



HPE - NVIDIA - 동국시스템즈

# AI Solution Day

2024년 4월 9일(화), 09:00~13:30

그랜드 인터컨티넨탈 파르나스, 로즈(5F)



# AI에 최적화된 플랫폼 아키텍처 소개

송규태 (동국시스템즈)

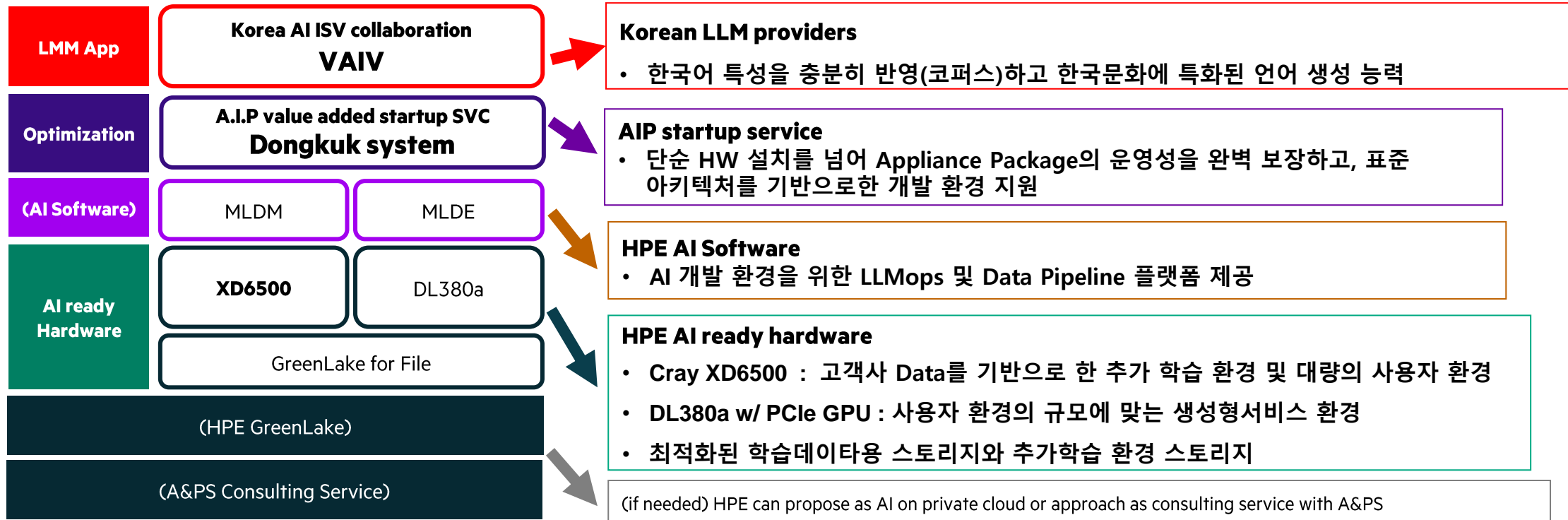
# LLM Package Appliance

“Build your own private LLM in your datacenter less than xx원”

- All in one package solution to build LLM app in datacenter

→ DL380 (생성형서비스) + framework start up service + Korean LLM model (VAIV)

## HPE Gen AI Package Stacks



# LLM Package Appliance

## Appliance Detail

### Model 소개



**v1.3B**  
**v5.8B**  
**v12.8B**

- 국내외 유명 Open source LLM 기반
- Task Oriented & Domain knowledge based
- 한국어 특화 모델

- **VAIV GeM**  
On-prem 기반의 Private LLM 구축

- **VAIV SearchGPT**  
RAG기반 질의 응답 서비스

- **VAIV SmartChat**  
LLM 기반 챗봇 서비스

- 12.8B 이상 대형 모델 + Finetune / Pretrain
- Prompt를 통한 다양한 Task 지원
- 데이터 전처리 및 Training 지원
- 컨설팅 및 결과보고서 포함

