CXL 4.0 표준 기반 메모리 풀링 아키텍처와 AI 팩토리 자원 공유 표준화 동향

안후영 한국전자통신연구원 책임연구원

1. 머리말

AI 기술 발전과 LLM(거대언어모델, Large Language Model) 및 생성형 AI 서비스 상용화는 데이터센터 인프라의 패러다임 전환을 요구하고 있다. 기존 범용 서버 중심 데이터센터는 고밀도·고효율 연산을 전담하는 'AI 팩토리'로 진화하고 있고, 동시에 AI 추론 워크로드가 학습과 함께 주요 부하로 부상하면서 대용량·고대역폭 메모리 수요가 기존 서버 아키텍처의 수용 한계를 빠르게 초과하고 있다. GPU/NPU 등 AI 가속기의 연산 성능은 비약적으로 향상됐으나, 장치에 탑재된 온보드 메모리만으론 급증하는 메모리 용량 수요를 충족하기 어려우며, 메모리 자원의 단편화와 활용 비효율 문제 또한 심화되고 있다.

이러한 문제를 해결할 핵심 기술로 주목받는 것이 CXL(Compute Express Link)[1] 기반 메모리 풀링(Memory Pooling)이다. CXL은 PCIe 물리 계층 기반의 캐시 일관성을 보장하는 인터커넥트 표준으로, 2019년 CXL 1.0 출시 이후 꾸준히 발전해 2025년 11월 PCIe 7.0 기반의 CXL 4.0으로 진화했다. CXL 4.0은 128 GT/s 전송 속도와 번들 포트(Bundled Port)를 통한 1.5 TB/s 대역폭을 제공하며 다중 랙 규모의 메모리 풀링을 현실화하는 중요한 이정표를 세웠다.

이번 원고에선 CXL 표준 발전 과정과 CXL 4.0 핵심 기술을 정리하고, AI 팩토리 환경에서의 메모리 병목 해소 방안 및 국내외 표준화 동향, 그리고 한국형 AIDC(AI Data Center) 메모리 인프라 표준화 방향을 제언한다.

