이기종 AI 가속기 연동 및 통합 관리 기술

이강찬 TC10 의장, 한국전자통신연구원 전략표준연구실장
인민교 PG1003 부의장, 한국전자통신연구원 책임연구원

1. 머리말

생성형 AI와 LLM(거대언어모델, Large Language Model)의 급속한 확산은 AIDC(AI Data Center) 인프라 전반에 전례 없는 전환을 촉발하고 있다. 시장조사기관 IDC(International Data Corporation)에 따르면 2024년 전 세계 AI 인프라 투자 규모는 약 1,100억 달러에 달하며, 2028년까지 연평균 26.5% 성장해 3,500억 달러를 상회할 것으로 전망된다[1]. 과거 CPU 중심으로 운영되던 데이터센터는 이미 GPU 및 AI 가속기 중심 구조로 빠르게 재편되었으며, 이 흐름은 앞으로도 가속될 것이다.

변화는 단순히 연산 장치 교체에 그치지 않는다. GPU, NPU, FPGA, ASIC 등 아키텍처가 상이한 다양한 가속기가 동일한 AIDC 내에 함께 배치되는 이기종(Heterogeneous) 환경이 표준이 되고 있다. 여기에 온프레미스 데이터센터와 클라우드 인프라가 결합된 하이브리드 구조, 그리고 지리적으로 분산된 복수 AIDC를 연결하는 분산형(Distributed) 인프라까지 더해지면서, AI 인프라의 운영 복잡도는 이전과 비교할 수 없는 수준으로 높아지고 있다.

이 변화의 이면에는 구조적인 운영 문제가 자리하고 있다. 각 제조사는 자사 가속기에 최적화된 독자적 관리 인터페이스, 드라이버 스택, 모니터링 도구를 제공하고 있어, 운영자는 이기종 장비별로 분절된 관리 환경에서 인프라를 파편적으로 운용할 수밖에 없다. 더욱이 여러 데이터센터에 분산된 AI 가속기를 하나의 논리적 자원 풀로 묶어 효율적으로 활용하는 것은 통합된 관리 표준 없이는 실질적으로 제한적이다. 이는 고가 AI 인프라의 활용률 저하, 운영 비용 증가, 서비스 가용성 저하로 직결되는 실질적인 산업 문제다.

이번 원고에서는 이러한 문제의식을 바탕으로, AIDC 환경에서 이기종 AI 가속기를 효과적으로 연동·통합 관리하기 위한 기술 구조와 표준화 방향을 분석한다. AI 가속기 시장의 다양화와 이기종 혼재 환경의 현실을 살펴보고, 통합 관리에 요구되는 핵심 기술 개념과 참조 아키텍처를 제시한다. 이어서 레드피쉬(Redfish) API 기반 접근, 오케스트레이션, 네트워크 경로 최적화 등 주요 기술 요소를 설명한다.

2. AIDC 환경의 이기종 AI 가속기 현황

2.1 AI 가속기 시장의 다양화

AI 연산 수요의 폭발적 증가와 함께, 단일 GPU 중심이던 가속기 시장이 다양한 아키텍처로 분화되고 있다. 대규모 병렬 연산에 최적화된 GPU는 여전히 LLM 학습의 핵심 장치로 자리하고 있으

나, 특정 AI워크로드에 특화된 대안 아키텍처들이 실제 운영 환경에 빠르게 도입되고 있다. 그랜드 뷰 리서치(Grand View Research)에 따르면 글로벌 데이터센터 가속기 시장은 2023년 약 490억 달러에서 2030년까지 연평균 29.2% 성장할 것으로 전망되며[2], 이 성장의 상당 부분을 이기종 가속기 확산이 견인할 것으로 분석된다. 아키텍처별로 살펴보면 다음과 같다. GPU는 CUDA/ROCm 기반 범용 병렬 연산 생태계를 갖추고 있어 AI 학습에 가장 널리 활용되고 있는데, 엔비디아(NVIDIA) H100, H200, B200과 AMD MI300X 등이 대표적이다. NPU는 행렬 곱셈과 같이 AI 연산에 특화된 연산 유닛을 집중 배치해 GPU 대비 에너지 효율을 높인 구조로, 구글(Google) TPU, 아마존(Amazon) Trainium·인텔(Intel) Gaudi, 국산 리벨리온 ATOM, 사피온 X330 등이 이에 해당한다. FPGA는 하드웨어 수준의 프로그래머블 구조로 지연에 민감한 추론 서비스나 전처리 가속에 활용되며, 인텔 Stratix, Xilinx Alveo 등이 대표적이다. ASIC은 특정 모델·워크로드에 완전히 특화된 고정 회로로 최고 수준의 에너지 효율을 달성하지만 유연성이 낮다는 특성이 있다.