- Private data를 통한 RAG 및 Inferencing 구축
- API 기반 서비스 연동
- 데이터 수집 및 전처리 지원

- 일상 대화 및 시나리오 대화 기능
- VAIV SearchGPT 연동
- OpenAPI 기반 연계 서비스 제공

### HW spec by CCU

#### <50 CCU>

- **HPE ProLiant DL380a w/ 4x L40S 48GB**
- Intel Gold 6430 32c with 512GB RAM
- NVIDIA ConnectX-7 NDR 400Gb 1port
- 3-year Tech Care Basic

#### <150 CCU>

- **HPE ProLiant DL380a w/ 4 x H100 94GB**
- Intel Gold 6430 32c with 1TB RAM
- NVIDIA ConnectX-7 NDR 400Gb 1port
- 3-year Tech Care Basic

#### <300 CCU>

- **HPE Cray XD670 w/ H100 SXM5 8 way**
- Intel Platinum 8462Y+40c with 1TB RAM
- NVIDIA ConnectX-7 NDR 400Gb 1port
- 3-year Tech Care Basic

### Pretrain Package

- **HPE Cray XD670 w/ H100 SXM5 8 way**  
- 고객사 요구사항(구축기간, 데이터셋 규모)에 따른 가변적 클러스터 구성

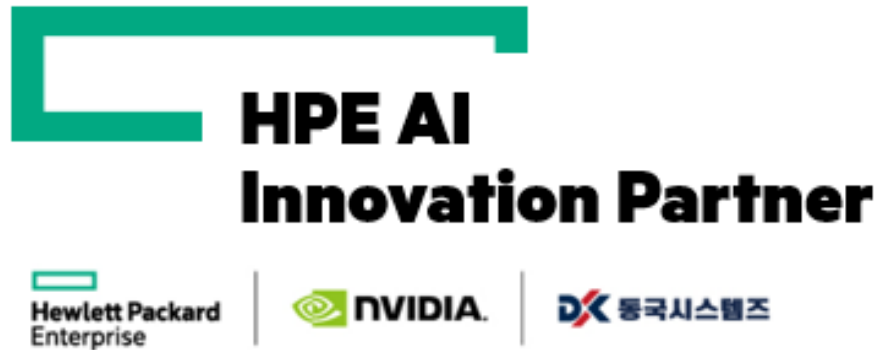
- **MLDE as LLMOps**  
- 다양한 모델의 필요에 따라 LLMOps 환경 구축 필요시 선택적 제안

- **GLFS as AI storage**  
- sLM check point saving 성능 충족  
- Model 인프라 Bursting 시 탁월한 성능 제공

# Detail Configuration

Category	Concurrent user	Server Model	Memory	CPU	Disk	GPU	Nic	Transceiver	Power	Management	Service
RAG only/ Chat Bot/	50 user	DL380a	32GB x 16 (512GB)	32-Core x 2 (Gold 6430)	-1.92TB SSD x 2 (OS) -Array Controller : MR416i-o - 7.68TB SSD x 5	L40S 48GB 4unit	-NDR 1port 400G x 2 -10G 4port x1 - 1G 4port x1	NDR/Ethernet 1x400Gb	1800- 200W x 4	Compute Ops	Tech care Basic 3y
	150 user	DL380a	64GB x 16 (1024GB)	32-Core x 2 (Gold 6430)	-1.92TB SSD x 2 (OS) -Array Controller : MR416i-o - 7.68TB SSD x 5	H100 94GB PCIe 4unit	-NDR 1port 400G x 2 -10G 2port x1 - 1G 4port x 1	NDR/Ethernet 1x400Gb	1800- 200W x 4	Compute Ops	Tech care Basic 3y
	300 user	XD670	64GB x 16 (1024GB)	40-Core x 2 (P-8462Y+)	- 1.92TB M.2 x 2 (Samsung PM9A3) -Gigabyte 4p M.2 Adptr(RAID) - 7.68TB U.3 SSD x 5	H100 SXM 8 way	-NDR 1port 400G x 4 -10G & 1G Embedded	NDR/Ethernet 1x400Gb	-	-	Tech care Basic 3y
Pretrain용 추가장비	Computing Node	XD670	64GB x 16 (1024GB)	40-Core x 2 (P-8462Y+)	-1.92TB M.2 x 2 (Samsung PM9A3) -Gigabyte 4p M.2 Adptr(RAID) - 7.68TB U.3 SSD x 5	H100 SXM 8 way	-NDR 1port 400G x 4 -10G & 1G Embedded	NDR/Ethernet 1x400Gb	-	-	Tech care Basic 3y
	Storage GLFS	GLFS 4C-5D-2S		R104GB / W16GB		C Node – 4 / D node - 5		U667TB			
		GLFS 8C-10D-6S		R208GB / W32GB		C node – 8 / D node – 10		U1.35PB			
		GLFS 23C-32D-12S		R599GB / W102GB		C Node – 23 / D Node – 32		U4.3PB			
LLM OPs	MLDE	To be updated									