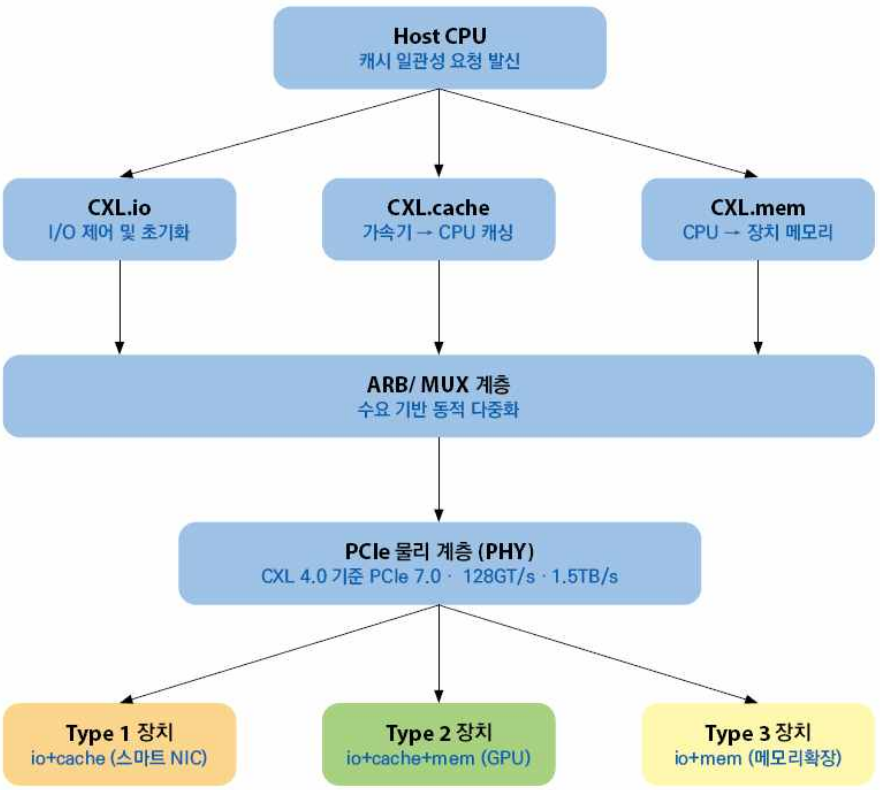
2. 배경 지식

2.1 CXL 개요

데이터센터는 CPU, GPU, FPGA, NPU 등 이기종 프로세서가 공존하는 환경으로 전환되고 있으나, 기존 PCIe 인터페이스는 캐시 일관성을 보장하지 않아 가속기 간 데이터 공유 시 복사·동기화 오버헤드가 불가피했다. 이를 극복하고자 인텔(Intel) 주도하에 주요 기업들이 2019년 CXL 컨소시엄을 설립했으며, CXL은 PCIe 물리 계층을 활용하면서 캐시 일관성을 보장하는 개방형 인터커넥트 표준으로 PCIe 인프라와 하위 호환성을 유지하고 있다.

[그림 1]과 같이 CXL 트랜잭션 계층은 세 가지 서브 프로토콜(CXL.io, CXL.cache, CXL.mem)로 구성되며, ARB/MUX 계층이 수요에 따라 프로토콜 간 자원을 동적으로 조정한다. CXL.io는 I/O 제어 프로토콜로 모든 디바이스가 필수 지원하며, CXL.cache는 가속기의 캐시 일관성 있는 호스트 메모리 접근을, CXL.mem은 CPU의 디바이스 메모리 직접 접근을 지원하는 메모리 풀링의 핵심

프로토콜이다. 세 프로토콜은 PCIe PHY를 공유하며 디바이스 유형(Type 1~3)에 따라 선택적으로 조합 적용된다. CXL 메모리는 DRAM(DDR5 기준 약 60~100 ns) 대비 소폭 높은 접근 지연(약 150~400 ns)을 가지나, 수십~수백 TB 규모의 용량 확장이 가능하고 NUMA 노드로 OS에 인식돼 기존 소프트웨어 스택과 호환성을 유지한다는 점이 핵심 강점이다.



[그림 1] CXL 서브 프로토콜 종류와 CXL 장치 유형

2.2 메모리 풀링 개요

메모리 풀링이란 다수의 호스트가 CXL 스위치를 통해 공유 메모리 자원을 동적으로 할당·해제하며 사용하는 아키텍처다. 기존에는 메모리가 각 노드에 고정 탑재돼 재분배 불가능하며, 데이터센터 평균 메모리 활용률은 40~60% 수준에 그쳐 피크 대응을 위한 메모리 오버프로비저닝이 만연했다. CXL 메모리 풀링은 컴퓨팅 노드로부터 메모리를 물리적으로 분리해 공유 풀로 구성함으로써 동적 재할당을 가능케 하고, TCO(Total Cost of Ownership) 절감과 대용량 메모리 접근을 동시에 실현한다.

CXL 메모리 풀은 호스트, CXL 스위치, 메모리 확장장치 세 계층으로 구성되며, 패브릭 매니저가 동적 할당을 관리한다. AI 팩토리 환경에선 LLM 추론시 KV 캐시가 장치의 온보드 메모리 용량을 초과하는 경우가 빈번한데, CXL 메모리 풀 활용 시 GPU VRAM 대비 약 4~5배 비용 절감과 저지연 접근이 가능하다. CXL 2.0에서 메모리 풀링이 표준에 포함됐고, CXL 3.0 이후엔 BI(Back-Invalidation) 메커니즘을 통해 복수의 호스트가 동일 메모리 영역을 캐시 일관성 있게 동시 접근하는 멀티 호스트 공유 메모리가 지원됐다. 덕분에 분산 컴퓨팅 워크로드의 노드 간 데이터 공유 오버헤드를 크게 줄일 수 있다.

3. CXL 표준 발전 동향

CXL 표준은 2019년 첫 규격 공개 이후 약 6년간 메모리 확장, 풀링, 공유, 다중 랙 규모 확장이라는 방향을 따라 빠르게 진화해 왔다. <표 1>은 버전별 주요 특성을 요약한 것이다.

