회사소개

1995년 설립, 업종-전산&정보통신, 국내 주요거점 도시에 영업소운영

일반현황

- 설립연도 : 1995년 9월
- 자 본 금 : 20억원
- 대표이사 : 이기호
- 업 태 : 도·소매업, 제조, 서비스
- 종 목 : 컴퓨터/주변기기, SW개발/공급, 정보통신공사 외

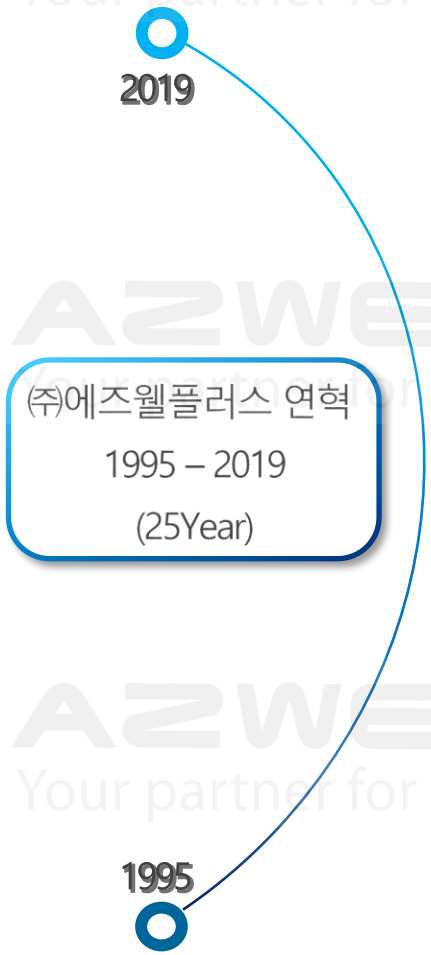
종업원 수

- 전체 : 113
  - 엔지니어 : 68
  - 영업대표 : 28
  - 관리부서 : 17

본사 및 지사 사무실

회사소개

25년의 업력, NVIDIA, 오라클, EMC 외 건실한 제조사와의 파트너쉽 체결



2019

- **NVIDIA Enterprise 최대 매출 NPN**
- 중견기업 자격 획득

2018

- **매출 2,184억원 달성**
- Authorized NVIDIA Elite Partner 계약

2016

- Authorized Dell Tier-1 Partner 계약
- Authorized NVIDIA Preferred Partner 계약

2011 ~ 2014

- Authorized Virtensys (VAR) Partner 계약
- Authorized VMware Enterprise Partner 계약
- EMC Isilon-partner of the Year award, Asia Pacific and Japan 수상
- Authorized ePapyrus 총판 계약

2004 ~ 2010

- **회사명 변경: (주)에즈웰플러스 (2004년 7월)**
- Sun Iforce Partner(Service & Sales 부분) 체결
- Sun StorageTEK 관련 지방 LSP 선정
- Authorized Oracle DMP/Gold Partner 계약
- Isilon 총판 계약

1995 ~ 2003

- **(주)써니벨 시스템 설립**
- Authorized Sun Partner 계약
- 정보통신부 산하 정보통신 연구진흥원 기술개발 사업자 선정
- 병역특례 (전문 연구 기관) 업체 지정 (12. 28)

NVIDIA DGX A100과 함께하는
AI 인프라 구축 방안

2020. 10.

AZWELL PLUS
Your partner for e-Business

목 차

1. Why GPU?

1. Why GPU?

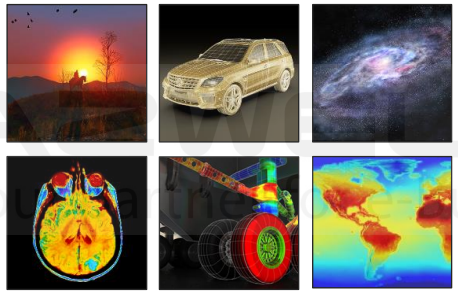
✓ AI 핵심기술, 딥러닝

- 딥러닝은 **알파고** 열풍 이후, 세상을 혁신하기 위한 가장 중심적인 기술로 각광받고 있음

AlphaGo

- 딥러닝 기술은 주로 **자율주행/자연어처리/영상탐지** 등에서 두각을 나타내고 있으나, 사실상 가공된 데이터만 있다면 **모든 산업**에서 적용 가능함

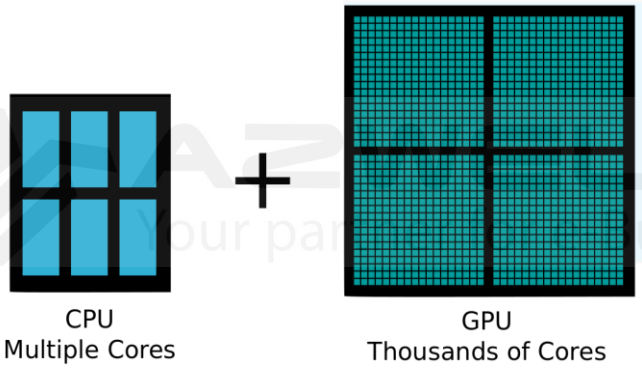
- **Google, FaceBook, Microsoft, Baidu** 등 글로벌 규모의 IT 업체들이 딥러닝 기술 개발에 앞서 있으며, **GPU**를 활용하여 다양한 성과를 발표하고 있음



1. Why GPU?

✓ AI에서 GPU는 왜 필요한가?

