AI 워크로드를 위한 고성능 컴퓨팅(HPC) 프로세서 기술 발전 동향

Technology Trends in HPC Processors for Supporting AI Workloads

박유미 (Y.M. Park, parkym@etri.re.kr) 기업성장지원전략실 책임연구원
기문철 (M.C. Kee, mooncheol.kee@supergate.cc) 수퍼게이트(주)/연구소장
한우종 (W.J. Han, antla21@yahoo.com) 슈퍼컴퓨팅시스템연구실 연구위원

ABSTRACT

High performance computing (HPC), once reserved for large-scale scientific research and industrial simulations, has rapidly evolved because of the massive computational demands of modern artificial intelligence (AI) workloads. This paper examines the architectural evolution of HPC processors to maximize the processing efficiency of these AI-driven tasks, focusing on advances in architecture, packaging, interconnect, and memory interface technologies. We conducted a comprehensive analysis of the latest HPC processors from Intel, AMD, and NVIDIA, highlighting their structural characteristics and interface innovations. The analysis reveals how chiplet-based packaging, high-bandwidth memory integration, and high-speed interconnects are being implemented in modern HPC processor designs. Based on these findings, this paper presents the emerging directions of next-generation HPC infrastructure architectures optimized for AI workloads.

KEYWORDS HPC 프로세서, 고대역폭 메모리, 고성능컴퓨팅 프로세서, 고속 인터커넥트

I. 서론

일반적으로 고성능 컴퓨팅(HPC: High Performance Computing)은 여러 대의 고사양 컴퓨팅 서버를 대규모로 연결해 복잡한 계산을 빠르게 수행하는 기술을 말한다[1]. 이러한 HPC 기술은 슈퍼컴퓨터로 구현되어 기상 예측, 신약 개발, 핵융합 등 대규모 계산이 필요한 과학 연구와 산업 시뮬레이션에 활용되어 왔다. 이에 'HPC'라는 용어는 응용 분야와 이를 뒷받침하는 컴퓨팅 인프라 전체를 포괄하는 개념으로 확장되어 사용되고 있었다.

최근 전 세계적으로 급성장하고 있는 인공지능(AI: Artificial Intelligence)은 HPC 기술에 새로운 전환점을 가져오고 있다. 특히 수조 개 파라미터를 갖는

* DOI: https://doi.org/10.22648/ETRI.2025.J.410107
* 이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. RS-2025-02304483, 온디바이스 AI 최적화 칩렛 기반 허브 SoC 개발, 90%)과 2025년도 한국전자통신연구원 내부연구과제(25ZV1100, 사업화본부 사업, 10%)의 지원을 받아 수행된 연구임

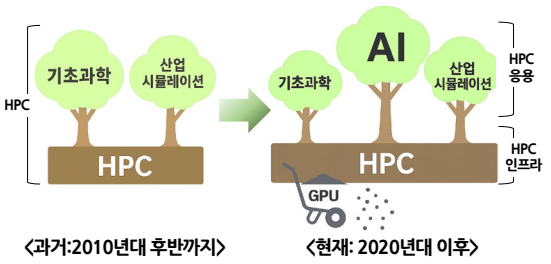

그림 1 HPC 인프라와 응용 분야

최신 대형 언어 모델(LLM: Large Language Model)의 등장으로 AI는 더 이상 단순한 응용이 아니라, 전통적인 과학 계산에 버금가는 연산 성능을 요구하는 사실상의 새로운 HPC 워크로드로 부상하였다. 그림 1에서 보듯, 과거의 HPC가 기초과학 연구와 산업 시뮬레이션을 성장시키는 기술적 토양 역할을 해왔다면, 오늘날 AI 역시 GPU를 중심으로 한 HPC 인프라를 기반으로 성장하고 있다.

이처럼 새로운 HPC 응용 분야로 자리 잡은 AI의 처리 성능은 단일 프로세서만으로는 확보할 수 없고 프로세서-메모리-인터커넥트로 구성된 전체 HPC 인프라의 최적화에 의해 좌우된다. 즉, AI의 고도화는 HPC 인프라의 발전에 기반하므로 HPC 기술의 중요성은 과거 어느 때보다도 더 주목받고 있다.

이에 본고에서는 HPC를 '인프라' 관점에서 고찰하고, 특히 HPC 인프라의 핵심 요소인 프로세서를 중심으로 AI를 효율적으로 처리하기 위한 구조적, 기술적 발전 방향을 분석한다.

본고의 구성은 다음과 같다. 먼저 Ⅱ장에서는 AI 워크로드 처리를 위한 HPC 프로세서의 구조적 진화를 살펴보고, Ⅲ장에서는 HPC 프로세서와 연동하여 대용량 데이터를 효율적으로 처리하기 위한 고속 인터커넥트 기술을 설명한다. Ⅳ장에서는 최신 HPC 프로세서 사례를 기반으로 AI 워크로드를 지원하는 기술 동향을 분석하며, 마지막으로 Ⅴ장에서 향후 HPC 프로세서 기술의 발전 방향을 전망한다.