<표 1> CXL 버전별 주요 특성 비교

버전	출시	PHY	전송속도	주요 특징
CXL 1.0/1.1	2019	PCIe 5.0	32 GT/s	단일 디바이스 연결, Type 1/2/3 정의
CXL 2.0	2020	PCIe 5.0	32 GT/s	스위칭, 메모리 풀링, MLD
CXL 3.0	2022	PCIe 6.0	64 GT/s	멀티레벨 스위칭, 공유 메모리, BI
CXL 3.1	2023	PCIe 6.0	64 GT/s	패브릭 관리 개선, 멀티 랙 확장
CXL 3.2	2024	PCIe 6.0	64 GT/s	RAS 고도화, TSP 보안 확장
CXL 4.0	2025	PCIe 7.0	128 GT/s	번들 포트, 4-Retimer, 1.5 TB/s

3.1 CXL 1.0 / 1.1

CXL 1.0은 2019년 인텔 주도로 공개된 최초의 CXL 규격으로, PCIe 5.0(32 GT/s) 기반의 호스트-단일 디바이스 점대점 연결을 정의했다. 디바이스 유형을 Type 1(CXL.io+cache, 스마트 NIC), Type 2(3개 프로토콜 모두, GPU·FPGA), Type 3(CXL.io+mem, 메모리 확장 전용)으로 분류해 이후 메모리 풀링의 기반을 확립했다.

3.2 CXL 2.0

CXL 2.0은 2020년 스위칭 기능을 도입해 복수 호스트가 복수 메모리 디바이스에 접근하는 메모리 풀링 구조를 표준화했다. 이는 하나의 물리 장치를 최대 16개 논리 디바이스(MLD)로 분할함으로써 다수 호스트가 독립적으로 메모리 영역을 공유할 수 있게 했으며, CXL이 데이터센터의 메모리 자원 관리 기술로 발전하는 전환점이 됐다.

3.3 CXL 3.0 / 3.1

CXL 3.0은 PCIe 6.0(64 GT/s)으로 대역폭을 2배 향상시키고, BI 메커니즘 기반 공유 메모리를 도입해 복수 호스트가 동일 메모리 영역의 캐시 일관성을 유지하며 동시 접근하는 방식을 실현했다. 이는 분산 컴퓨팅 환경의 노드 간 데이터 동기화 오버헤드를 획기적으로 절감하는 핵심 기능이다. 또한 포트 기반 라우팅(Port Based Routing)과 최대 4,096 노드 확장을 지원하는 GFAM(Global Fabric Attached Memory) 모드를 도입했으며, CXL 3.1에서 패브릭 관리 기능 개선 및 멀티 랙 확장 시나리오가 보강됐다.

3.4 CXL 3.2

CXL 3.2는 물리 계층 속도는 유지하면서 운영·보안 성숙도를 높이는 데 초점을 맞췄다. OS와 애플리케이션이 CXL 메모리 디바이스 상태 정보에 정밀하게 접근할 수 있도록 모니터링 인터페이스가 확장됐으며, TSP(Trusted Security Protocol) 기능 강화를 통해 기밀 컴퓨팅 환경에서의

TEE(Trusted Execution Environment)를 CXL 메모리 디바이스 및 가속기로 확장 지원해 멀티 테넌트 환경의 메모리 격리 및 암호화를 강화했다.

3.5 CXL 4.0

CXL 4.0[2]은 2025년 11월 출시됐으며, AI 팩토리의 급증하는 메모리 대역폭 수요에 대응하기 위한 핵심 기능들을 포함한다. 물리 계층을 PCIe 7.0으로 전환해 전송 속도를 128 GT/s로 2배 향상시켰으며, 기존 256B Flit 포맷을 유지해 지연 증가 없이 대역폭 확장을 실현했다. 가장 주목할 기능은 번들 포트로, 복수의 물리적 CXL 포트를 하나의 논리적 연결로 묶어 x16 링크 기준 최대 양방향 1.5 TB/s의 집합 대역폭을 제공한다. 소프트웨어 관점에서 번들된 포트는 단일 장치로 인식돼 기존 소프트웨어 모델의 복잡도 증가 없이 고성능 연결이 가능하다.

