Hewlett Packard Enterprise | NVIDIA | DX 동국시스템즈

HPE - NVIDIA - 동국시스템즈

AI Solution Day

2024년 4월 9일(화), 09:00~13:30

그랜드 인터컨티넨탈 파르나스, 로즈(5F)

Hewlett Packard Enterprise | NVIDIA | DX 동국시스템즈

HPE - NVIDIA - 동국시스템즈
AI Solution Day

AI에 최적화된 플랫폼 아키텍처 소개

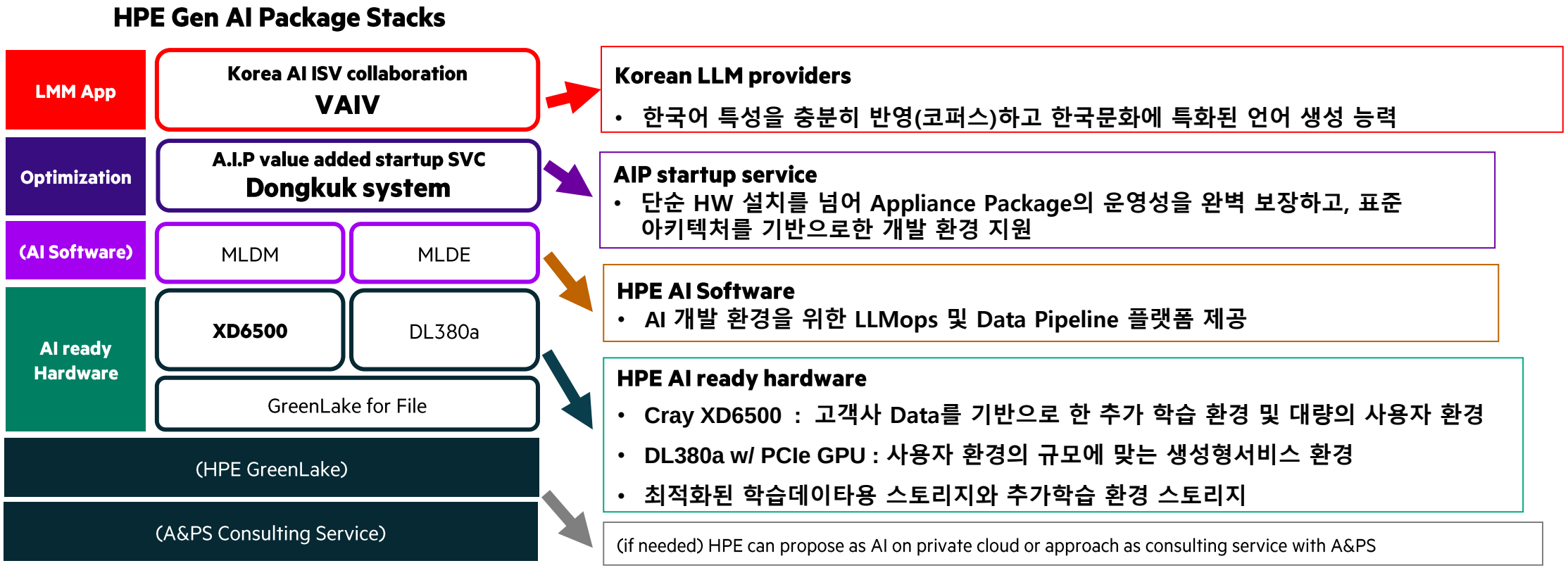
송규태 (동국시스템즈)

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)



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

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

- **VAIV SearchGPT**
  RAG기반 질의 응답 서비스

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

- **VAIV SmartChat**
  LLM 기반 챗봇 서비스

- 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)

