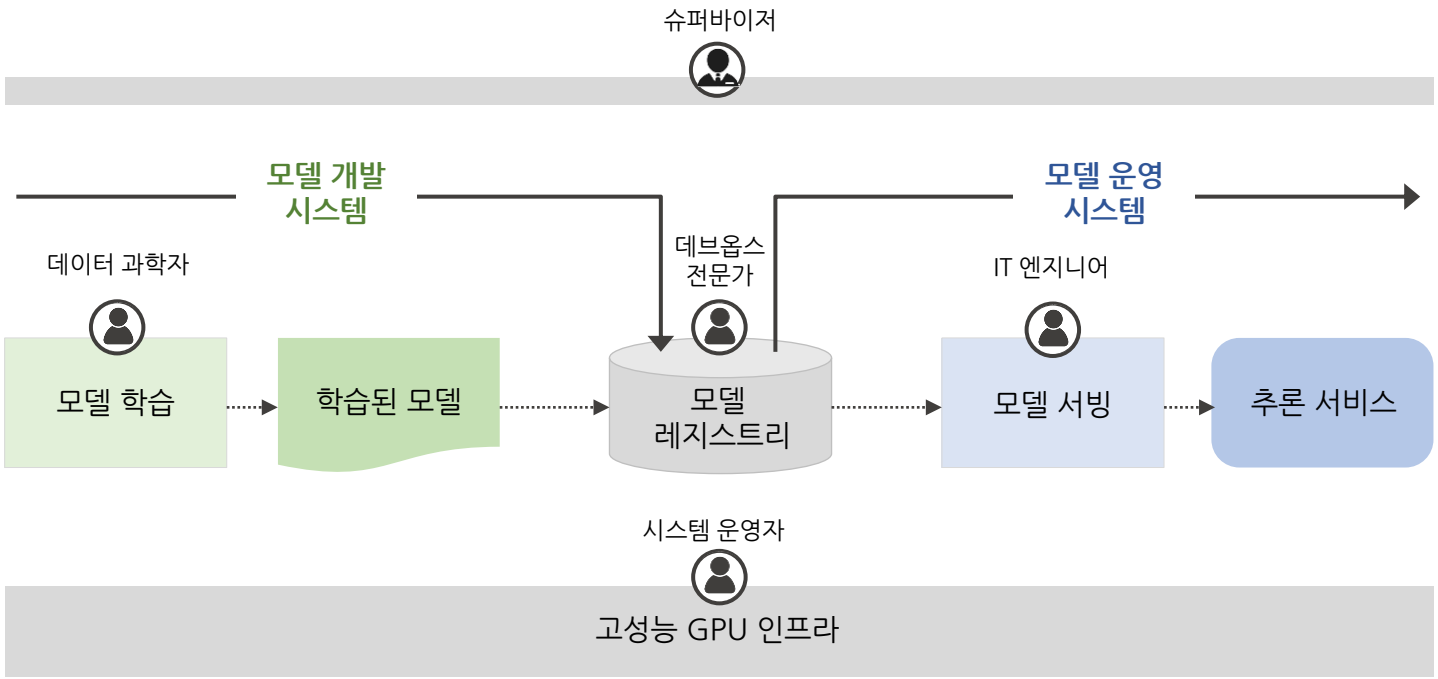
The average time to develop from prototype to production

47%

The average rate for ML models to go into production.

*Survey Analysis: Moving AI Projects From Prototype to Production . Gartner Research ID: G00729313. July 6.

AI 모델 개발에 국한된 것이 아닌 엔지니어링, IT 운영을 포괄하는 관점의 전문 역량 수급 문제

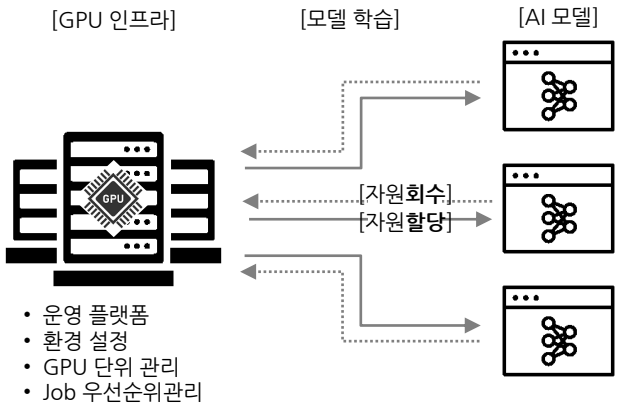


- MLOps 기능을 모두 Open source 기반으로 구축하려면 많은 인원의 엔지니어를 채용해야 함
- ML 개발과 운영을 위해 데이터 과학자, 데이터 엔지니어, DevOps 엔지니어, IT 엔지니어, 시스템 운영자 등 다양한 전문가를 채용해야 하는 문제 발생
- 기존 IT 시스템과 ML을 위한 시스템의 구조적 차이를 이해하고 서로 다른 전문 영역 간 협업하고 R&R을 관리해야 하는 문제 존재

모델 학습과 모델 배포 시 요구되는 시스템과 인프라 관리 포인트가 상이한 문제

모델 학습을 위한 GPU 인프라 문제

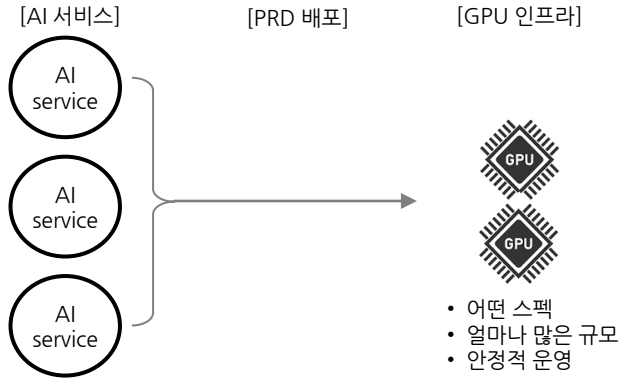
→ 24 X 7, GPU 자원 가동률 극대화



- 온프레미스 인프라 구축 시 모델 학습 특성(Resource-hungry)에 최적화된 운영 플랫폼 필요
- 기존 플랫폼 도입 시 서버 단위로 자원을 할당하므로 GPU 단위의 유휴와 가동률 파악 불가
- 자원 할당 과정에서 서버 설정(OS, Drivers, Libraries등)을 개발자 별로 변경해야 하는 오버헤드 발생
- 모델 학습 단위의 Job 스케줄링 필요