또한 CXL 4.0에선 리타이머(Retimer) 지원 수가 기존 최대 2개에서 4개로 확대돼 다중 랙으로의 연결 확장이 표준 규격 내에서 가능해졌으며, AI 팩토리에서 요구되는 랙 간 메모리 풀링 구현 기반이 마련됐다. 이에 더해 패트롤 스크럽 오류 보고 세분화, 호스트 주도 PPR, 유연한 메모리 스페어링 등 RAS 기능을 강화해 대규모 AI 서비스의 연속성을 높였다. CXL 4.0은 CXL 1.0~3.x와의 하위 호환성을 보장하며, 상용 배포는 2026년 말~2027년으로 전망된다.

4. AI 팩토리 자원 공유 표준화 동향 및 방향

4.1 AI 팩토리 환경의 메모리 자원 공유 요구사항

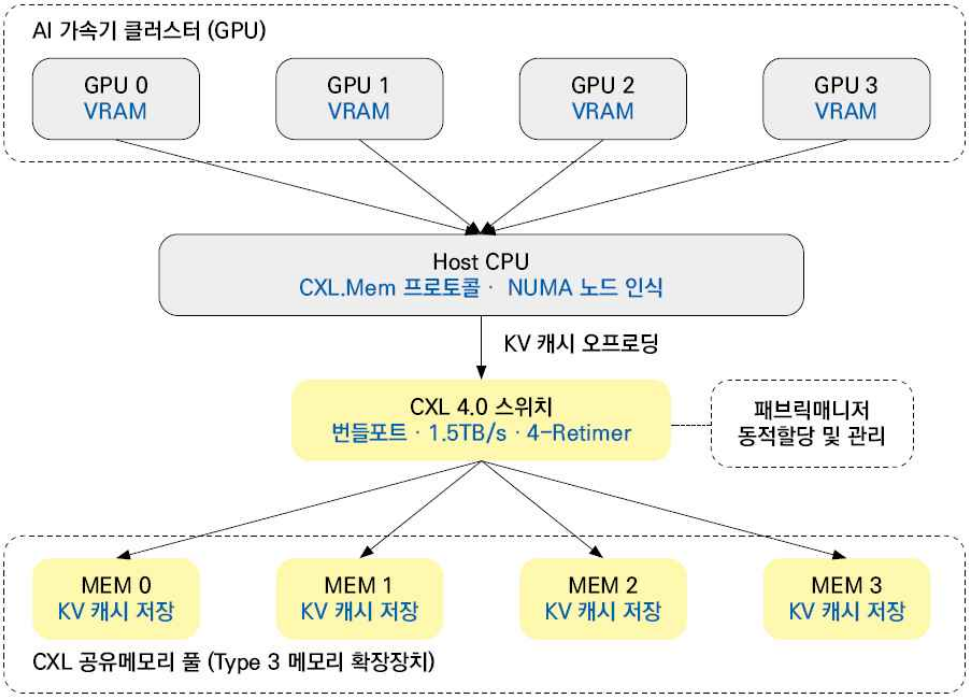
AI 팩토리 환경에서 메모리 자원 문제는 용량 병목과 메모리 단편화라는 두 가지 양상으로 나타난다. 용량 측면에선 70B 이상 LLM을 128K 토큰의 긴 컨텍스트로 처리할 경우 KV 캐시만으로 150GB 이상 메모리가 요구돼 최고급 AI 가속기의 HBM 용량(H100 80GB에서 B200·MI300X 192GB 수준)을 초과하며, SSD 오프로딩 시 수백 마이크로초~밀리초 접근 지연으로 추론 성능이 급격히 저하된다[3][4].