이러한 다양화는 기술적 요인과 지정학적 요인이 복합적으로 작용한 결과다. 기술적으로는 트랜스포머 아키텍처 기반 AI 워크로드가 범용 GPU와 다른 연산 특성을 요구하면서 DSA(도메인 특화 아키텍처, Domain-Specific Architecture)의 경쟁력이 높아졌다. 지정학적으로는 미국 AI 반도체 수출 규제 강화가 자국 AI 반도체 역량 강화를 위한 각국의 정책 지원을 촉진했다. <표 1>은 AIDC에 실제 배치되거나 도입이 임박한 주요 AI 가속기의 특성을 비교한 것이다.

<표 1> 주요 AI 가속기 유형 및 관리 인터페이스 현황

유형	대표 제품	주요 용도	관리 인터페이스
GPU	NVIDIA H100/B200, AMD MI300X	LLM 학습 대규모 병렬 연산	NVML, DCGM, ROCm SMI
NPU	Google TPU v5p, Amazon Trainium2, Intel Gaudi 3, 리벨리온 ATOM, 사피온X330	추론·학습 에너지 효율 최적화	플랫폼 전용 SDK, HL-SMI, 독자 CLI
FPGA	Intel Stratix, Xilinx Alveo	저지연 추론 전처리 가속	OpenCL, Vitis 독자 관리 도구
ASIC	특정 모델 특화	고정 워크로드 최고 에너지 효율	제조사별 상이

2.2 이기종 혼재 환경의 운영

단일 AIDC에 이기종 AI 가속기가 혼재하는 상황은 이미 현실이 되고 있다. 실제 AI 서비스 사업자들은 LLM 학습에 엔비디아 GPU 클러스터, 추론 서비스에 비용 효율이 높은 NPU, 전처리·후처리 파이프라인에 FPGA를 함께 운용하는 복합 구조를 채택하고 있다. 여기에 온프레미스 데이터센터와 클라우드 인프라가 결합된 하이브리드 환경, 그리고 지리적으로 분리된 복수 AIDC를 연결하는 분산형 구조가 더해지면서 운영 복잡도가 크게 증가한다. 이처럼 서로 다른 세대의 GPU, 다양한 NPU, FPGA가 온프레미스와 클라우드에 걸쳐 분산 배치된 환경에선 자원 상태관리, 작업 배치, 성능 모니터링이 근본적으로 복잡해진다.

운영 현장에서 가장 먼저 드러나는 문제는 통합 가시성(Unified Visibility) 부재다. 운영자는 엔비디아 가속기를 보기 위해 N VML 기반 대시보드를, AMD 가속기를 보기 위해 ROCm 모니터링 도구를, 국산 NPU를 보기 위해 또 다른 독자 CLI를 각각 실행해야 한다. 장비별 관리 인터페이스

차이로 인해 통합 운영 자동화가 어려워지고, 전력 소비, 온도, 메모리 사용량, 연산 활용률 등 핵심 운영 지표를 클러스터 단위로 통합해 파악하는 것이 사실상 불가능하다. 업계 조사에 따르면 이기종 AIDC 환경에서 가속기 평균 활용률은 단일 제조사 환경 대비 15~25%가량 낮은 것으로 보고된다[2].