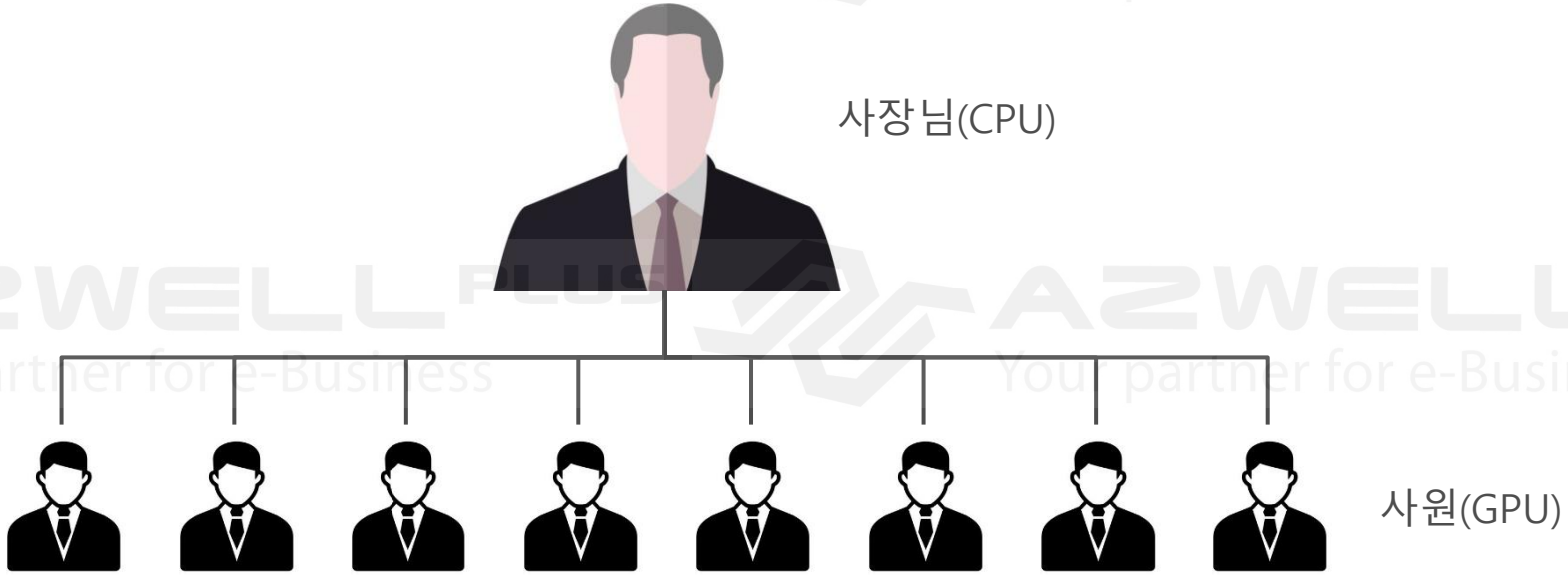
딥러닝은 반복학습을 통해 결론을 도출해 내는 알고리즘 하나의 정밀한 연산보다 여러 프로세스를 동시 처리하는 병렬 연산이 요구됨



- 데이터 학습량이 많으면 많을수록, 정교한 딥러닝 알고리즘 구현 가능!
- GPGPU(General-Purpose computing on Graphics Processing Units) 기술의 발달로 **반복학습** 위주의 워크로드에서 많은 Core를 가진 GPU가 CPU를 압도하기 시작
  - CPU Core는 최대 128 Core, GPU Core는 약 **6000~8000 Core**
- **AI 인프라에서 고성능 GPU는 이제 선택이 아닌 필수**

1. Why GPU?

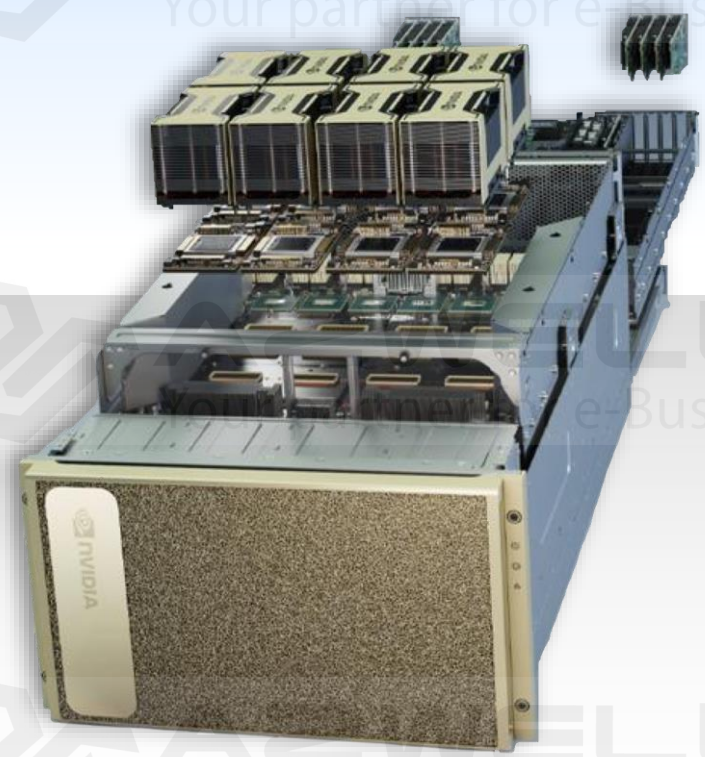
✓ GPU- System에서 CPU-GPU와의 관계



- 반복적 연산 위주의 DL Workload에 있어 CPU는 GPU에 **Order 역할 수행**
  - CPU의 Order를 받은 **GPU**들이 실제 연산을 수행하고, **결과값을 CPU에 제출**
  - 각 최종 결과값을 기반으로 CPU에서 DL 모델을 완성

2. NVIDIA

DGX A100 제품 소개



2. NVIDIA DGX A100 제품 소개

✓ AI 인프라 투자 담당자의 고민

'점점 발전하는 DL 기술,
But 어떤 System을 도입해야 하는지?'

'DL 워크로드는 수집/학습/추론 등,
But 각각 다른 GPU를 별도 구성 및 운영 필요'

'크고 작은 Job 필요한 다수의 사용자'
But 기존 GPU 할당은 1GPU 당 1Job만 가능

'DL 분야의 S/W는 매 달마다 새로운 Version 출시
But 대부분 오픈소스, 어떻게 관리하고 운영해야 하는지?'

2. NVIDIA DGX A100 제품 소개

✓ DGX A100 특장점

A. 최신의 NVIDIA 고성능 GPU A100

8x Tesla A100 40G SXM4
For NVLink

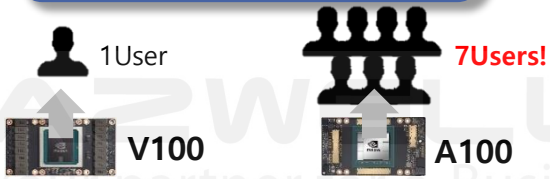
- 전세대 대비 약 5배 성능 향상된 최고성능 GPU Tesla A100 8장 장착!
- 세계 최고 성능의 600GB/s 8-Way NVLink 지원으로 대규모 딥러닝 최적화 연산 수행

B. 모든 워크로드 통합

데이터 수집/학습/추론 통합 수행

- 딥러닝에 필요한 모든 워크로드를 별도 구축하지 않고 A100만으로 통합 수행 가능!

C. 다수사용자 배포 최적화



- 다수 사용자에게 배포가 불가능하던 기존 세대 GPU의 단점을 보완
- DGX A100(8x A100)은 최대 56User에게 리소스 할당 가능!