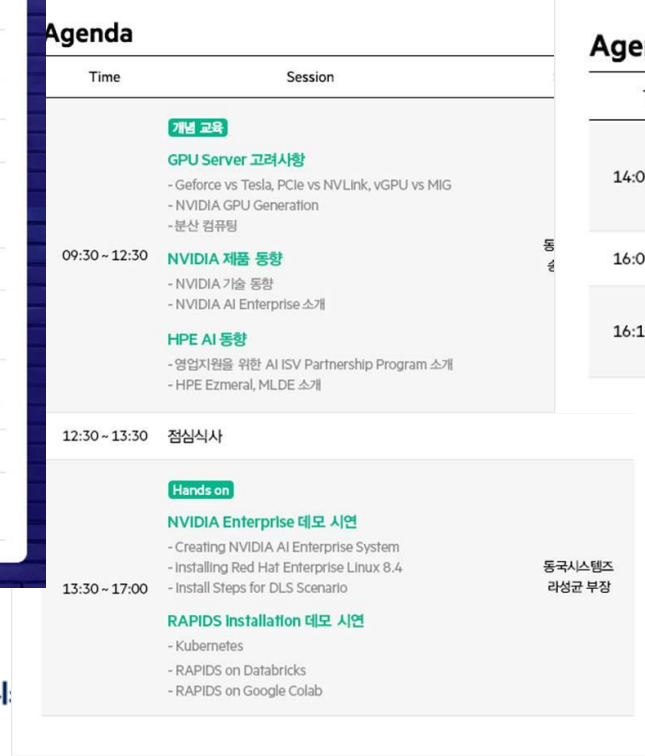
AI ISV Partner Program

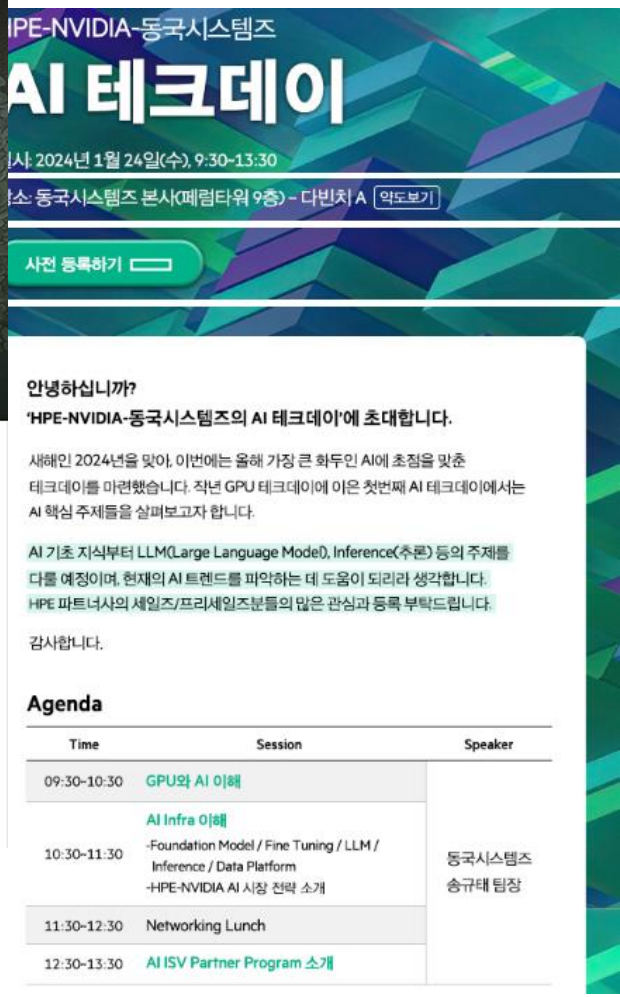
넥타르소프트 Nectarsoft Co.,Ltd	AIoT 사고 예방, 음성인식	바로가기
바이브컴퍼니 VAIV Company inc	생성형 AI, 검색/챗봇	바로가기
테스트웍스 Testworks	영상 인식, 굿데이터 생성	바로가기
부뜰정보시스템 Buttle Information Systems Co., Ltd.	RPA, 상담사 관리, 대화형 AI 봇	바로가기
스칼라웍스 ScalaWox	실시간 영상 분석	바로가기
한솔인티큐브 Hansol Inticube	긴급출동 음성봇, 챗봇	바로가기
씨에스리 CSLEE	노코드 데이터분석	바로가기
씨이랩 XIILAB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

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

A.I.P (AI Innovation Partner)

GPU Teck Day





Hewlett Packard Enterprise | NVIDIA | DX 동국시

HPE-NVIDIA-동국시스템즈
AI Solution Day

Hewlett Packard Enterprise | NVIDIA | DX 동국시스템즈

HPE - NVIDIA - 동국시스템즈

AI Solution Day

THANK YOU