두 번째 문제는 워크로드 오케스트레이션의 한계다. 쿠버네티스(Kubernetes), SLURM 등 클러스터 오케스트레이터는 이기종 가속기의 세부 특성을 표준화된 방식으로 인식하지 못하기 때문에, 각 가속기의 유휴 용량과 성능 특성을 고려한 동적 작업 스케줄링이 어렵다. 특히 여러 데이터센터에 분산된 가속기 자원을 하나의 논리적 풀로 묶어 AI 작업을 최적 배치하기 위해선, 가속기 자원의 위치·성능·상태 정보를 공통 모델로 표현하고 교환할 수 있는 표준화된 인터페이스가 반드시 필요하다.

세 번째 문제는 장애 대응과 수명주기 관리의 파편화다. 이기종 가속기에서 발생하는 하드웨어 오류, 온도 초과, 메모리 오염 등의 이벤트는 제조사별로 상이한 형식과 채널로 보고된다. 통합된 이벤트 수집체계 없이는 심각한 장애도 신속한 대응이 어렵다. 펌웨어 업데이트, 드라이버 패치 등 수명주기 관련 작업도 제조사별로 개별 수행해야 하므로, 관리 인력 부담이 가속기 종류 수에 비례해 선형 증가한다. 결론적으로, 이기종 AI 가속기 환경에선 표준 기반의 통합 관리 구조가 선택이 아닌 필수 조건이다.

3. 이기종 AI 가속기 연동·통합 관리 핵심 개념과 구조

3.1 통합 관리의 핵심 요구사항

이기종 AI 가속기 통합 관리체계를 설계하기 위해선 운영 현장의 요구사항을 명확히 정의해야 한다. 이 요구사항들은 단순한 기능 목록이 아니라, 통합 관리시스템의 아키텍처 결정에 직접 영향을 미치는 핵심 제약 조건이 된다.

첫 번째는 가속기 자원 인벤토리 관리다. 이기종 환경에서는 GPU, NPU, FPGA, ASIC 등 각 가속기의 종류, 위치, 수량, 성능 사양, 현재 상태를 정확하게 파악하고 관리하는 것이 모든 자동화의 출발점이다. 새로운 가속기가 네트워크에 연결될 때 자동으로 식별·등록되는 자동 디스커버리(Auto-Discovery) 기능이 필수적이며, 가속기 자원의 위치, 성능, 상태 정보를 공통 데이터 모델로 표현할 수 있어야 한다.

두 번째는 통합 자원 상태 모니터링이다. 수천 개 이기종 가속기 노드에 대한 전력, 온도, 메모리 사용률, 연산 활용률 등의 지표를 표준화된 포맷으로 실시간 수집하고, 이를 단일 관리 뷰에서 제공해야 한다.

세 번째는 가속기 자원 할당 및 스케줄링이다. AI 오케스트레이션 시스템이 이기종 가속기의 특성을 인식하고 워크로드를 최적으로 배치할 수 있도록 장비관리 시스템과 오케스트레이션 시스템 간의 연동 인터페이스가 표준화돼야 한다.

네 번째는 제조사 독립성(Vendor-Agnosticism)이다. 새로운 가속기가 도입될 때 기존 관리 소프트웨어의 대규모 수정 없이 수용할 수 있는 확장성은 표준 기반인터페이스를 통해서만 달성 가능하다.

3.2 상호운용성 확보를 위한 기술 요구사항

통합 관리체계가 이기종 환경에서 실제 작동하려면, 각 가속기와 상위 관리 소프트웨어 사이에 명확히 정의된 상호운용성 기술 요구사항이 충족돼야 한다. <표 2>는 이를 영역별로 정리한 것이다. 이기종 환경에서 상호운용성 확보가 중요한 이유는 하나의 관리도구가 제조사와 무관하게 동일한 방식으로 모든 가속기를 제어할 수 있어야 전체 인프라 운영 자동화가 가능해지기 때문이다.