NVIDIA DGX A100
DL/HPC 전용 슈퍼컴퓨터

D. GPU 전문 프리미엄 서비스 제공

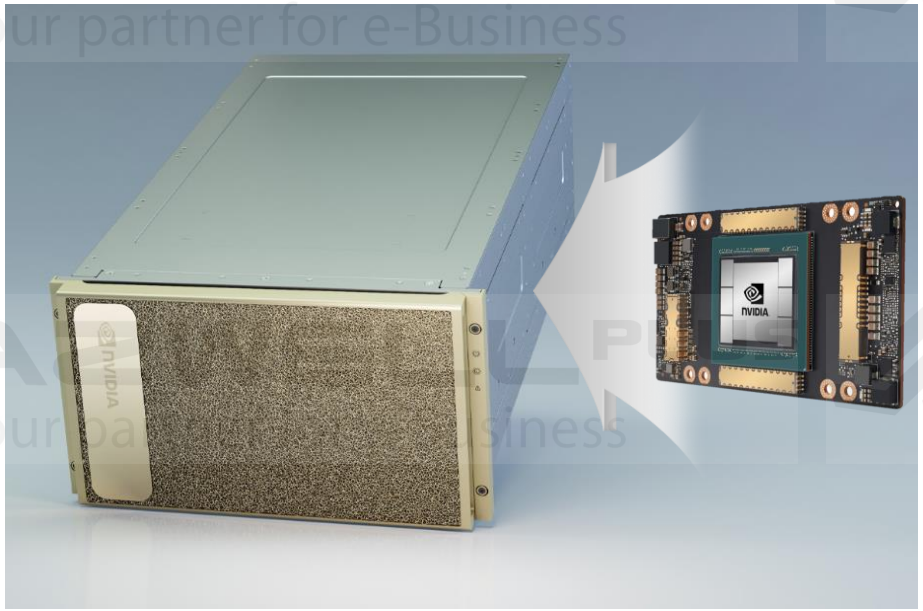
NVIDIA ELITE PARTNER + NVIDIA

DL/GPU 전문 Vender, Elite Partner의 프리미엄 기술지원 서비스 제공

- DL/GPU에 특화된 전문 Vender/Partner의 기술지원 서비스 제공
- Linux의 오픈소스 기반 DL 생태계에서 지속적인 S/W 업데이트 제공

2. NVIDIA DGX A100 제품 소개

✓ A. 최신의 NVIDIA 고성능 GPU A100



GPUs	8x NVIDIA A100 Tensor Core GPUs
GPU Memory	320 GB total
Performance	5 petaFLOPS AI 10 petaOPS INT8
NVIDIA NVSwitches	6
System Power Usage	6.5kW max
CPU	Dual AMD Rome 7742, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)
System Memory	1TB
Networking	8x Single-Port Mellanox ConnectX-6 VPI 200Gb/s HDR InfiniBand 1x Dual-Port Mellanox ConnectX-6 VPI 10/25/50/100/200Gb/s Ethernet
Storage	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 15TB (4x 3.84TB) U.2 NVME drives
Software	Ubuntu Linux OS
System Weight	271 lbs (123 kgs)
Packaged System Weight	315 lbs (143kgs)
System Dimensions	Height: 10.4 in (264.0 mm) Width: 19.0 in (482.3 mm) MAX Length: 35.3 in (897.1 mm) MAX
Operating Temperature Range	5°C to 30°C (41°F to 86°F)

- **현존하는 System 중 가장 최고성능의 컴퓨팅 리소스 제공!**
  - **5PetaFLOPS** DL 성능 제공
  - DGX-1 대비 **5배** DL 성능 제공
  - **55,312** CUDA Core, **3,456** Tensor Core
  - 최대 **56User** 사용 가능**(Multi Instance GPU)**

2. NVIDIA DGX A100 제품 소개

✓ A. 최신의 NVIDIA 고성능 GPU A100 – DGX SYSTEM Spec 비교

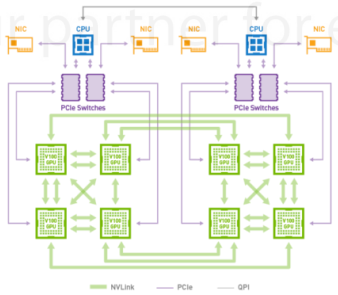
구분	DGX-1V	DGX-2	DGX A100
GPUs	8x Tesla V100 32GB SXM2	16x Tesla V100 32GB SXM3	**8x Tesla A100 40GB SXM4**
GPU Memory	256 GB total System	512 GB Total System	**320 GB Total System**
CPUs	2x Intel Xeon E5-2698 v4 2.2 GHz 20-Cores, Support PCIe 3.0	2x Intel Xeon Platinum 8168, 2.7 GHz 24-cores, Support PCIe 3.0	**2x AMD ROME 7742 Processor 2.25 GHz, 64Core, Support PCIe 4.0**
CUDA Cores	40,960	81,920	**55,312**
Tensor Cores	5,120	10,240	**3,456 (Next Generation)**
DL Performance (Training)	1 PetaFLOPS (FP16, Mixed)	2 PetaFLOPS (FP16, Mixed)	**5 PetaFLOPS (FP16, With Sparsity)**
INT8 / INT4 (Inference)	X	X	**9,984 TOPS / 19,968 TOPS**
System Memory	512 GB DDR4	1.5 TB DDR4	**1TB DDR4**
Storage	1x 480 GB SATA SSD (OS) 7 TB(4x 1.92TB) SATA SSD RAID 0 (DATA)	2x 960 GB NVME SSDs RAID 1 (OS) 30 TB(8x 3.84TB) NVME SSDs RAID 0 (DATA)	**2x 1.9 TB M.2 NVMe SSD RAID 1 (OS) 15 TB(4x 3.84TB) U.2 NVMe SSD RAID 0 (DATA)**
Network	4x IB EDR 100 Gb (VPI) 2x 10 Gb ETH	8x IB EDR 100 Gb (VPI) 2x 10/25 Gb ETH	**8x IB HDR 200 Gb (VPI) 2x 10/25/50/100/200 Gb ETH**
TDP	3 kW	10 kW	**6.5 kW**