모델 운영을 위한 GPU 인프라 문제

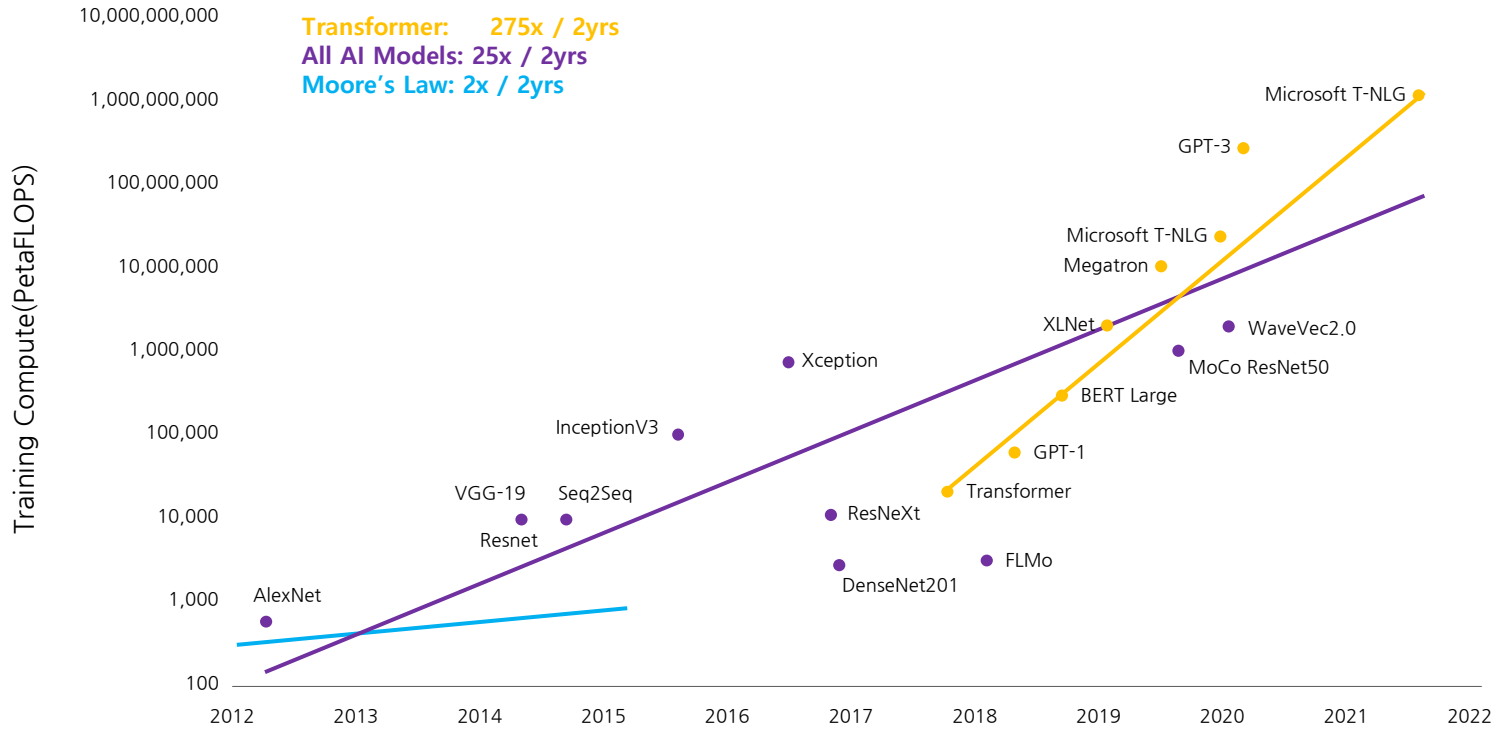
→ 최소한의 GPU 사용으로 안정적인 운영 품질 확보



- 모델 배포 시 서비스 운영을 위한 GPU 자원 스펙과 규모 산정 불가
- 여러 개의 서비스 운영 시 GPU 자원 내 서비스 간 간섭으로 안정성 확보 불가
- 지속적으로 증가하는 모델의 리소스 사용량에 따른 비용 증가의 문제