# A.I.P (AI Innovation Partner)



- AI ISV Partner Program
- GPU Tech Day
- GPU Stock Biz

# A.I.P (AI Innovation Partner)

## AI ISV Partner Program

 <p><b>넥타르소프트</b> Nectarsoft Co.,Ltd</p> <p>AIoT 사고 예방, 음성인식</p> <p><a href="#">바로가기</a></p>	 <p><b>바이브컴퍼니</b> VAIV Company inc</p> <p>생성형 AI, 검색/챗봇</p> <p><a href="#">바로가기</a></p>	 <p><b>테스트웍스</b> Testworks</p> <p>영상 인식, 굿데이터 생성</p> <p><a href="#">바로가기</a></p>
 <p><b>부들정보시스템</b> Buttle Information Systems Co., Ltd.</p> <p>RPA, 상담사 관리, 대화형 AI 봇</p> <p><a href="#">바로가기</a></p>	 <p><b>스칼라웍스</b> ScalaWox</p> <p>실시간 영상 분석</p> <p><a href="#">바로가기</a></p>	 <p><b>한솔인티큐브</b> Hansol Inticube</p> <p>긴급출동 음성봇, 챗봇</p> <p><a href="#">바로가기</a></p>
 <p><b>씨에스리</b> CSLEE</p> <p>노코드 데이터분석</p> <p><a href="#">바로가기</a></p>	 <p><b>씨이랩</b> XIIILAB</p> <p>AI 영상 분석 플랫폼, GPU 관리</p> <p><a href="#">바로가기</a></p>	 <p><b>티쓰리큐</b> T3Q</p> <p>인공지능/빅데이터 분석 시스템</p> <p><a href="#">바로가기</a></p>
 <p><b>알체라</b> Alchera Inc</p> <p>비대면 신원 확인, 얼굴인식</p> <p><a href="#">바로가기</a></p>	 <p><b>엠아이큐브솔루션</b> MICUBE Solution, Inc.</p> <p>공정 최적화, 불량 사진예측</p> <p><a href="#">바로가기</a></p>	
 <p><b>위세아이텍</b> WISEITECH CO.,LTD</p> <p>머신러닝 자동화 플랫폼</p> <p><a href="#">바로가기</a></p>	 <p><b>인이지</b> INEEJI Corp</p> <p>공정 최적화 AI 예측</p> <p><a href="#">바로가기</a></p>	

- 13 개 분야별 ISV
- 2 개사 조인 완료 단계
- '24내 20개사 확대 예정



# A.I.P (AI Innovation Partner)

## GPU Teck Day

### HPE-NVIDIA-동국시스템즈 GPU Tech Day

2023년 3월 23일(목) 09:30 - 17:00 | 페럼타워 2층 세미나실

#### Agenda

구분	Time	Session	Speaker
개념교육	09:30 - 10:30	GPU의 이해 • GPU와 CPU의 차이 • 그래픽을 위한 GPU가 AI업무에 필수가 된 이유 • NVIDIA GPU 특성 소개	송규태 팀장 동국시스템즈
	10:30 - 11:30	AI Workload의 이해 • 기본적인 AI 이해, AI 업무의 특성 소개 • Data 분석 및 AI 인프라 구성 시 검토/주의사항	송규태 팀장 동국시스템즈
	11:30 - 11:45	휴식 시간	
	11:45 - 12:45	GPU 상세 • GPU 서버 구축과 관련한 상세 기술정보 • GPU Architecture 이해	송규태 팀장 동국시스템즈
	12:45 - 13:45	점심 시간	
Hands_on	13:45 - 14:45	O/S 구성 • Ubuntu 설치 / Driver 설치	라성균 부장 동국시스템즈
	14:45 - 15:45	NVIDIA Library 설치 • CUDA / cuDNN 설치	라성균 부장 동국시스템즈
	15:45 - 16:00	휴식 시간	
	16:00 - 17:00	ML Tool 설치 • Python / Jupyter 설치	라성균 부장 동국시스템즈