<표 2> 이기종 AI 가속기 상호운용성 기술 요구사항

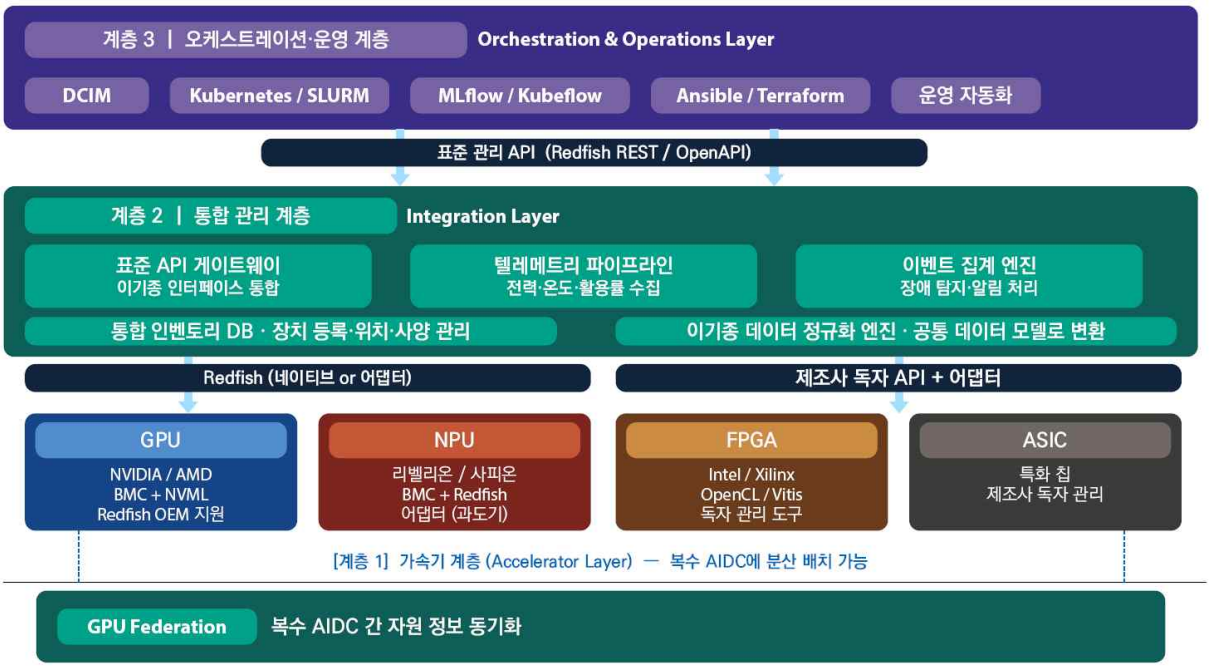
영역	요구사항 내용	구현 기술/표준 후보
디스커버리	장치 자동 감지 및 인벤토리 자동 등록	SMBIOS, Redfish Discovery
모니터링	전력·온도·활용률 표준 메트릭 수집·제공	Redfish Telemetry, Prometheus exporter
수명주기 관리	펌웨어 업데이트·롤백 표준 절차 지원	Redfish Update Service
이벤트·알림	이상 이벤트 표준 포맷 탐지·전달	Redfish Event Service, SNMP Trap
보안	TLS 암호화 통신, 역할 기반 접근 제어(RBAC)	OAuth2, OpenID Connect, X.509
오케스트레이션 연동	가속기 특성 메타데이터 제공, 작업 배치 인터페이스	Kubernetes Device Plugin, Redfish + 확장
AI 특화 지표	TOPS, MFU, 메모리 대역폭 활용률 등	OEM Extension
분산 자원 연합	복수 AIDC 간 자원 정보 모델 공유·동기화	신규 표준화 영역

3.3 분산 AI 가속기 통합 관리 참조 아키텍처

분산 AI 환경에서는 단순히 단일 데이터센터 내 이기종 가속기를 관리하는 것을 넘어, 여러 데이터센터에 걸쳐 분산된 GPU, NPU 자원을 하나의 논리적 자원 풀로 구성해 활용할 수 있어야 한다. 이러한 구조는 GPU 연합(GPU Federation) 개념으로 설명될 수 있다. GPU 연합은 지리적으로 분산된 다수 AIDC 내 가속기 자원을 단일 오케스트레이션 계층에서 통합 관리하며, AI 워크로드의 특성에 따라 최적의 자원 위치에 작업을 배치하는 분산 컴퓨팅 패러다임이다.