지속적으로 증가하는 모델의 리소스 수요에 대응한 효율적 인프라 구성 및 운영의 어려움

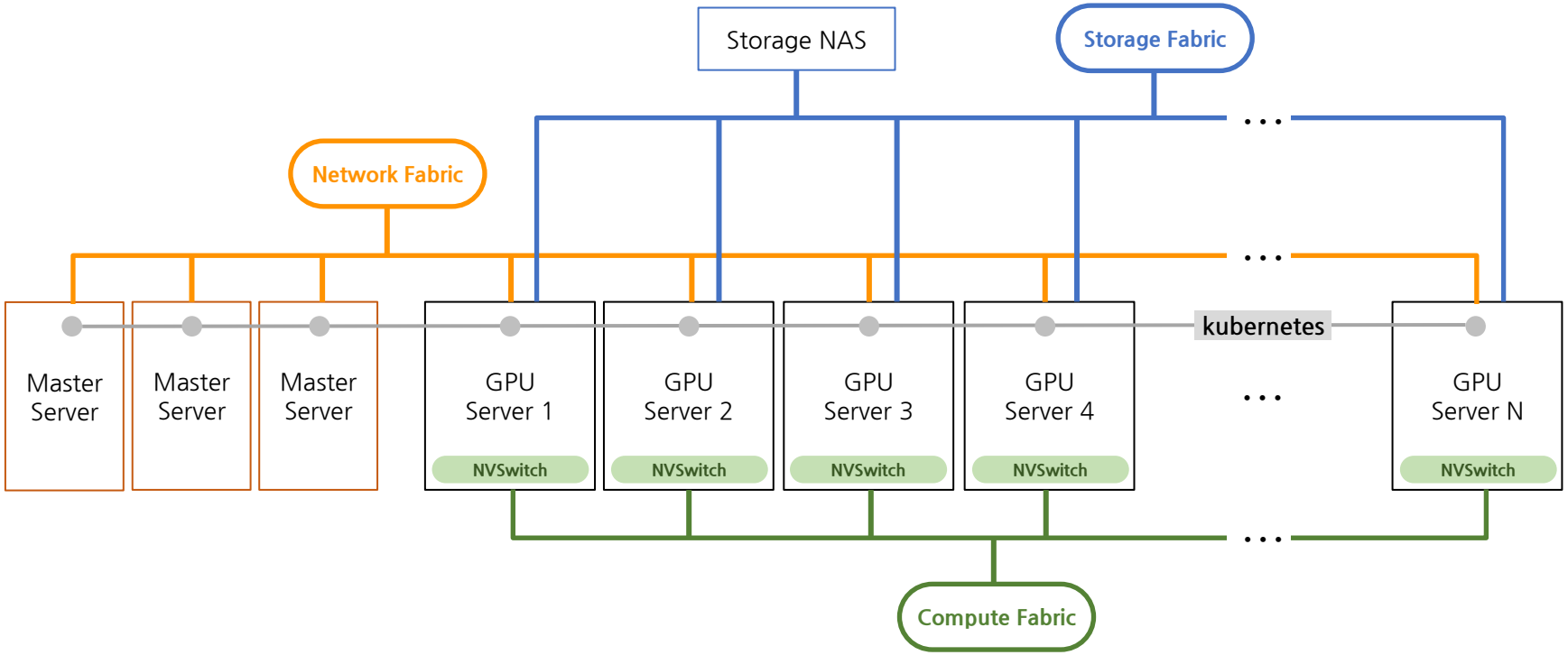
[NVIDIA GTC Keynote, Nov 2021]



- 인공지능 모델의 크기가 연간 10배로 빠르게 증가하여 이를 학습하고 운영하는데 필요한 자원의 가격은 지속 증가 추세[1)]
- AI 모델의 크기는 빠르게 증가하는 추세. 특히, 자기 지도 학습은 트렌스포머 모델로 귀결되며, 모델 크기의 초 대규모화로 인한 학습/추론 컴퓨팅 요구량은 2년에 275배씩 기하 급수적으로 증가 중

1)Big Ideas 2021: ARK Investment Management LLC, 26 Jan 2021

GPU 머신이 증가할 수록 AI 전용 인프라 구성의 복잡도 또한 급격히 증가



• 증가하는 AI 모델 계산량을 커버하기 위해서는 다수의 GPU 머신을 도입해야 하며, 이에 따라 AI 전용 인프라를 구성하기는 더욱 복잡하고 어려움

Our Solutions

Our Solutions

AI lifecycle의 운영(MLOps)과 인프라 활용 문제 해결에 집중하는 완전 관리형 솔루션 'AI Pub'

*AI Pub: Pub(사람들이 편하게 모이는 장소), Public(대중적인), Publisher(AI 배포 도구)

완전 관리형 통합 AI 플랫폼 'AI Pub' 구조

(1) AI Pub

- 비 전문가도 AI 모델을 배포하고 유지보수 할 수 있도록 쉬운 UI의 서비스를 제공하여 비즈니스 도입 속도 제고
- 데브옵스와 인프라 관리 등 특정 영역의 기술 업무를 대체 할 수 있는 기능을 제공하여 기술 전문성 부족 문제 해소

