

AiNA 산업융합위원회 제9회 정례회의 개최 결과

('26.06.22.(월), AINA 사무국)

I. 회의 개요

- (일시/장소) '26.05.21.(목), 14:00~17:00/AWS 코리아 사옥(센터필드 12층)
- (논의내용) AWS 및 관련 기업(브로드컴, 퍼블릭에이아이) 전문가 세미나
- (참 석 자) 산업융합위원회, 관련 기업, 사무국 등 26명
 - 김태운 책임, 이창수 책임, 성태현 책임, 국수진 선임(이상 LG전자), 장준희 책임(NIA), 김동구 교수(연세대), 정희상 실장(ETRI), 박영완 상무(퀄컴), 오충근 수석(TTA), 고병조 매니저, 송선석 상무(이상 키사이트), 이라미 부문장, 박정민 본부장(이상 LIG Accuver), 전형준 연구소장(유캐스트), 고의준 총괄, 안수호 매니저, 장석재 매니저(이상 AWS), 최양수 전무(이루온), 박은복 상무(진인프라), 양대영 이사(Newgens), 정상철 상무(브로드컴), 김성중 대표(퍼블릭에이아이), 황인수 상무(레드햇), 송광묵 이사, 박성혁 상무이사(이상 우리넷), 김현수 과장(KANI),

II. 주요 결과

□ 5G Core on AWS : NTT Use case : 안수호 상무(AWS)

발표자료 비공개

- 통신업계 현황과 과제
 - (트렌드 전환) 클라우드 마이그레이션 → 소버린 클라우드·에이전트 AI로 화두 이동. 장애·보안사고의 근본 원인이 '사람의 개입'인 만큼, 제로 트러스트 기반 제로 터치 자동화가 핵심 과제로 부상
 - (데이터 파편화 문제) C360 다면 분석을 시도하나 고객센터·CRM·데이터 허브에 흩어진 데이터가 정제되지 않아 실사용 가능한 데이터가 부족한 상태(데이터 파운데이션 구축이 자동화의 전제 조건)

- (국내외 접근 방식 차이) 국내는 BSS 전면 교체를 시도하는 경향, 해외(BT, NTT 도코모)는 작은 워크로드부터 단계적으로 검증하며 역량을 쌓는 방식으로 성과를 내는 중

○ 해외 혁신 사례

- (BT, 브리티시 텔레콤) 수작업 과다·데이터 사일로·레거시 혼재 문제 해결을 위해 데이터 도메인화·클러스터링 기반 지식 기반 구축 : 이상 탐지·RCA·서비스 영향도 분석 자동화로 비용 50% 절감, 변화 속도 25% 개선
- (NTT 도코모) 멀티벤더 사일로 및 복잡한 인시던트 RCA 한계를 해결하기 위해 네트워크 데이터를 토폴로지화·지식 기반화하여 에이전트 AI 학습에 활용 : 특정 지역 5G 코어를 AWS에 올려 검증, 전력 소비 70% 절감 달성
- (공통 시사점) 두 사례 모두 '에이전트 AI를 쓰기 위해 클라우드로 간 것'이 아니라, 데이터·인프라 기반을 유연하게 만들어 놓으니 에이전트 AI가 작동할 수 있는 운동장이 자연스럽게 만들어진 것

○ 주요 질의응답(Q&A) 정리

- Q1) 에이전트 AI 활용을 위해 코어망 가상화와 오케스트레이션 플랫폼이 필요한데, AWS에 관련 기술이 있는가?
- A1) 서비스 포트폴리오는 준비되어 있으며, 그래프 DB 기반 구성 시 효율성이 높음, 국내는 법·컴플라이언스 이슈가 더 큰 장벽으로 현재 OSS 차세대화→DR→인증·IoT→코어망 순으로 단계적 접근 중
- Q2) AI-RAN 위주 논의에서 지능형 코어로도 관심을 넓혀야 하고, AWS와 협력 시 비용 부담도 크다는 의견이 있음
- A2) 코어·클라우드 환경에서의 에이전트 성능·에너지 절감 실증이 중요하다는 데 공감함, 얼라이언스 가입 취지에 맞게 비용 부담 완화 및 협력 방안을 지속 강구할 것임

□ OutPost 소개 및 사례 공유 : 장석재 매니저(AWS)

상세내용 발표자료 참고

○ Outposts 개요 및 구조

- (전진 배치 개념) 지연 시간에 민감하거나 현장 데이터를 로컬 처리해야 하는 워크로드를 위해 AWS 리전의 랙을 고객 데이터 센터에 설치하는 완전관리형 서비스, 하드웨어 관리는 AWS가 전담하며 리전과 동일한 콘솔·API 사용 가능
- (네트워크 구조) 로컬 네트워크(현장 장비 연결)와 서비스 링크(리전 VPN 연결) 두 개로 구성, 고객 데이터는 리전으로 전송되지 않으며 장비 운영 상태정보만 암호화 전송