이 참조 아키텍처는 [그림 1]과 같이 세 개의 수직적계층으로 구성된다. 최하단 가속기 계층(Accelerator Layer)에는 다양한 제조사의 GPU, NPU, FPGA,ASIC이 위치하며, 각 장비는 BMC(자체 관리 컨트롤러, Baseboard Management Controller) 또는 이에 상응하는 온보드 관리 칩을 통해 관리 신호와 텔레메트리를 노출한다. 이상적으론 각 가속기가 레드피쉬 등 표준 관리 인터페이스를 네이티브로 지원하는 것이 최선이지만, 현실적으론 어댑터 또는 사이드카 방식으로 표준 인터페이스를 추가하는 과도기적 구현이 병행된다. 중간 통합 관리 계층(Integration Layer)은 이 아키텍처의 핵심이다. 이기종 가속기로부터 수집한 관리 정보를 표준화된 데이터 모델로 정규화(Normalize)하고, 상위 오케스트레이션 및 DCIM(데이터센터 인프라 관리, Data Center Infrastructure Management) 소프트웨어에 일관된 API를 제공한다. 이 계층은 표준 관리 API 게이트웨이, 이기종 텔레메트리 수집 파이프라인, 이벤트 집계 및 상관 분석 엔진, 통합 인벤토리 데이터베이스로 구성된다. GPU 연합 구조에선 이 계층이 복수의 AIDC에 걸쳐 분산 배치되며, 글로벌 자원 뷰를 제공하기 위한 계층 간 정보 동기화 메커니즘이 추가된다.

최상단 오케스트레이션·운영 계층(Orchestration & Operations Layer)에는 DCIM 소프트웨어, 쿠버



[그림 1] 이기종 AI 가속기 연동·통합 관리 참조 아키텍처

네티스, SLURM 등 워크로드 오케스트레이터, AI/ML 플랫폼(MLf low, Kubef low 등), 운영 자동화 도구(Ansible, Terraform 등)가 위치한다. 이 계층은 통합 관리 계층이 제공하는 표준 API를 통해 이기종 AI 가속기를 단일한 추상화된 자원 풀로 인식하고, 자원 탐색(Discovery), 자원 매칭(Matching), 실행 관리(Execution Management)를 자동화할 수 있다. 표준화에서 가장 중요한 요소는 가속기 계층과 통합 관리 계층 사이의 인터페이스로, 이 인터페이스가 표준화되지 않으면 통합 관리 계층의 복잡도가 가속기 종류 수에 비례해 증가하는 악순환이 반복된다.

4. 이기종 AI 가속기 연동·통합 관리 핵심 기술

4.1 레드피쉬 API 기반 통합 관리 체계

레드피쉬는 DMTF(Distributed Management Task Force)에서 정의한 데이터센터 장비 관리 표준으로, RESTful API 기반 관리 인터페이스를 제공한다. 기존 IPMI(Intelligent Platform Management Interface)의 기능적 한계와 보안 취약성을 극복하기 위해 2015년 첫 발표 이후, 델(Dell), 휴렛팩커드 엔터프라이즈(Hewlett Packard Enterprise), 레노보(Lenovo), 슈퍼 마이크로 컴퓨터(Super Micro Computer) 등 주요 서버 제조사 전반이 채택하면서 서버 관리의 사실상 표준(de facto standard)으로 자리 잡았다. 레드피쉬는 HTTP/HTTPS 기반 JSON 포맷, OpenAPI 스펙 기반 자동화 도구 생성, SSE(Server-Sent Events) 기반 실시간 이벤트 스트리밍, 세분화된 RBAC(역할 기반 접근 제어, Role-Based Access Control) 등 현대적인 API 설계 원칙을 준수한다.