(2) COASTER

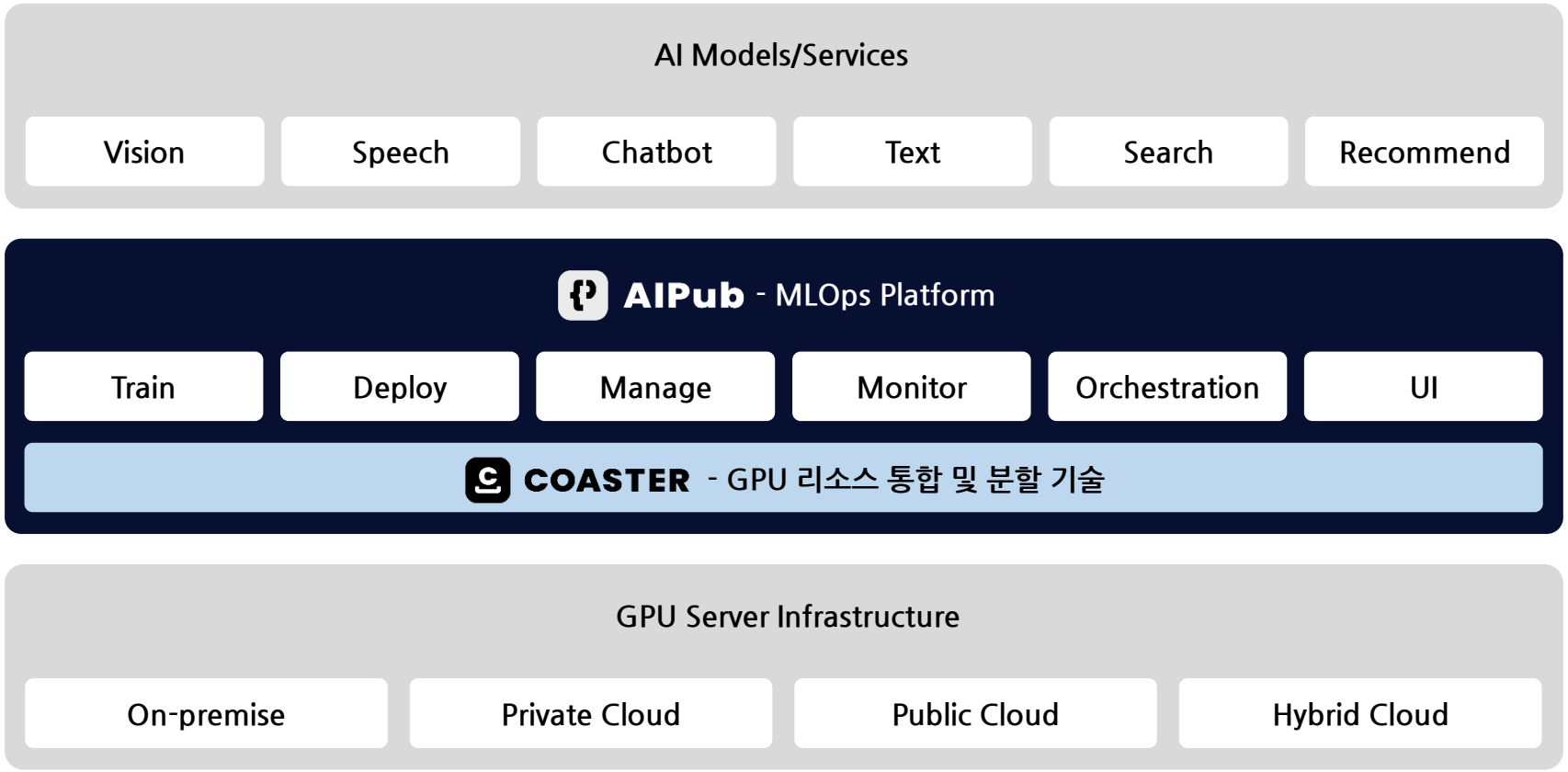
- 값비싼 AI 인프라를 최고의 효율과 최소의 비용으로 활용 할 수 있도록 핵심 기술 적용

(3) Plug-ins

- AI 인프라의 핵심 벤더인 NVIDIA 와 NetApp의 소프트웨어 기능을 연동하여 고객들의 편의성 제고

Our Solutions

‘AI Pub’은 AI 모델 학습 및 운영 업무를 위한 도구와 필요한 인프라의 효율적 서빙 기능을 지원



Our Solutions

GPU 자원 운용 기술 기반 MLOps 플랫폼 'AI Pub'을 통해 AI 개발과 운영의 효율성 극대화

서비스 명	주요 서비스	서비스 상세
AIPub COASTER를 코어로 하여 AI 개발과 운영 업무를 지원하는 완전 관리형 서비스	**AI 개발 환경 세팅**	사용자의 개발 환경을 이미지의 형태로 관리
	컨테이너 워크스페이스	AI 인프라를 팀과 사용자 단위로 할당, 모니터링, 과금 가능한 관리자
	모델 학습	AI 학습 별로 필요한 자원을 자동으로 할당하여 작업 수행, 우선순위 변경 가능
	모델 운영	비 개발자도 서비스 생성, 중지, 삭제, 업데이트, 롤백 가능한 AI 운영
	컨테이너와 서비스 관리	클라우드 네이티브를 위한 관리 및 운영
	인프라와 서비스 모니터링	AI 인프라 관리자를 위한 자원과 서비스 모니터링
COASTER Kubernetes를 확장하여 GPU 인프라 운용과 사용자 관리 기능을 강화한 컨테이너 플랫폼	**GPU 자원의 분할 사용**	GPU 1개의 Utilization과 Memory를 100개 블록으로 나누어 활용
	GPU 자원의 조회와 할당	K8s의 확장 명령어로 클러스터 전체의 컴퓨팅 자원 조회
	User 권한 관리 - Group	리소스 접근 권한을 사용자 그룹단위로 설정 및 관리
	스케줄러 대기열 관리	작업 대기열 상의 우선 순위 변경
Plug-ins 고성능 컴퓨팅 자원을 더욱 효율적으로 활용할 수 있도록 플러그 인 제공	**NVIDIA MIG(Multi-instance GPU)**	MIG로 분할된 Instance에 COASTER 설치 하여 더 밀도 높게 자원 분할
	NVIDIA Triton Inference Server	COASTER와 함께 이용 시 1개 GPU에 더 많은 서비스 집적 가능
	NetApp Trident	COASTER와 연동하여 사용자 별 격리된 Volume 기반 Multi-tenancy 환경 지원
	NetApp Astro	COASTER와 연동하여 여러 개의 마이크로 서비스로 구성된 어플리케이션 관리

COASTER

TEN

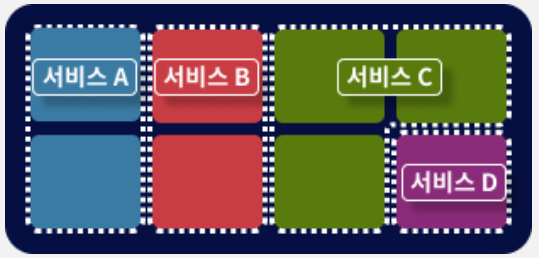
Kubernetes를 확장하여 기능을 강화한 컨테이너 플랫폼 Coaster의 주요 기능