자원 단편화 측면에선 워크로드별 메모리 수요 변동에도 불구하고 메모리가 서버 소켓에 고정 탑재돼 동적 재배분이 불가능하며, 데이터센터 평균 메모리 활용률은 40~60% 수준에 머문다. 메모리 비용이 서버 구매 비용의 약 40~50%를 차지하는 만큼 이는 막대한 TCO 손실로 직결된다.

CXL 기반 메모리 풀링은 [그림 2]와 같이 AI 팩토리 환경의 메모리 용량 병목과 자원 단편화 문제를 동시에 해소하는 구조적 해법을 제공한다. LLM 추론시 발생하는 대용량 KV 캐시를 CXL 메모리 풀로 오프로딩하면 200~500ns 저지연을 유지하면서 GPU VRAM 대비 약 4~5배 낮은 비용으로 대용량 메모리를 확보할 수 있으며, 200G와 100G RDMA 방식 대비 3.8[1]~6.5배[2]의 처리량 향상을 달성할 수 있다. 또한 패브릭 매니저를 통해 훈련 작업 종료 후 해당 메모리 영역을 추론 클러스터로 즉시 재할당하는 동 적 운영이 가능해, 데이터센터 전체의 메모리 활용률을 크게 향상시킬 수 있다.

4.2 글로벌 표준화 동향

CXL 컨소시엄은 280개 이상 회원사를 보유하며 Gen-Z, OpenCAPI, CCIX 등 경쟁 표준을 통합해 사실상 단일 개방형 표준으로 확정됐다. CXL 2.0 기반 메모리 확장 솔루션은 2025년 양산 단계에



[그림 2] LLM 추론 성능 향상을 위한 CXL 4.0 메모리 풀 자원 공유 예

진입했으며, CXL 3.x 기반 풀링은 2026~2027년, CXL 4.0 기반 다중 랙 풀링은 2027년 이후 프로덕션 적용이 전망된다. 연관 표준화 측면에서 JEDEC는 CXL 컨소시엄과 MOU를 체결하고 CXL 메모리 모듈 인터페이스 표준(JESD317)을 제정했으며, OCP는 CXL 메모리 확장 모듈의 폼팩터·전원·냉각 표준화를 추진 중이고, PCIe 7.0은 CXL 4.0의 물리 계층 기반으로 직접 활용되고 있다.

산업 생태계 측면에서 SK 하이닉스는 CXL 2.0 기반 CMM-DDR5로 AI 워크로드 토큰 처리 성능 31% 향상을 제시했으며, 삼성전자와 마이크론(Micron)도 CXL 2.0/3.x 호환 메모리 확장장치를 양산 중이다. 아스테라 랩스(Astera Labs)의 Leo CXL 컨트롤러는 마이크로소프트(Microsoft) 애저(Azure)의 CXL 클라우드 인스턴스에 채택됐고, 국내 스타트업 파네시아는 2025년 11월 업계 최초로 CXL 3.2 패브릭 스위치 샘플 제공을 시작했다. 엔비디아(NVIDIA) 블랙웰(Blackwell)과 AMD MI300X도 CXL을 통한 메모리 확장을 공식 지원하기 시작했으며, 2025년 11월 마이크로소프트 애저의 CXL 탑재 클라우드 인스턴스 공개 프리뷰 출시는 상용화 진입의 중요한 이정표로 평가된다.

4.3 국내 표준화 방향 및 TTA 역할

국내 AIDC 표준화는 몇 가지 구조적 과제를 안고 있다. CXL 핵심 특허와 표준 주도권이 미국 빅테크중심으로 형성돼 있어 국내 기업이 표준 추종자(Fast Follower) 위치에 머무를 위험이 있으나, SK 하이닉스와 삼성전자가 JEDEC 및 CXL 컨소시엄 내 DRAM 관련 규격에 적극 기여하고 있는 점은 긍정적이다. 또한 상용 CXL 하드웨어 연구 환경이 아직 성숙하지 않아 에뮬레이션 의존도가 높으며, OS 수준의 NUMA 인식, 패브릭 매니저 인터페이스 등 소프트웨어 스택의 국산화 기술 개발이 요구된다.