AIDC에서 레드피쉬를 활용하면 GPU 서버 및 AI 가속기 장비의 전력 상태, 온도, 펌웨어 버전, 메모리 용량 등 장비 상태 정보를 표준화된 방식으로 수집할 수 있다. 2023년부터 DMTF는 레드피쉬 스펙의 Processor 리소스 스키마를 GPU·NPU 등 이기종 가속기에 적용 가능하도록 확장하는 작업을 본격화했으며, 레드피쉬 1.19(2024) 기준으로 ProcessorCollection, ProcessorMetrics 리소스

를 통해 가속기 기본 속성과 실시간 성능 지표를 표준화된 JSON 스키마로 표현할 수 있게 됐다. 그러나 레드피쉬는 장비 관리 계층(Device Management Layer) 기술임을 명확히 이해해야 한다. 레드피쉬는 장비 상태 모니터링, 펌웨어 관리, 전원 제어 등 하드웨어 수명주기 관리에 특화돼 있으며, AI 워크로드 자원 스케줄링이나 오케스트레이션을 직접 수행하는 기술이 아니다. 따라서 레드피쉬만으로 이기종 AI 가속기 통합 관리의 모든 요구사항을 충족할 수는 없다. TOPS, MFU, 텐서 코어 활용률 등 AI 워크로드 특화성능 지표는 현 표준 스키마에 포함돼 있지 않으며, 멀티-GPU 토폴로지와 이기종 가속기 간 상호 연결정보의 표현도 미흡하다.

국산 NPU 제조사 관점에서 레드피쉬 지원은 선택이 아닌 전략적 필수 요소다. 국내 AIDC 운영자가 기존 DCIM 인프라와 연동되지 않는 가속기 도입을 꺼리는 현실에서, 레드피쉬 네이티브 지원 또는 검증된 레드피쉬 어댑터 제공은 시장 진입 장벽을 낮추는 핵심 요소가 된다. 단기적으로는 기존 독자 API에 대한 레드피쉬 어댑터 레이어를 개발하고, 중기적으로는 BMC 수준에서 레드피쉬를 내재화하며, 장기적으로는 AI 특화 메트릭을 국제 표준 스키마에 편입시키는 단계적 접근이 현실적이다.

4.2 연동 최적화와 오케스트레이션

통합 관리 API만으로는 이기종 AI 가속기 환경의 운영 효율을 충분히 높이기 어렵다. 분산 GPU 환경에서 AI 작업을 효율적으로 실행하기 위해선 자원 탐색(Resource Discovery), 자원 매칭(Resource Matching), 실행 관리(Execution Management)의 세 단계가 통합된 오케스트레이션 체계가 필요하다. 자원 탐색 단계에선 분산된 이기종 가속기의 가용 용량과 현재 상태를 실시간으로 파악하고, 자원 매칭 단계에선 AI 작업의 요구사항(필요 메모리, 연산 유형, 지연 요건 등)과 가속기 특성을 매칭해 최적의 배치 결정을 내린다. 실행 관리 단계에선 선택된 가속기에 작업을 배정하고 실행 상태를 모니터링하며 장애 발생 시 자동 재배치(Failover)를 수행한다.

가속기 인식 스케줄링(Accelerator-Aware Scheduling)은 이 오케스트레이션 체계의 핵심 기술이다. 쿠버네티스의 Device Plugin 프레임워크와 같이, 오케스트레이터가 이기종 가속기의 타입·용량·현재 부하·네트워크 토폴로지를 표준화된 속성으로 인식해 AI작업을 최적의 가속기에 배정하는 기술이다. 예를 들어, LLM 학습처럼 GPU 간 고대역폭 통신이 필수적인 작업은 NVLink로 연결된 엔비디아 GPU 클러스터에, 배치 추론처럼 전력 효율이 중요한 작업은 저전력 NPU에, 실시간 추론처럼 지연이 민감한 작업은 FPGA에 각각 배정하는 방식이다. 이 스케줄링의 품질은 결국 통합 관리 계층의 표준 API가 얼마나 풍부한 가속기 메타데이터를 제공하는가에 달려 있다.