GPU 자원의 분할 사용

[Kubernetes Native - GPU 할당]

GPU 1개 단위로만 컨테이너 할당 가능하여
1개 GPU에 여러 개의 컨테이너를 띄울 수 없음

[Coaster Extended - GPU 할당]



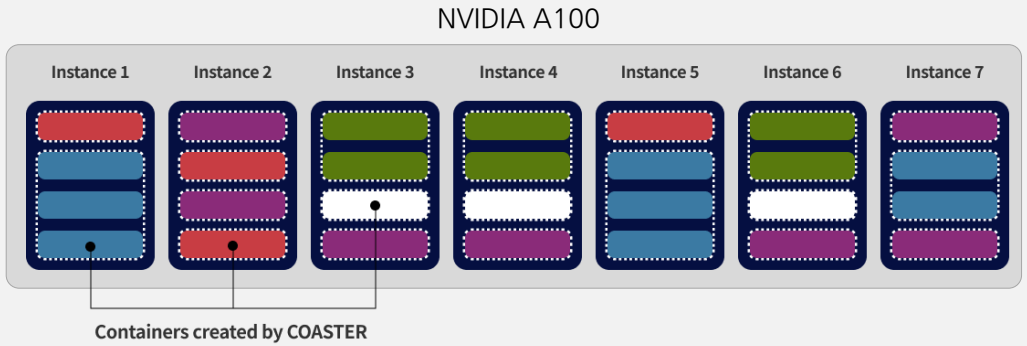
GPU 1개의 utilization와 memory를 1% 단위로
분할하여 100개 블록으로 나누어 활용 가능

- **Block 기반으로 컨테이너에 할당 시 GPU 1개에 여러 컨테이너를 띄울 수 있을 뿐만 아니라 컨테이너 간 리소스 사용량 침해를 막아 안정성 담보 가능**

NVIDIA 의 GPU 활용 기술과 COASTER를 연동하여 자원 활용 효율성 향상

NVIDIA MIG(Multi-Instance GPU)

COASTER X NVIDIA



- NVIDIA A100의 MIG로 분할된 하나의 Instance에 Coaster를 설치하여 더 세밀하게 자원을 분할하여 이용할 수 있음

Kubernetes를 확장하여 기능을 강화한 컨테이너 플랫폼 Coaster의 주요 기능

GPU 자원의 조회와 할당

[Kubernetes Native - GPU 조회]

K8s에서 클러스터 내 자원의 상태를
조회하기 위해 각각의 노드에 직접 접속하여
해당 노드의 정보만 확인 가능함

❯

[Coaster Extended - GPU 조회]

K8s의 확장 명령어를 사용해 클러스터
전체의 컴퓨팅 자원을 조회 가능

```
Resource Type                 Total     Request Max    Available
---------------------------   --------  -------------  -----------
cpu                           56        53.05          53.05
memory                        128313Mi  126949Mi       126949Mi
ten1010.io/tesla-t4           0         0              0
ten1010.io/tesla-t4-block-0   100       97             97
ten1010.io/tesla-t4-block-1   100       100            100


Block Type                    Total    Request Max    Available
----------------------------  -------  -------------  -----------
ten1010.io/tesla-t4-block      200            100          197
```

- K8s에서는 종류가 다른 GPU를 구별하여 컨테이너에 할당할 수 없음 그러나 Coaster에서는 적절한 타입의 GPU에 필요한 Block 수량을 컨테이너에 할당 가능

Kubernetes를 확장하여 기능을 강화한 컨테이너 플랫폼 Coaster의 주요 기능

User 권한 관리 - Group

[Kubernetes Native]

유저 생성 시 유저별로 모든 리소스 접근 권한을 따로 관리해야 함으로 관리가 복잡하고 어려움

[Coaster Extended]

하나의 Group에 정책을 공유할 유저와 네임 스페이스, 공유 저장소, 이미지 레지스트리, 서버 노드 등 리소스 접근 권한을 묶어 한 번에 관리 가능

Kubernetes를 확장하여 기능을 강화한 컨테이너 플랫폼 Coaster의 주요 기능

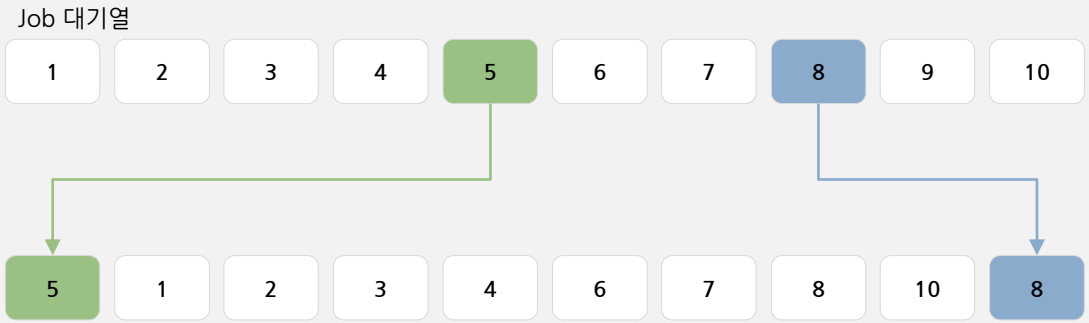
스케줄러 대기열 관리

[Kubernetes Native]