2. NVIDIA DGX A100 제품 소개

✓ A. 최신의 NVIDIA 고성능 GPU A100 - DGX System NVLink 비교

8x SXM2 GPU Cube Mesh Architecture

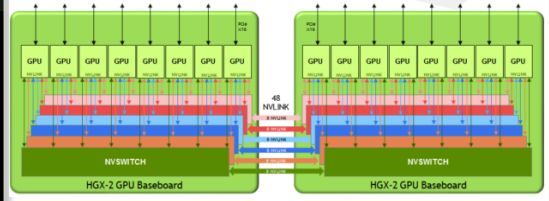
DGX-1V



특정구간 최대
100 GB/s P2P 통신
GPU-GPU 최대 2 NVLink

16x SXM3 GPU NVSwitch Architecture

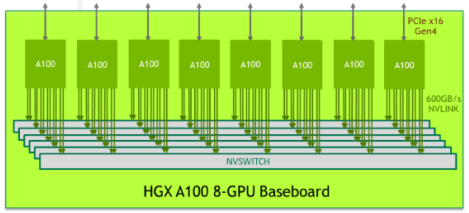
DGX-2



모든구간
300 GB/s P2P 통신
GPU-GPU 6 NVLink

8x SXM4 GPU NVSwitch Architecture

DGX A100

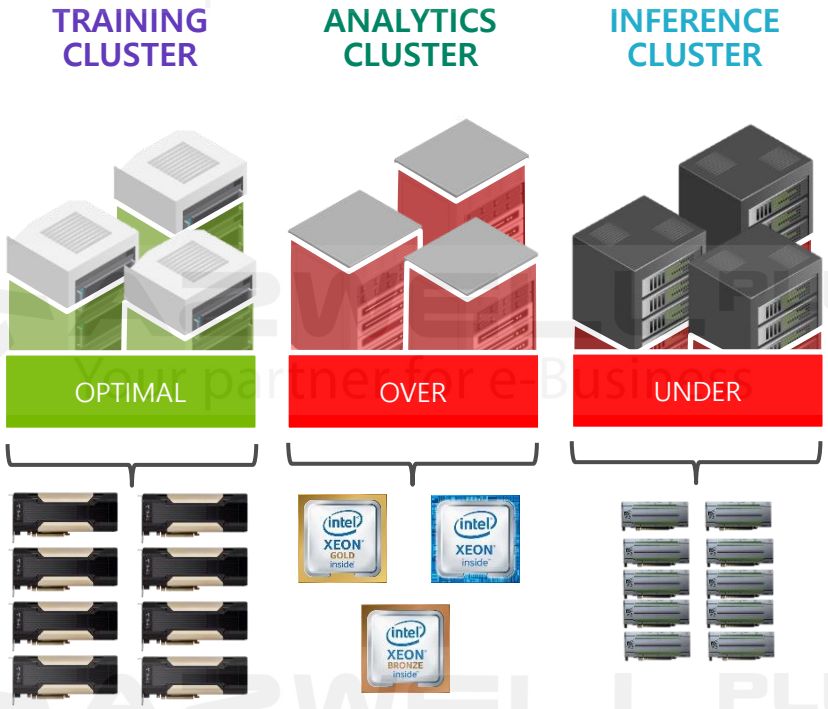


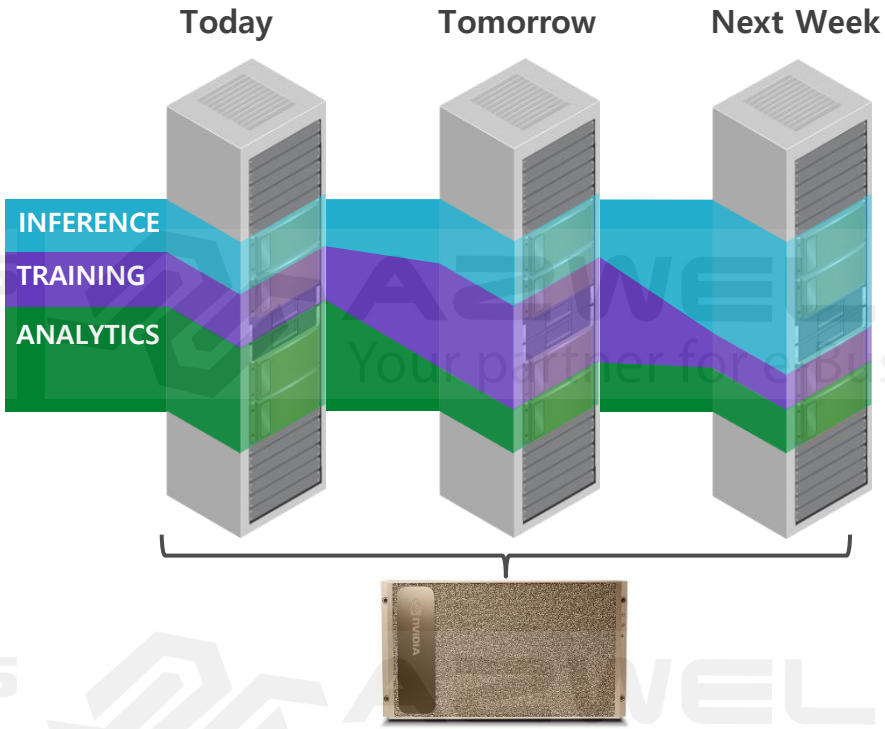
모든구간
600 GB/s P2P 통신
GPU-GPU 12 NVLink

- V100 NVLink2.0(GPU당 6개 NVLink) 대비 최소 **2배** 대역폭 제공
- A100 NVLink3.0은 GPU당 **12개 NVLink** 제공