텔레메트리 기반 예측적 관리(Predictive Management)도 중요한 기술 요소다. 이기종 가속기에서 수집되는 대량의 운영 메트릭을 AI/ML 모델로 분석해 장애를 사전에 예측하고 선제적 조치를 취하는 접근이다. GPU 온도 상승 패턴 및 메모리 오류율 증가 추세를 분석해 하드웨어 교체 시점을 예측하거나, 특정 가속기 성능 저하를 조기에 탐지해 해당 가속기에 배정된 작업을 자동으로 다른 가속기로 이전하는 자동화가 가능해진다. AI 인프라 운영 자동화를 완성하려면 레드피쉬 기반의 장비 관리와 이러한 오케스트레이션·예측 관리 기술이 표준 인터페이스를 매개로 긴밀하게 연동돼야 한다.

4.3 네트워크·데이터 경로 최적화

대규모 AI 모델 학습에서는 GPU·NPU 간 그래디언트 교환과 모델 파라미터 동기화에 사용되는 네트워크 대역폭이 전체 학습 성능을 결정짓는 핵심 요소다. 동종(Homogeneous) 환경에선 엔비디아 NVLink/NVSwitch나 AMD xGMI와 같이 제조사가 최적화한 전용 인터커넥트를 사용할 수 있지만, 이기종 환경에선 이러한 전용 인터커넥트를 사용할 수 없어 표준 네트워크 패브릭에 의존해야 한다.

이기종 가속기 환경에서의 표준 인터커넥트로는 현재 InfiniBand와 고속 이더넷(100GbE/400GbE/800GbE)이 주로 사용된다. RDMA(원격 직접 메모리 접근, Remote Direct Memory Access) 기술은 CPU 개입 없이 가 속기 메 모리 간 데이터를 직접 전송해 지연을 최소화하고 CPU 부하를 줄이는 핵심 기술이다. RoCE(RDMA over Converged Ethernet)와 InfiniBand 기반 RDMA를 이기종 가속기 환경에서 일관성 있게 활용하기 위해선 네트워크 패브릭과 가속기 드라이버 간 상호운용성 검증이 필수적이다. 또한 여러 데이터센터에 걸친 GPU 연합 구조에서는 데이터센터 간 연결 네트워크의 대역폭과 지연이 전체 AI 학습 성능에 직접적인 영향을 미치므로, 데이터 경로 최적화와 트래픽 우선순위 관리가 더욱 중요해진다.

최근에는 CXL(Compute Express Link)이 이기종 가속기 연동의 새로운 기술 기반으로 주목받고 있다. CXL은 PCIe 기반의 캐시 코히어런트(Cache-Coherent) 인터커넥트 표준으로, CPU, GPU, NPU, 메모리 등 이기종 컴퓨팅 자원 간 메모리 공유와 직접 접근을 가능케 한다. CXL이 본격 상용화되면 이기종 가속기들이 메모리 자원을 효율적으로 공유하고 데이터 복사 없이 직접 협력하는 아키텍처가 가능해진다. 다만 CXL 기반 이기종 가속기 연동은 현재 표준화와 제품화가 진행 중인 단계로, 실제 AIDC 적용까지는 추가적인 시간이 필요할 것으로 전망된다.

5. 맺음말

AIDC 환경에선 GPU, NPU, FPGA, ASIC 등 다양한 AI 가속기를 효율적으로 활용하기 위한 통합 관리 기술이 필수적으로 요구된다. 이기종 AI 가속기 연동·통합 관리 기술은 AIDC 운영 효율성을 높이고 AI 인프라 활용도를 향상시키는 핵심 기술로, 향후 AIDC 확산과 함께 관련 기술 및 표준화 활동이 더욱 활발히 진행될 것이다.

이번 원고에서 살펴본 바와 같이, 이 문제는 단일 기술로 해결되지 않는다. 레드피쉬는 장비 관리 계층의 핵심 표준으로서 중요한 역할을 담당하지만, 그것만으론 충분하지 않다. 가속기 인식 오케스트레이션, 고성능 네트워크 인터커넥트, 텔레메트리 기반 예측적 관리, GPU 연합 구조에서의 분산 자원 연합 관리 등이 참조 아키텍처의 각 계층에서 유기적으로 결합될 때, 실질적인 운영 효율화가 가능해진다. 각 기술의 역할과 한계를 함께 이해하고, 전체 아키텍처 관점에서 통합하는 접근이 중요하다.