TTA는 국내 기업의 해외 시험인증 기관 의존 구조를 탈피하기 위해 CXL 전기·프로토콜 적합성 및 다양한 벤더 간 상호운용성 시험 역량을 조기에 확보해야 하며, CXL 기반 메모리 풀링 도입을

위한 배포 아키텍처 가이드라인·성능 평가 방법론·보안 요구사항을 표준으로 제정해 국내 산업 생태계의 도입 장벽을 낮출 필요가 있다.

산·학·연 역할 분담 측면에서, 산업계는 CXL 4.0 호환 디바이스 양산 및 컨트롤러·패브릭 매니저 국산화에, 학계는 CXL 활용 소프트웨어 연구와 에뮬레이션 기반 선행 연구에, 출연연은 AIDC 참조 아키텍처 구축 및 ITU-T·IEEE 등 국제표준화 기구에서의 기고 활동 강화에 집중해 글로벌 표준화 논의에서 국내 입지를 높여야 한다.

5. 맺음말

이번 원고는 2025년 11월 출시된 CXL 4.0 표준의 핵심 기술 내용과 CXL 기반 메모리 풀링이 AI 팩토리 환경에서 갖는 의미를 분석하고, 관련 국내외 표준화 동향 및 국내 대응 방향을 정리했다. 기술적 관점에서 CXL 4.0은 단순한 속도 향상을 넘어 AI 팩토리 인프라의 구조적 전환을 가능케 하는 이정표적 표준이다. 이는 PCIe 7.0 기반 128 GT/s 전송 속도, 번들 포트를 통한 1.5 TB/s 대역폭, 4-리타이머 지원을 통한 다중 랙 확장 능력을 바탕으로 기존 단일 서버 종속 메모리 구조를 랙 간 규모의 공유 메모리 풀 아키텍처로 전환할 수 있는 표준적 기반을 처음으로 완성한 것이다.

다만 CXL 4.0 기반 다중 랙 시스템의 본격 상용 배포는 2027년 이후로 전망되며, 그 이전까지 OS 수준의 CXL-aware 메모리 관리, 패브릭 매니저 표준 인터페이스 정착, AI 프레임워크의 CXL 메모리 최적화 지원 등을 통해 소프트웨어 생태계 성숙이 선행돼야 CXL 4.0의 하드웨어 성능이 실질적인 서비스 가치로 전환될 수 있다.

국내 대응 관점에선 메모리 반도체 강국이라는 강점을 표준화 주도권으로 연결하는 전략적 전환이 요구되며, SK 하이닉스·삼성전자의 표준화 기여 확대, TTA 중심의 CXL 시험인증체계 수립, 산·학·연 협력기반 응용 소프트웨어 경쟁력 확보가 병행돼야 한다. AI 팩토리 시대의 메모리 인프라 표준화는 데이터센터 경제성과 국가 AI 경쟁력을 좌우하는 핵심 전략 과제임을 인식하고, 글로벌 표준화 논의에 적극 참여해 나가야 할 것이다.

[참고문헌]

[1] Compute Express Link, "https://computeexpresslink.org/"

[2] CXL Specification, "https://computeexpresslink.org/cxl-specification/"

[3] Y. et al., "ARKV: Adaptive and Resource-Efficient KV Cache Management under Limited Memory Budget for Long-Context Inference in LLMs," arXiv, 2025. https://arxiv.org/abs/2603.08727

[4] D. Liu and Y. Yu, "CXL-SpecKV: A Disaggregated FPGA Speculative KV-Cache for Datacenter LLM Serving," ACM/SIGDA FPGA '26, Feb. 2026. https://arxiv.org/pdf/2512.11920

[5] S. Li et al., "Rcmp: Reconstructing RDMA-Based Memory Disaggregation via CXL," ACM Transactions on Architecture and Code Optimization, 2024. https://dl.acm.org/doi/10.1145/3634916

[6] XConn Technologies & MemVerge, "Breakthrough Scalable CXL Memory Solution to Offload KV Cache in AI Inference Workloads," Supercomputing 2025 (SC25) Demo, Nov. 2025.

※ 출처: TTA 저널 제224호