### HPE-NVIDIA-동국시스템즈 GPU Tech Day

일시 : 2023년 6월 26일(월) 09:30 - 17:00  
장소 : 페럼타워 2층 세미나실

#### Agenda

Time	Session	Speaker
	<b>개념 교육</b> <b>GPU Server 고려사항</b> - Geforce vs Tesla, PCIe vs NVLink, vGPU vs MIG - NVIDIA GPU Generation - 분산 컴퓨팅	
09:30 - 12:30	<b>NVIDIA 제품 동향</b> - NVIDIA 기술 동향 - NVIDIA AI Enterprise 소개	
	<b>HPE AI 동향</b> - 영입지원을 위한 AI ISV Partnership Program 소개 - HPE Ezmeral, MLDE 소개	
12:30 - 13:30	점심식사	
	<b>Hands on</b> <b>NVIDIA Enterprise 데모 시연</b> - Creating NVIDIA AI Enterprise System - Installing Red Hat Enterprise Linux 8.4 - Install Steps for DLS Scenario	동국시스템즈 라성균 부장
13:30 - 17:00	<b>RAPIDS Installatfon 데모 시연</b> - Kubernetes - RAPIDS on Databricks - RAPIDS on Google Colab	

### HPE-NVIDIA-동국시스템즈 AI를 위한 핵심 GPU 테크데이

일시 : 2023년 10월 24일(화), 14:00 - 17:10  
장소 : 동국시스템즈 본사 (페럼타워 9층) - 다빈치 A

#### Agenda

Time	Session	Speaker
14:00 - 16:00	<b>AI를 위한 Power &amp; Cooling 설계</b> - 전력 산출과 대용량 전력 소요에 따른 장비 배치 - 다양한 Cooling 방식과 HPE 장비의 Cooling 설계	동국시스템즈 송규태 팀장
16:00 - 16:10	휴식시간	
16:10 - 17:10	<b>AI를 위한 Software</b> - HPE Ezmeral 과 MLDE / MLDM 소개 및 Demo	동국시스템즈 라성균 부장

### HPE-NVIDIA-동국시스템즈 AI 테크데이

일시 : 2024년 1월 24일(수) 9:30-13:30

장소 : 동국시스템즈 본사(페럼타워 9층) - 다빈치 A [약도보기](#)

[사전 등록하기](#)

#### 안녕하십니까?

'HPE-NVIDIA-동국시스템즈의 AI 테크데이'에 초대합니다.

새해인 2024년을 맞아, 이번에는 올해 가장 큰 화두인 AI에 초점을 맞춘 테크데이를 마련했습니다. 작년 GPU 테크데이에 이은 첫번째 AI 테크데이에서는 AI 핵심 주제들을 살펴보고자 합니다.

AI 기초 지식부터 LLM(Large Language Model), Inference(추론) 등의 주제를 다룰 예정이며, 현재의 AI 트렌드를 파악하는 데 도움이 되리라 생각합니다. HPE 파트너사의 세일즈/프리세일즈분들의 많은 관심과 등록 부탁드립니다.

감사합니다.

#### Agenda

Time	Session	Speaker
09:30-10:30	<b>GPU와 AI 이해</b>	
	<b>AI Infra 이해</b> - Foundation Model / Fine Tuning / LLM / Inference / Data Platform +HPE-NVIDIA AI 시장 전략 소개	동국시스템즈 송규태 팀장
11:30-12:30	Networking Lunch	
12:30-13:30	<b>AI ISV Partner Program 소개</b>	





HPE - NVIDIA - 동국시스템즈

**AI Solution Day**

**THANK YOU**