국내 산업계·연구기관·정부에 대한 제언을 세 가지로 정리한다.

첫째, 국산 NPU 개발 로드맵에 레드피쉬 등 표준 관리 인터페이스 지원 일정을 명시적으로 포함해야 한다. 하드웨어 성능이 우수해도 기존 DCIM 생태계와 연동되지 않는 가속기는 시장에서 선택받기 어렵다.

둘째, 이기종 AI 가속기 통합 관리 분야는 국제표준화 경쟁이 본격화되기 전 초입 단계에 있어,

지금이 국내 주도 표준을 선점할 수 있는 결정적 시점이다. ITU-T SG13에 분산 AI 자원 연합 관리 관련 신규 표준화 과제를 선제적으로 제안하고, DMTF, OCP 등 주요 국제표준화 기구에서도 국내 기술 요구사항과 국산 NPU의 특성이 표준 스키마에 반영될 수 있도록 적극 대응해야 한다. 표준화 기여 없이는 국산 가속기가 글로벌 DCIM 생태계에서 '지원 대상 외 장치'로 남을 수밖에 없으며, 이는 아무리 우수한 하드웨어 성능도 시장 경쟁력으로 이어지지 못하는 구조적 장벽이 된다.

셋째, 정부는 국산 AI 가속기의 공공 시장 진입을 위한 제도적 기반을 적극적으로 조성해야 한다. 공공AI 인프라 조달 기준에 표준 관리 인터페이스 준수 여부를 단계적으로 포함하고, 국산 가속기가 표준화된 인터페이스를 갖출 수 있도록 R&D 지원 과제 기획단계부터 관리 소프트웨어 생태계 요건을 명시하는 방향으로 정책 설계가 이뤄져야 한다. 미국과 유럽이 자국 AI 반도체 생태계를 정부, 군 수요를 통해 육성한 사례를 벤치마크해, 국내에서도 공공 조달이 국산 AI 가속기 표준화 생태계 형성의 마중물 역할을 담당할 수 있어야 한다.

AI 인프라의 경쟁력은 가속기 칩 하나의 성능만으로 결정되지 않는다. 수천·수만 개의 이기종 가속기를 통합된 시스템으로 운용할 수 있는 소프트웨어, 표준, 생태계 역량이 함께 갖춰질 때 비로소 진정한 AI 인프라 경쟁력이 완성된다. 국산 AI 반도체 개발에 쏟아지는 투자가 실질적인 산업 성과로 이어지기 위해서는 하드웨어 개발과 표준화, 생태계 구축이 처음부터 함께 설계돼야 한다. 산업계, 연구기관, 표준화 기구, 정부가 한 방향으로 협력하는 것이 어느 때보다 중요한 시점이다.

[이 연구는 과학기술정보통신부와 정보통신기획평가원이 지원하는 AI 학습·추론을 위한 분산 GPU 자원 연동 및 관리 표준 개발과제를 통해 발생한 결과임]

[참고문헌]

[1] IDC, "Worldwide Artificial Intelligence Infrastructure Forecast, 2024-2028", IDC #US52177524, June 2024.

[2] Grand View Research, "Data Center Accelerator Market Size, Share & Trends Analysis Report", GVR-4-68039-768-7, 2024.

[3] DMTF, "Redfish Scalable Platforms Management API Specification", DSP0266 Version 1.19.0, DMTF Standard, 2024.

[4] DMTF, "Redfish Data Model Specification", DSP0268 Version 2024.1, DMTF Standard, 2024.

[5] ITU-T, "Recommendation ITU-T Y.3507: Requirements and framework for management of heterogeneous computing resources in cloud computing", ITU-T Study Group 13, 2022.

[6] TTA, "인공지능 반도체 표준화 현황 및 추진 방향", TTA 표준화 이슈 리포트 제68호, 2024.

[7] OCP, "Open Accelerator Infrastructure (OAI) Architecture Specification v2.0", Open Compute Project, 2023.

[8] IEEE, "P2941: Standard for AI/ML Model Representation, Compression, Distribution and Management", IEEE Working Group Draft, 2024.

※ 출처: TTA 저널 제224호