○ 세대별 주요 변화

- (1→2세대 전환) 단일 랙에서 네트워크 랙·컴퓨트 랙 분리 구조로 변경, 최대 24개 랙까지 스케일아웃 가능. 최신 인스턴스 탑재로 1세대 대비 최대 40% 성능 향상
- (특화 인스턴스 추가) 통신사 5G 코어·미디어 배포용 고대역폭 인스턴스(최대 800Gbps)와 금융·거래소용 마이크로초 단위 초저 지연 인스턴스가 2세대에서 신규 제공
- (서비스 범위 제한) 컴퓨트·스토리지·컨테이너·DB 등 핵심 서비스는 로컬 구동 가능. Bedrock·SageMaker 등 AI 서비스는 Outposts 에서 직접 구동 불가, 리전 연동 방식 사용 필요

○ 국내 공공·연구 도입 사례

- NIPA 5G MEC(판교)/GERI(구미) : VR/AR 테스트 환경의 서버 용도로 도입, 판교는 초기 지연 목표 미달(SKT 장비 이슈), 구미 Nokia 5G 연동에서 10ms 이하 안정 확보
- AICA(광주) : NPU 서버와 Outposts 간 연동으로 상호운영성 실증, 1년 단기 계약으로 도입

- 국토부 자율주행 과제 : 랙 2대를 3년 계약으로 운영 중(각 월 약 \$14K~14.6K), 연구·실증 단계는 현행 인증으로 진입 가능하나 과제 종료 후 상용 전환 시 보안 인증 재검토 필요

○ 주요 질의응답(Q&A) 정리

- Q1) 서비스 링크로 고객 데이터가 유출되지 않는지? EKS 마스터 노드 위치는?
- A1) 고객 데이터는 리전으로 전송되지 않으며 장비 운영 상태정보만 암호화 전송. EKS는 리전에 컨트롤 플레인을 두는 기본 모드와, 컨트롤 플레인까지 로컬에 배치하는 '로컬 클러스터 모드' 두 가지 지원
- Q2) CSAP 인증 범위와 비용 지원 방안은?
- A2) AWS 서울 리전은 CSAP 하(다) 등급 보유, 연구 과제는 현인증으로 진입 가능하나 상용 전환 시 재검토 필요, 할인 프로그램·PoC 크레딧 등 내부 지원 방안이 있어 과제 기획 단계부터 AWS 영업팀 협의 권장

□ UEC 관련 Ethernet Scale Up : 정상철 상무(브로드컴)

발표자료 비공개

○ 브로드컴 제품 포트폴리오

- (사업 구조) 매출의 약 60%는 반도체(네트워크 실리콘), 40%는 소프트웨어(VMware 등), 엔비디아와 함께 반도체 업체 중 70%대 영업이익률을 유지하는 극소수 기업
- (공급 범위) 데이터센터 내 스위치·NIC·광 트랜시버까지 포괄하는 풀스택 포트폴리오로 데이터센터 진입부터 최종 단말까지 모든 경로를 거침
- (경쟁 구도) 스위치 시장에서 엔비디아와 직접 경쟁. 브로드컴은 이더넷 오픈 스탠다드, 엔비디아는 독점 프로토콜(NVLink·InfiniBand)로 각각 시장 공략

○ 이더넷의 재정의 : Tomahawk Ultra (Now Shipping)

- (목표) "이더넷은 느리고 손실이 많다"는 시장 고정관념을 하드웨어 혁신으로 완전히 타파, 초저지연·무손실·저오버헤드를 동시에 실현
- (AI 스케일업 지원(SUE)) NVLink 대비 최대 4배 많은 XPU를 연결 가능한 오픈 스탠더드 패브릭, 집합 연산(AllReduce 등)을 스위치 내 엔진에서 직접 처리해 AI 작업 완료 시간 단축
- (신뢰성) 링크 레이어 재전송(LLR)과 크레딧 기반 흐름 제어(CBFC)를 하드웨어에서 구현하여 대규모 트래픽 환경에서도 패킷 유실 원천 차단

○ XPU(맞춤형 AI 가속기) 시장 및 글로벌 동향

- (XPU 시장 전망) 구글·메타·AWS 등 글로벌 빅테크 10여 곳이 브로드컴 ASIC 기반 자체 XPU를 양산 중, 내년 기준 XPU 시장이 GPU 시장을 초과할 전망으로, 예측 시점이 계속 앞당겨지고 있음
- (국내 동향) 네이버·카카오 등이 자체 반도체 테스트 및 다변화 진행 중, 도입 지연의 주요 원인은 하드웨어가 아닌 엔비디아 CUDA 중심의 소프트웨어 전환 장벽

○ 데이터센터 간 연결(DCI)

- (Jericho4(Now Shipping)) 100km+ 장거리 데이터센터 간에도 손실 없는 데이터 전송 지원. 장거리 구간에서 발생하는 대량의 인-플라이트 데이터를 처리하기 위해 덤 버퍼 탑재
- (국내 적용 조건) 데이터센터 내 GPU/XPU 최소 1만 개 이상 보유 시 경제적 효용 발생. 현재 국내 사례는 없으나 투자 확대에 따라 도입 논의 예상

○ 주요 질의응답(Q&A) 정리

- Q1) PCIe 스위치 기반 스케일업과의 차이는?