2. NVIDIA DGX A100 제품 소개

✓ B. 모든 워크로드 통합

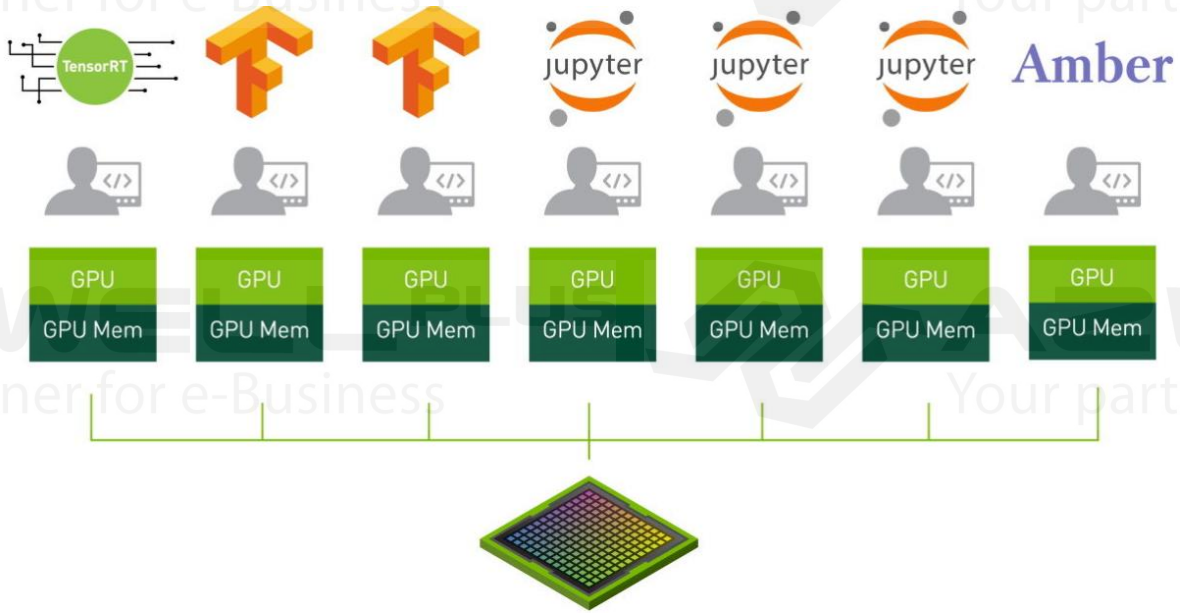




- 기존의 DL을 위한 전통적인 클러스터 구성방식은 Training / Analytics / Inference 로 **별도 구축**
  - **유휴자원**이 필연적으로 발생할 수 밖에 없음
- Training / Analytics / Inference의 **모든 성능**을 제공하는 **DGX A100**에 **MIG** 기능을 적용
  - 최소한의 **전력/상면/비용** 투자로 유휴자원 없이 **모든 워크로드를 통합 수행!**

2. NVIDIA DGX A100 제품 소개

✓ C. 다수 사용자 배포 최적화



- 이전 세대까지의 1JOB-1GPU 할당 기술을 보완, 하나의 GPU를 분할하여 사용할 수 있는 **MIG** 기능 제공
- 40GB VRAM의 A100을 5GB 단위로 최대 **7 User**까지 할당 가능
  - DGX-A100에서 최대 **56 User**까지 **Training/Inference/HPC** 워크로드 수행 가능!
  - **작은 규모의 워크로드**에도 유연한 리소스 제공

2. NVIDIA DGX A100 제품 소개

✓ D. GPU 전문 프리미엄 서비스 제공

1

OS/Firmware/Driver/ Framework

모든 단계의 DL Stack이 사전 최적화 설치 제공

GPU-accelerated containers

PyTorch, MXNet, TensorFlow

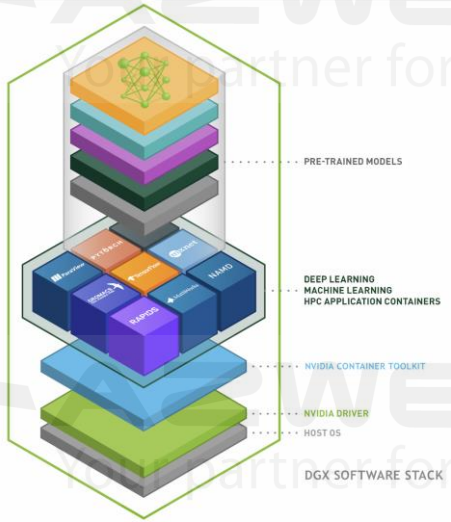
TensorRT

RAPIDS

CUDA, CUDA-X HPC, OpenACC

NVIDIA CUDA Toolkit

*** 출고 시 모두 사전 최적화 설치되어 제공, 주기적 업데이트**



2

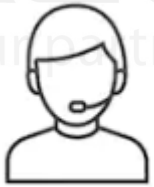
DL/GPU만 담당하는 Vender/Partner의 전문화된 기술지원 서비스 제공

