호실	AI 102호	호실명	AI Data Center Room
용도	고성능 컴퓨터 자원을 통한 AI 연구 활동 지원		

주요 기자재

1) AI-X Cluster
2) MobileX Digital Twin Server
3) HCI Server

- **AI-X Cluster (AI대학원 / 담당자: 김서인)**
  - AI 교육·연구 및 창의활동 프로그램 실습을 위해서 GPU 8장과 200G 및 100G 급의 네트워킹 카드를 보유한 고성능 GPU 서버인 NVIDIA DGX-A100 5대를 이용해 DevOps 기반 AI 컴퓨팅 클러스터 컴퓨팅 환경 구축
  - Slum 기반의 스케줄링 관리를 통해 저렴한 가격으로 대여
- **MobileX Digital Twin Server (AI대학원 / 담당자: 김서인)**
  - A10 GPU 10장이 탑재된 서버 및 L40 서버로 구성하여 Digital Twin 연구에 사용
- **HCI Server (AI대학원 / 담당자: 김서인)**
  - 꿈꾸는 Dream-AI LMS 서버로, AI대학원 자체 제작 콘텐츠 및 인공지능 관련 강의를 AI대학원 학생뿐만 아니라 외부에서 접근할 수 있도록 공유 및 개방

	A100 80GB PCIe	A100 80GB SXM
FP64	9.7 TFLOPS	
FP64 Tensor Core	19.5 TFLOPS	
FP32	19.5 TFLOPS	
Tensor Float 32 (TF32)	156 TFLOPS | 312 TFLOPS*	
BFLOAT16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
FP16 Tensor Core	312 TFLOPS | 624 TFLOPS*	
INT8 Tensor Core	624 TOPS | 1248 TOPS*	
GPU Memory	80GB HBM2e	80GB HBM2e
GPU Memory Bandwidth	1,935GB/s	2,039GB/s
Max Thermal Design Power (TDP)	300W	400W***
Multi-Instance GPU	Up to 7 MIGs @ 10GB	Up to 7 MIGs @ 10GB
Form Factor	PCIe dual-slot air cooled or single-slot liquid cooled	SXM
Interconnect	NVIDIA® NVLink® Bridge for 2 GPUs: 600GB/s ** PCIe Gen4: 64GB/s	NVLink: 600GB/s PCIe Gen4: 64GB/s
Server Options	Partner and NVIDIA-Certified Systems™ with 1-8 GPUs	NVIDIA HGX™ A100-Partner and NVIDIA-Certified Systems with 4,8, or 16 GPUs NVIDIA DGX™ A100 with 8 GPUs

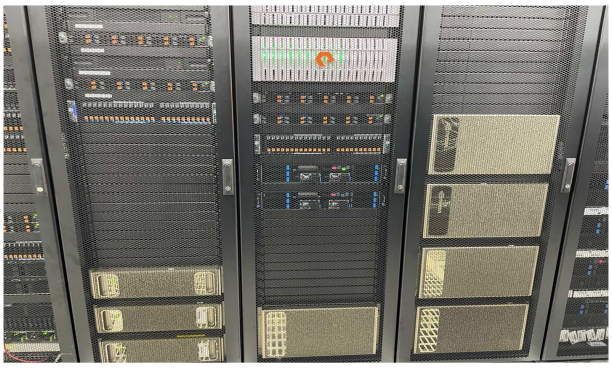
[A100 specification]

FP32	31.2 TF
TF32 Tensor Core	62.5 TF | 125 TF*
BFLOAT16 Tensor Core	125 TF | 250 TF*
FP16 Tensor Core	125 TF | 250 TF*
INT8 Tensor Core	250 TOPS | 500 TOPS*
INT4 Tensor Core	500 TOPS | 1000 TOPS*
RT Cores	72
Encode / Decode	1 encoder 2 decoders (+AV1 decode)
GPU Memory	24 GB GDDR6
GPU Memory Bandwidth	600 GB/s
Interconnect	PCIe Gen4: 64 GB/s
Form Factor	1-slot FHFL
Max TDP Power	150W
vGPU Software Support	NVIDIA vPC/vApps, NVIDIA RTX™ vWS, NVIDIA AI Enterprise
Secure and Measured Boot with Hardware Root of Trust	Yes (optional)
NEBS Ready	Level 3
Power Connector	PEX 8-pin

[A10 specification]

GPU Architecture	NVIDIA Ada Lovelace architecture
GPU Memory	48GB GDDR6 with ECC
Memory Bandwidth	864GB/s
Interconnect Interface	PCIe Gen4x16: 64GB/s bi-directional
NVIDIA Ada Lovelace architecture-based CUDA Cores	18,176
NVIDIA third-generation RT Cores	142
NVIDIA fourth-generation Tensor Cores	568
RT Core performance TFLOPS	209
FP32 TFLOPS	90.5
TF32 Tensor Core TFLOPS	90.5 | 181**
BFLOAT16 Tensor Core TFLOPS	181.05 | 362.1**
FP16 Tensor Core	181.05 | 362.1**
FP8 Tensor Core	362 | 724**
Peak INT8 Tensor TOPS	362 | 724**
Peak INT4 Tensor TOPS	724 | 1448**
Form Factor	4.4" (H) x 10.5" (L) - dual slot
Display Ports	4 x DisplayPort 1.4a
Max Power Consumption	300W
Power Connector	16-pin
Thermal	Passive
Virtual GPU (vGPU) software support	Yes
vGPU Profiles Supported	See Virtual GPU Licensing Guide¹
NVENC | NVDEC	3x | 3x (Includes AV1 Encode & Decode)
Secure Boot with Root of Trust	Yes
NEBS Ready	Level 3
MIG Support	No
NVLink Support	No

[L40 specification]



[AI대학원 Server 장비]