- A1) PCIe는 처리 오퍼레이션이 많아 레이턴시가 높고 확장성에 한계가 있음, 구글·메타·AWS 등은 PCIe와 별도로 전용 스케일업 패브릭을 독립 구성하여 사용
- Q2) CXL 기술 적용 계획은?
- A2) 구글 등 핵심 고객사가 CXL 미채택 입장을 밝혔고, 브로드컴도 최신 PCIe 스위치에서 CXL 관련 기능을 삭제함
- Q3) 국내 오픈 이더넷 도입이 더딘 이유는?
- A3) 하드웨어 성능이 아닌 소프트웨어 전환 장벽, 글로벌 탑티어는 자체 SW 스택을 보유해 전환이 용이하지만 국내 기업은 CUDA 의존도가 높음

□ 폐쇄망 LLM 기반 AI Agent 운영 인프라 부터 Application까지 소개:
김성중 대표(퍼블릭에이아이) **발표자료 비공개**

○ 회사 개요 및 솔루션 체계

- 회사 개요: 기업 환경에 최적화된 폐쇄망 기반 AI 시스템 전문 기업, RAG-LLM 분야 국내 유일 TTA 신뢰성 인증 보유 국세청·코트라·하이닉스 등 공공·민간 다수 구축 실적
- Devstream (GPU 인프라 관리): CPU·GPU 자원 분할 및 상세 모니터링, 하나의 GPU를 여러 프로세스에 효율적으로 배분하는 자동 슬라이싱 제공, 별도 엔지니어 수동 지정 없이 자원 최적화 가능
- AIMP (MLOps): AI 모델 학습·배포·평가·모니터링을 통합 관리. 데이터 누적에 따른 모델 자동 고도화 및 재학습 지원, 기존 AI 시스템이 사업 완료 후 고도화 시 별도 발주가 필요한 문제를 해결
- LIFI (LLM Ops / AI Agent): 폐쇄망·공개망 LLM Agent 통합 관리, 다중 Agent 협업 설계 및 팀별 권한 제어, AI Agent 운영 시 최대 80개 이상의 sLLM 모델을 자동으로 운용

○ 핵심 기술 및 운영 전략

- **RAG + AI Agent 전략**: 무리한 파인튜닝보다 RAG 시스템과 다중 Agent 오케스트레이션이 효과적, LLM의 컨텍스트가 길어지면 중간 정보가 소실되는 한계를 작업 분산으로 극복
- **할루시네이션 대응**: 할루시네이션은 LLM 고유 특성으로 접근, RAG 시스템으로 사실 오류를 줄이고, 가드레일로 부적합 출력을 차단, 데이터 접근 단계의 권한 제어(KMS)와 결합이 최근 중요해지고 있음
- **LLM to Physical AI (LAM)**: 언어 인지·추론에 그치지 않고 실제 기기(PLC/밸브 등)를 제어하는 Large Action Model로 확장
- **MCP(Model Context Protocol)**: 서로 다른 앱·API를 LLM으로 표준화하여 제어, Agentic LLM 구현의 핵심 인터페이스

○ 주요 구축 사례

- **국세청 LLM 기반 AICC**: 세법 기반 RAG 시스템으로 전화 상담 자동화, 개인정보 정규식 변환·WAF 보안·API 인증키 암호화 적용
- **코트라 수출비서**: 폐쇄망 LLM + Reasoning & Action AI Agent, 기관 내 분산 데이터 일괄 처리 및 미활용 데이터 활용 전환
- **KLID 지방재정 업무 혁신**: 자연어 질의로 지자체 재정정보를 조회·즉시 시각화, 기존 필터 검색 대비 활용성 증대, 자동 레포트 작성으로 확대
- **하이닉스 My비서 Agent**: Supervisor Agent가 메시지 요약·회의실 예약·담당자 조회·지식 조회 등 4개 전문 Agent를 오케스트레이션하는 AI Agentic RAG 시스템

○ 주요 질의응답(Q&A) 정리

- **Q1)** 비즈니스 모델이 클라우드인지, 온프레미스인지?
- **A1)** 특정 환경을 가리지 않음, 고객이 필요로 하는 형태(폐쇄망 구축형·클라우드·하이브리드)에 맞춰 최적 시스템을 구축
- **Q2)** 공공기관 한글 문서를 AI 에이전트에 활용할 수 있는지?

- A2) 가능함, HWPX는 XML 구조로 이루어져 있어 데이터 파싱 및 정보 추출에 문제 없음
- Q3) 소규모 PoC 규모와 기간은
- A3) 활용할 Tool의 복잡도와 범위가 업무마다 크게 다름, 구체적인 상담을 통해 범위를 산정이 필요

III. 회의 전경



(회의 배경 및 안건 소개)
AINA 김태운 산업융합위원장(LG전자)



(5G Core on AWS : NTT Use case)
안수호 상무(AWS)



(OutPost 소개 및 사례 공유)
장석재 매니저(AWS)



(UEC 관련 Ethernet Scale Up)
정상철 상무(브로드컴)



(전체 전경)



(단체 사진)