- 일반 System과 DL/GPU를 모두 담당하는 엔지니어의 기술지원

WAS/Storage/GPU /Backup, ETC Support

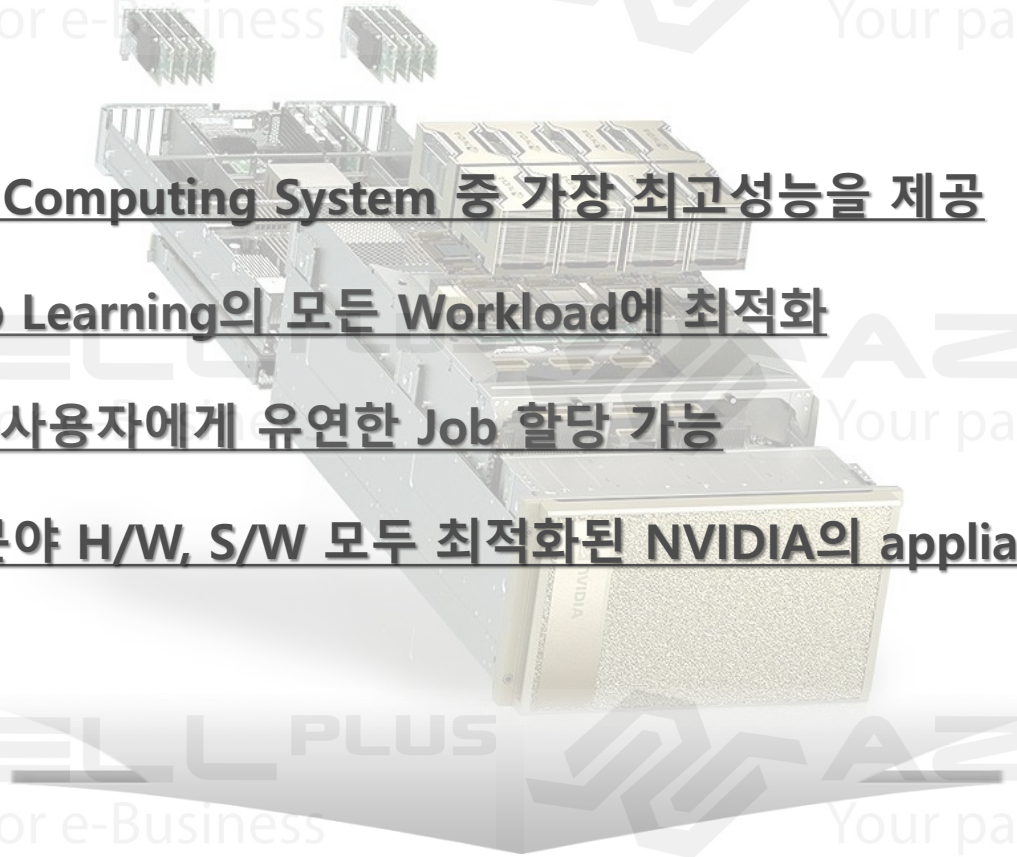
- DL/GPU System만 담당하는 엔지니어의 전문 기술지원

NVIDIA DGX System Support



2. NVIDIA DGX A100 제품 소개

✓ DGX A100 SUMMARY



GPU Computing System 중 가장 최고성능을 제공

Deep Learning의 모든 Workload에 최적화

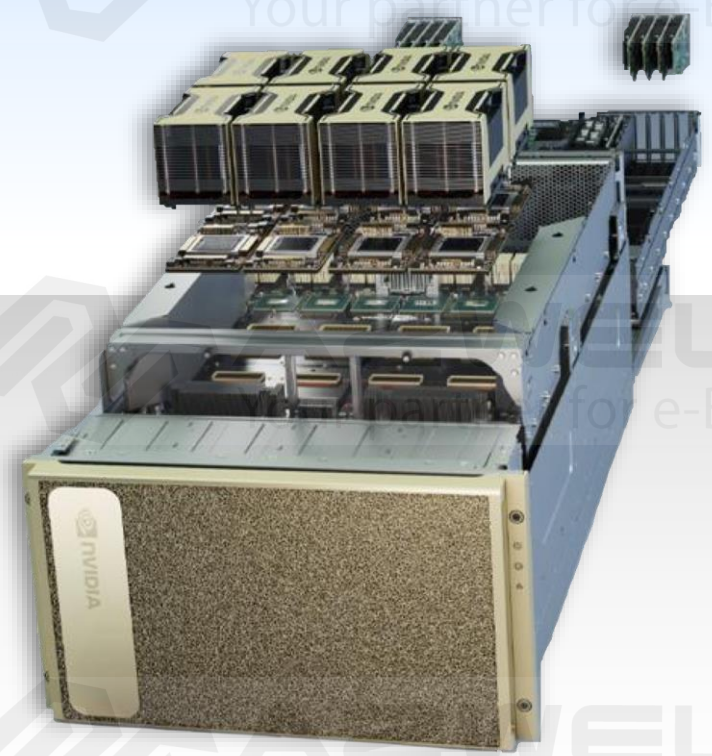
다수 사용자에게 유연한 Job 할당 가능

DL 분야 H/W, S/W 모두 최적화된 NVIDIA의 appliance System

최종 투자비용 절감, 성능 및 효율 극대화!

3. DGX A100

구축 가이드



3. DGX A100 구축 가이드

✓ 각 분야별 AI 인프라 구성 방법

학교/연구 환경 AI 인프라

심도깊은 DL 연구를 위해 큰 규모의 GPU 자원 필요	학생들의 실습환경 구성을 위해 작은 규모의 GPU 자원 필요

DGX A100 + Job Scheduler(인프라 운영/관리자)

Enterpirse 환경 AI 인프라

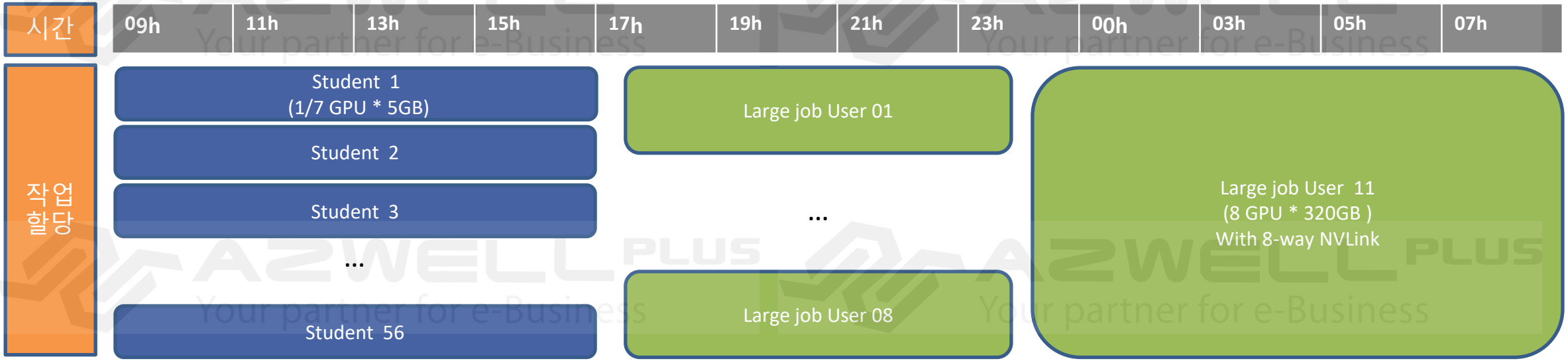
자율주행, 음성인식과 같은 큰 프로젝트 여러 개의 GPU를 하나의 Job에 할당하며, 여러 노드를 마치 하나의 노드처럼 활용할 수 있어야 함	딥러닝은 수많은 대용량 데이터를 실시간으로 수용하는 워크로드, 최적화된 스토리지 설계가 중요

DGX A100 + Job Scheduler(인프라 운영/관리자)

- 각 분야별로 수행 프로젝트 규모에 따라 DGX A100을 맞춤형으로 구성
- Job Scheduler로 인프라를 효율적으로 관리

3. DGX A100 구축 가이드

✓ 학교/연구 환경 AI 인프라 – 용도별 GPU 사용 시나리오



학생들 강의 실습을 위한 GPU 할당, GPU 8개를 MIG 분할하여 56User에게 할당

교수님, 연구원들의 소규모 연구를 위한 GPU 할당, GPU 1개씩 8User에게 할당

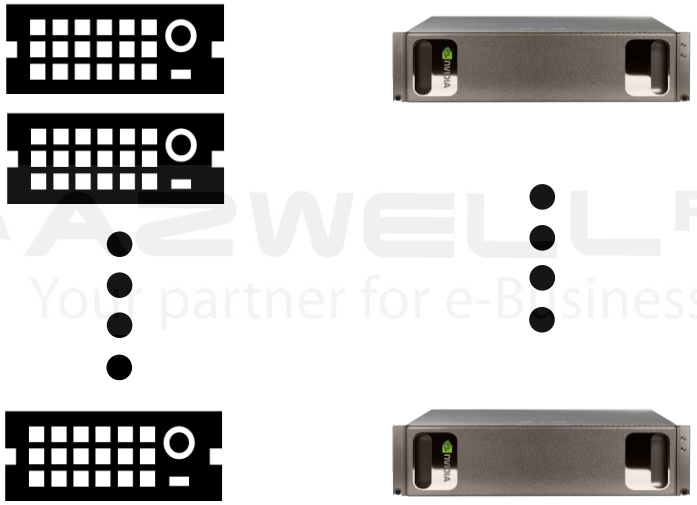
교수님, 연구원들의 대규모 연구를 위한 GPU 할당, GPU 8개를 1User에게 할당 NVLink 기능을 활성화하여 8GPU를 하나의 GPU처럼 사용