Kubernetes의 기본 스케줄러는 FIFO(First In First Out) 방식으로 대기열 관리 함

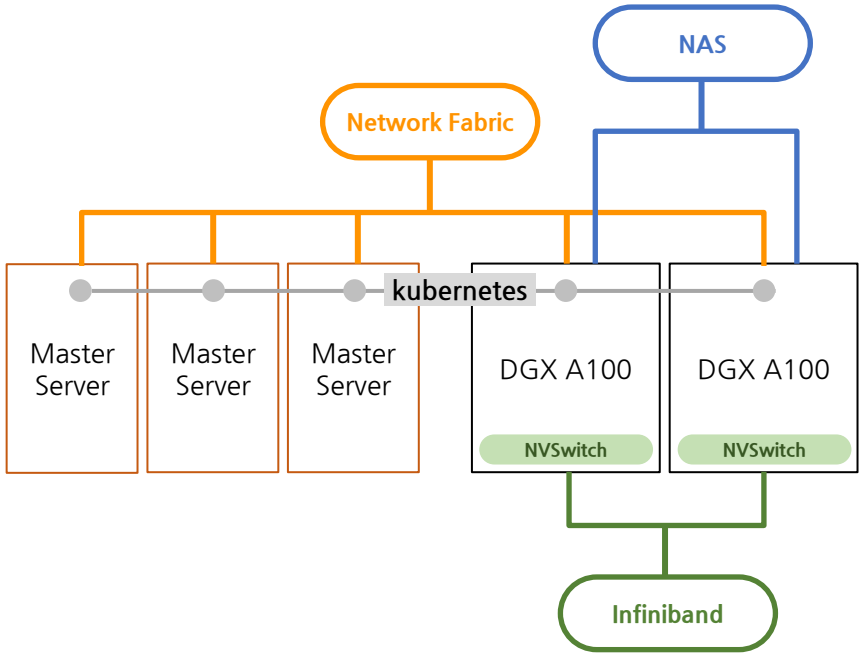
[Coaster Extended]

Coaster의 스케줄러는 Queue에 있는 작업들의 우선순위를 자유롭게 변경 가능



*Coaster 스케줄러 이용 시 K8s의 기본 스케줄러의 동시 사용은 불가합니다.

Coaster 시연 환경



- Kubernetes를 기반으로 GPU 자원을 조회 및 할당
- GPU 블록을 쉽게 컨테이너에 할당
- Infiniband를 통해서 멀티 노드 학습 지원

TEN

COASTER를 코어로 하여 AI 개발과 운영 업무 지원하는 완전 관리 형 서비스 AI Pub

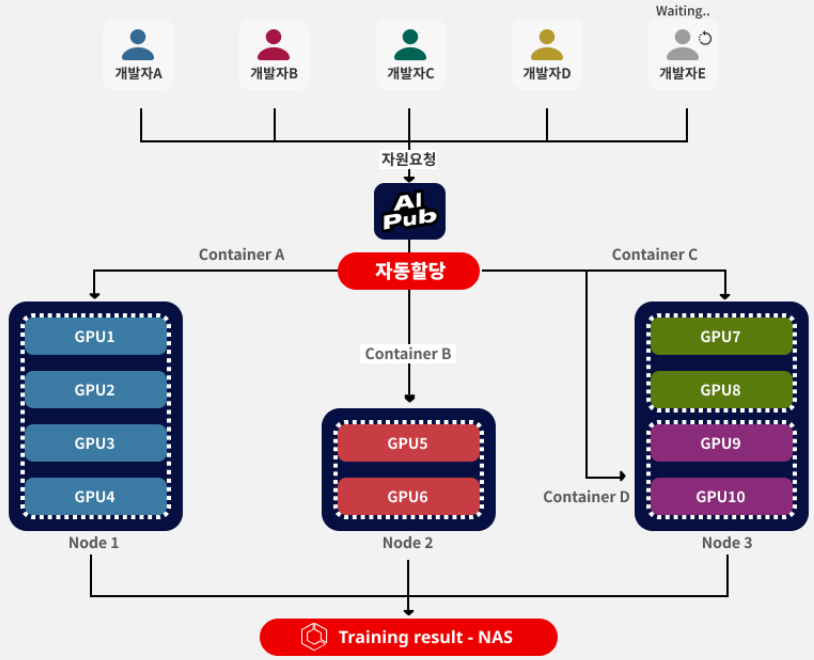
Container Workspace

[Pain Point]

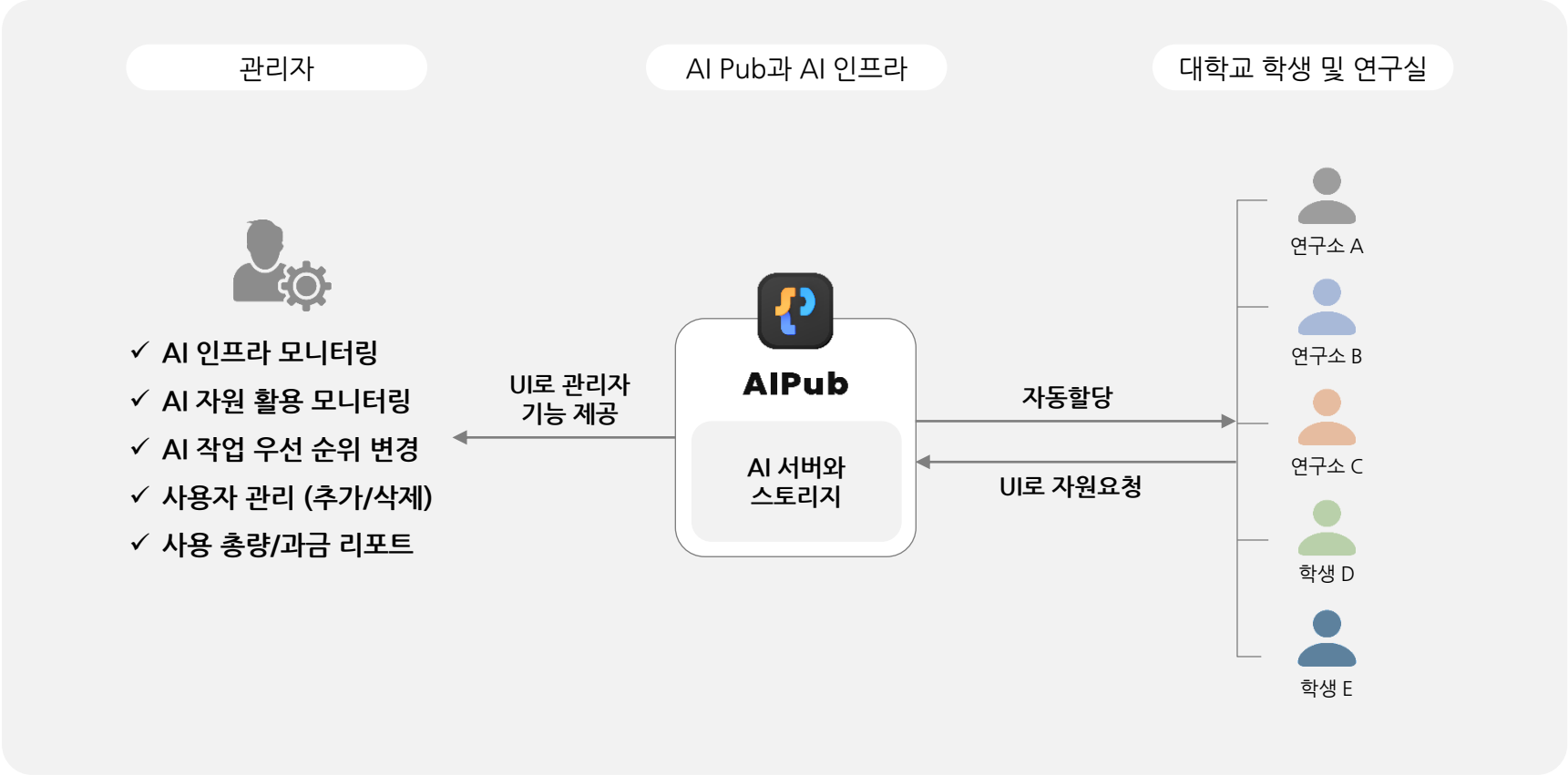
인공지능 개발용 인프라를 구축하여 중앙 관리 할 경우
개발자나 팀 단위로 자원을 할당 및 관리 방법 부재,
각 단위 별 사용량을 측정할 수 있어야 함

[AI Pub Solution]

AI 인프라를 팀 혹은 사용자 단위로 할당 가능
팀 내 개발자들의 자원 사용량 및 팀 별 사용량
집계 및 과금 가능



AI 개발을 위한 대학의 AI Pub 도입 사례 - AI 자원의 가동률 극대화



- 1인 1서버 개발이 어려운 고가의 AI 서버 특성 상 중앙에서 GPU 클러스터를 구축하고 나누어 쓰는 환경 필요
- 온프레미스 환경에서도 재정적, 관리적 측면에서 클라우드와 같은 '서비스형 인프라' 제공
- AI Pub이 제공하는 GPU자원 관제/분배/회수/과금 서비스를 통한 중앙 관리로 자원의 가동률을 극대화 함

COASTER를 코어로 하여 AI 개발과 운영 업무 지원하는 완전 관리 형 서비스 AI Pub

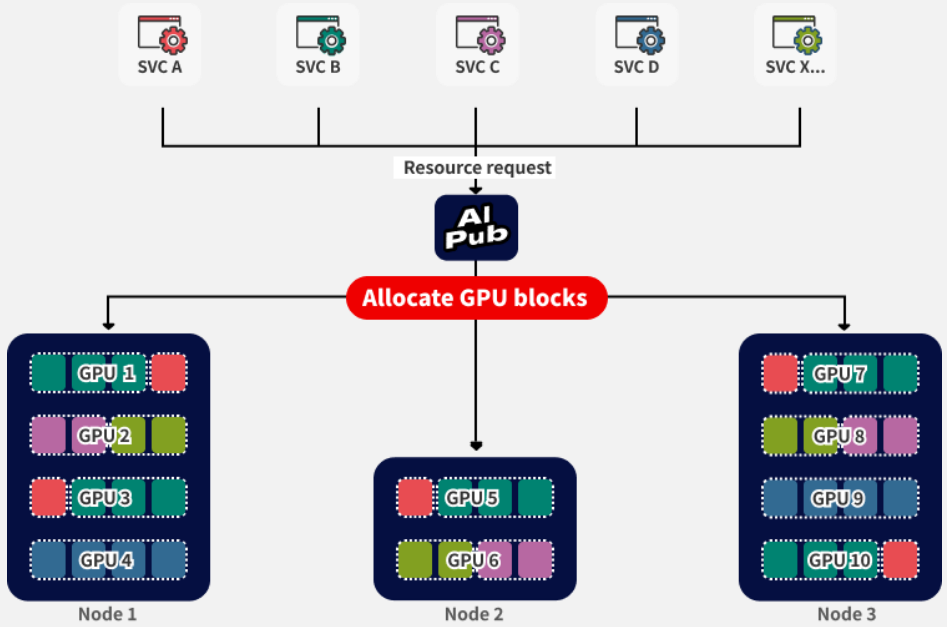
Model Operation

[Pain Point]

AI 서비스 운영 시 최소한의 자원만을 사용하여
비용 절감 필요

[AI Pub Solution]

사용자의 서비스를 이미지의 형태로 관리
GPU 1개를 100개의 Block으로 나누어 최소 단위로 운영
비개발자도 서비스 생성, 중지, 삭제, 업데이트, 롤백 가능



COASTER를 코어로 하여 AI 개발과 운영 업무 지원하는 완전 관리 형 서비스 AI Pub

Container and Service Management

[Pain Point]

온프레미스 서버 운영을 위한 서비스가 필요 함

[AI Pub Solution]