- **MIG, NVLink** 기능을 활용하여 유연한 사용자 할당
- 정해진 타임 테이블에 따라 GPU를 분할-할당-회수-재할당 진행
  - ➢ 09 ~ 17시 학생 실습용 56User 사용, 17 ~ 09시 연구용 User 사용
- **Job Scheduler** 도입 시 운영자의 실시간 개입 없이도 자동화 운영 가능 (맨 뒤의 장표에 추가설명)

3. DGX A100 구축 가이드

✓ 학교/연구 환경 AI 인프라 – 필요 서버 규모 비교

DGX A100 이전 환경 (As-is)



* 강의 실습용 인프라
2080Ti 서버 (8GPU)
7대 – 총 56GPU

* 심층 연구용 인프라
DGX-1 (8GPU)
3~5대

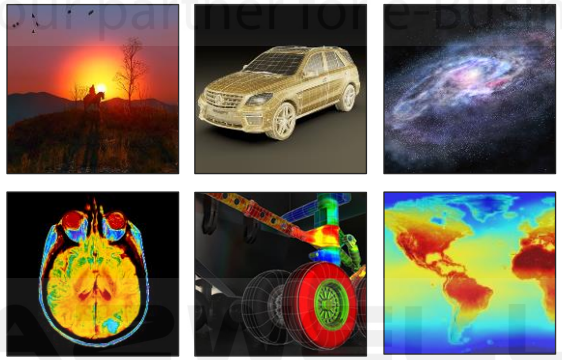
DGX A100 적용 환경 (To-Be)

* 강의 실습/심층 연구용 인프라
DGX A100 (8GPU)
1대

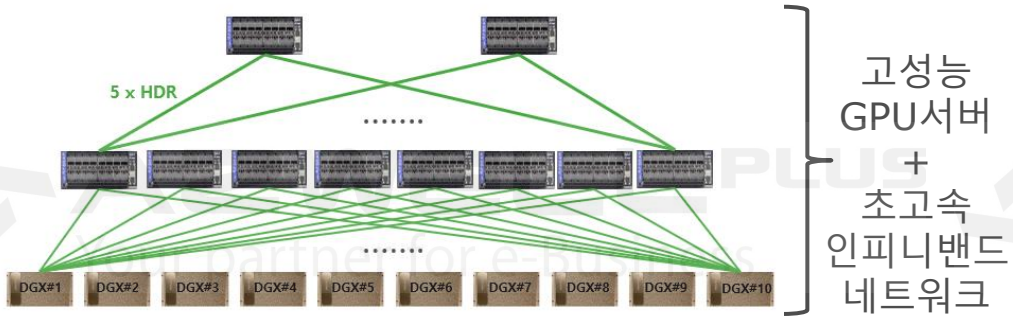
- A100의 유연한 활용
  - GPU MIG 기능: 최대 **56유저** 지원, 총 7대 서버가 필요한 실습용 인프라를 대체
  - NVLink 기능: 최대 **5 PetaFLOPS** 성능 지원, 기존 연구용 고성능 서버 3~5대를 대체
    - **가동률을 극대화하여 투자 대비 효율성을 향상, 전력 및 상면비용 비교불가**

3. DGX A100 구축 가이드

✓ Enterprise 환경 A100 인프라 – Multi Node Network 구성



- **자율주행, 음성인식** 등의 기업 단위 대규모 프로젝트에는 서버와 서버를 하나로 묶어 하나의 딥러닝 모델 학습을 수행할 수도 있어야 함

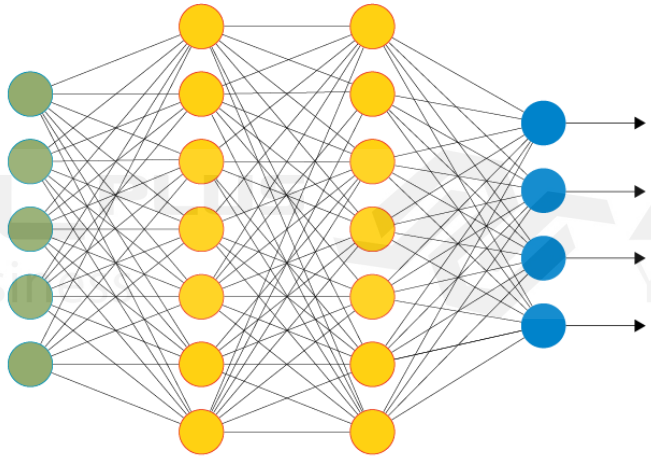


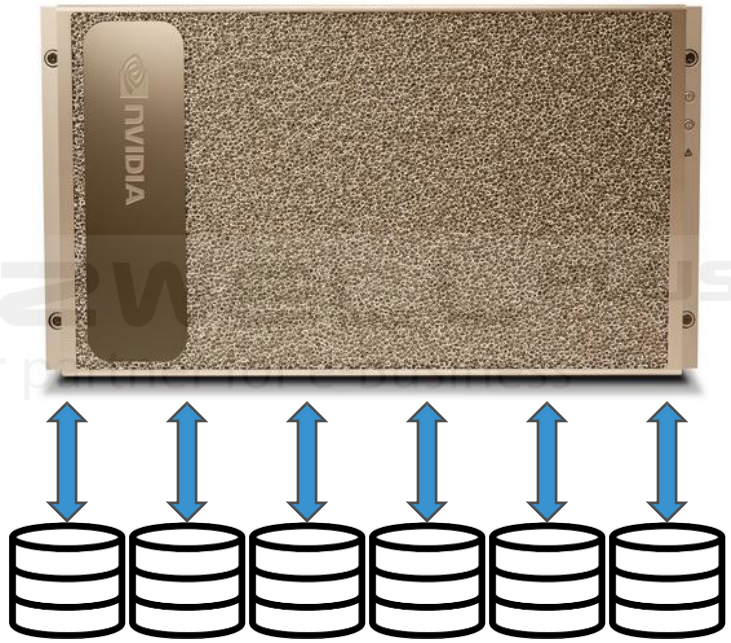
- 여러 서버가 마치 하나의 슈퍼컴퓨터처럼 작동하는 구조, **인피니밴드 네트워크**를 설계 및 구축해야 함

3. DGX A100 구축 가이드

✓ Enterprise 환경 A100 인프라 – AI 전용 고성능 NAS 구성

- Enterprise 대규모 프로젝트에서는 큰 DL 모델 학습에 필요한 대용량 데이터셋을 수용, 실시간으로 학습할 수 있는 **AI 맞춤형 스토리지** 필요





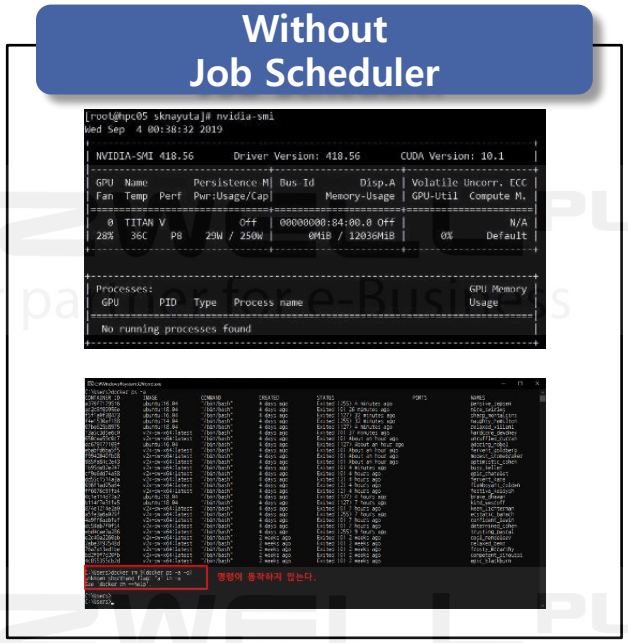
- DL 워크로드의 특성상 GPU-Storage 간 I/O가 동시다발적으로 발생, 병목을 효과적으로 수용할 수 있는 **병렬식 I/O 스토리지** 설정이 중요해짐

3. DGX A100 구축 가이드

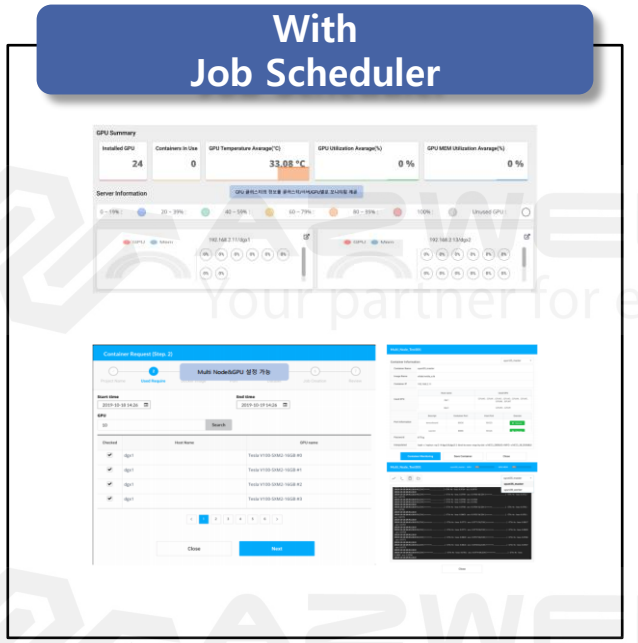
✓ GUI 기반 Job Scheduler 도입

<GPU 모니터링, Job 할당 및 회수>

Without Job Scheduler

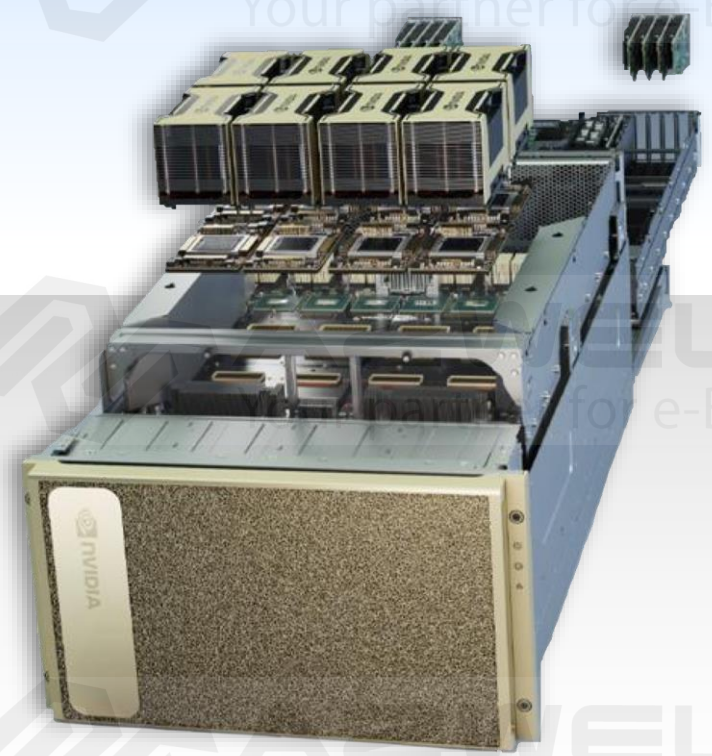


With Job Scheduler



- GPU 자원은 할당 및 회수가 빈번히 발생, 정해진 스케줄에 따라 Job을 사용자별로 사용현황을 파악, 할당하고 회수할 수 있는 자동화 시스템 필요
- 쉘 기반이 아닌 **GUI** 환경 구현으로 운영 편의성 제공(운영자가 꼭 **엔지니어**일 필요 X)

4. Why Azwell?



4. Why Azwell?

- NVIDIA 공식 **Elite Partner**
- DL 인프라는 단순히 HW만 좋아서 되는게 아님.
- 사용 용도에 따라 SW가 크게 달라지며, 모든 스택이 **최적화**되어야 함
- **DGX POD** 구성 래퍼런스 등 노하우를 기반으로, AI 인프라 구축을 컨설팅 단계에서부터 구축까지 **올인원 서비스** 제공가능
- Azwell은 **스토리지, 네트워크** 분야에 있어서도 국내 최고급의 기술력 및 엔지니어를 보유
- 많은 문의 주십시오!

NVIDIA ELITE PARTNER

AZWELL PLUS Your partner for e-Business

감사합니다