퍼블릭 클라우드에서 제공하는 서비스 운영 및 관리 기능 제공

High Availability

이중화를 통해
Single point of failure 제거

Rolling Update

서비스 중단 없이
업데이트 상시 가능

Load Balancing

서비스 요청을 바쁘지 않은
서버로 자동 분배

Scale-out

서비스 요청에 따라 자동으로
서버 수를 늘려 트래픽 처리

Fail over 대응

서비스가 죽은 경우 탐지 및
새로 서비스를 띄워 안정성 확보

이상징후 알림

중요 운영 이벤트 발생 시
카톡, 슬랙으로 실시간 알림

COASTER를 코어로 하여 AI 개발과 운영 업무 지원하는 완전 관리 형 서비스 AI Pub

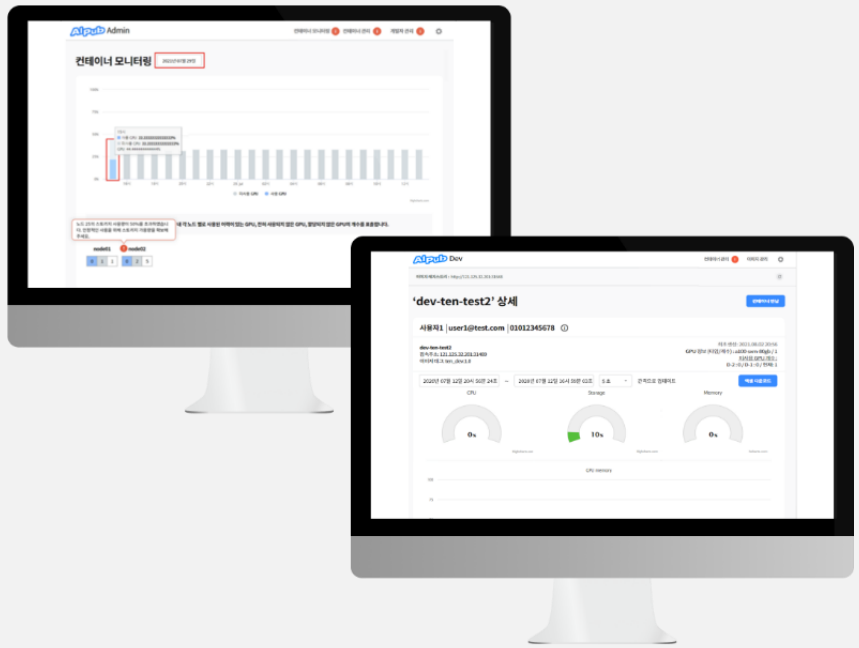
Infra and Service Monitoring

[Pain Point]

서비스와 시스템의 상태를 탐지하고 오류를
관리 할 수 있는 서비스 필요

[AI Pub Solution]

인공지능 자원 관리자와 서비스 운영자를 위한
모니터링 서비스 제공

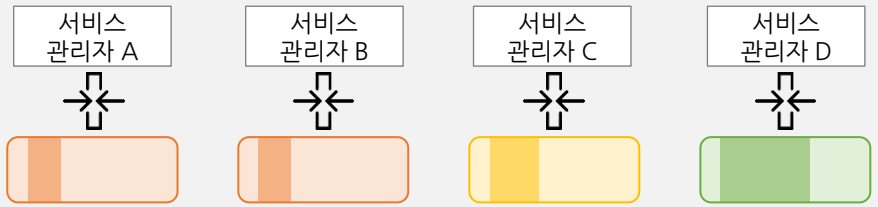


Solution Use Case

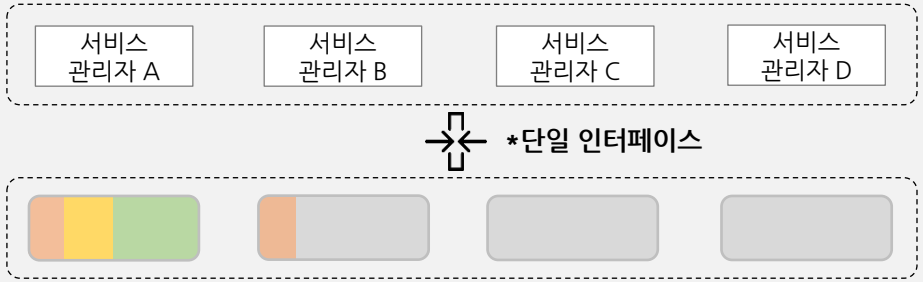
AI 운영을 위한 AI Pub 도입 사례 - AI 서비스 운영 비용을 10분의 1로 절감

서비스 관리 Interface GPU 자원

[AI Pub 적용 전] 서비스 별 각각의 인터페이스로 관리
→ 서비스 운영에 필요한 만큼의 분량이 아닌 개별 GPU 전체를 점유하게 됨



[AI Pub 적용 후] 단일 인터페이스로 관리
→ 서비스 별로 필요한 만큼만 GPU를 분할하여 사용함으로 적은 개수의 GPU에 집적 운영 가능



- 서비스 관리자(운영자)별로 **종류가 다른 AI 서비스**를 **다른 고객(부서)의 의뢰**를 받아 운영
  → 관리자 별로 서로 다른 운영 인프라를 받아서 운영
- 이런 경우 서비스 별로 GPU전체를 할당하여 운영하는 것이 손 쉬운 방법이지만 리소스 낭비가 막대해 짐

Overview

WE ARE HIRING!!

인공지능을 보편화 한다면 더 많은 혜택이
사회 구성원에게 돌아갈 것이라는 믿음으로
함께 세상을 AI롭게 하실 분들을 찾습니다!

개발자를 중심으로 전직군에 대한 채용 중이니 관심 있는 분들은 홈페이지를 통해 연락 부탁드립니다!

TEN

감사합니다. ☺

BN I&C TEN