II. HPC 프로세서 구조의 발전

HPC 프로세서는 대규모 계산과 고도의 병렬 처리 작업을 위해 고성능 연산 유닛과 고속 인터커넥트를 갖춘 고성능 컴퓨팅용 프로세서로 정의할 수 있다. 초기에는 주로 CPU(Central Processing Unit)가 이러한 HPC 프로세서 역할을 담당하였으나 AI의 부상으로 HPC 프로세서는 GPU와 AI 특화 가속기까지 포함하는 구조로 발전하고 있다. 본 장에서는 AI로 인하여 HPC 프로세서가 어떻게 기술적으로 진화하고 있는지부터 살펴본다.

1. 연산 중심 구조에서 연산+데이터 중심 구조로

최근 HPC 프로세서는 단순한 고성능 연산 칩이 아니라 AI 워크로드를 효율적으로 처리하기 위해 구조적으로 진화하고 있다. 일반적으로 AI 워크로드는 대규모 데이터에 대한 병렬 계산의 반복 작업들이기 때문에 HPC 프로세서는 계산 성능뿐만 아니라 대규모 데이터의 병렬, 분산처리를 극대화하기 위해 다음과 같은 구조적 특징을 가진다.

첫째, 코어 구조 측면에서 기존 프로세서는 수십 개의 범용 코어를 중심으로 순차적 명령 실행에 최적화된 구조인 반면, 최근 HPC 프로세서는 AI 학습의 대규모 병렬 연산을 처리하기 위해 수백~수천 개의 연산 유닛을 병렬로 구동하는 SIMD(Single Instruction Multiple Data)/SIMT(Single Instruction Multiple Thread) 구조를 채택한다.

둘째, 캐시 및 메모리 구조도 대용량 데이터의 효율적 처리를 위해 발전하고 있다. 기존 프로세서

가 L1~L3 중심의 계층형 캐시를 기반으로 하는 데 비해 HPC 프로세서는 고대역폭 메모리, 이종 프로세서 간 공유 메모리, 온칩 버퍼, 또는 대용량 레지스터 파일 등을 도입하여 데이터 병목을 최소화한다.

셋째, 최근 HPC 프로세서는 기존 프로세서 대비 대규모 연산 유닛을 연결해야 하므로 온칩 네트워크(NoC: Network-on-Chip) 기반의 고속 패브릭 구조를 채택하고 있다. NoC는 코어, 캐시, 메모리 컨트롤러 간 고대역폭/저지연 통신을 제공함으로써 버스 연결보다 대규모 병렬 처리의 확장성과 통신 효율을 크게 향상시킬 수 있다. 또한, 집적도 한계를 극복하고 연산유닛의 다양성도 확보하기 위해 칩렛 인터커넥트 채택도 활발히 일어나고 있다.

넷째, 메모리 접근 속도가 프로세서 연산 속도를 따라가지 못해 데이터 병목을 일으키는 메모리 월(Memory Wall) 문제는 일반 프로세서보다 대규모 데이터 계산 위주의 AI 워크로드를 다루어야 하는 HPC 프로세서에서 훨씬 심각해진다. 이를 완화하기 위해 메모리 칩 적층 패키징 기술, 고속 메모리 인터페이스 기술, 메모리 근접 계산 기술들이 발전하고 있다. 이 중 메모리 칩 적층 패키징 기술은 본장 4절에서, 고속 메모리 인터페이스 기술은 Ⅲ장 5절에서 상세히 설명한다.

2. CPU(멀티코어)에서 GPU(매니 코어)로

초기 HPC 프로세서들은 CPU의 클록 속도 향상과 파이프라인 최적화를 통해 성능을 높여왔다. 그러나 반도체 공정이 미세화의 한계를 보이면서 CPU의 성능 역시 더디게 향상되고 있었다. 더욱이 AI 모델이 요구하는 대규모 행렬 연산과 높은 병렬 처리를 CPU의 직렬 처리 구조만으로 감당할 수 없게 되었다.

이러한 상황에 수천 개의 연산 유닛을 동시에 구동할 수 있는 GPU가 HPC 프로세서의 새로운 대안으로 제시되었다. GPU는 SIMD/SIMT 기반의 대규모 병렬 처리 구조를 통해 연산 처리량을 획기적으로 높였으며, CUDA 등 병렬 프로그래밍 모델을 함께 제공함으로써 AI 모델을 위한 소프트웨어 생태계를 안정적으로 구축하였다. 이러한 점에서 GPU는 오늘날 대표적인 AI 처리용 프로세서로 자리매김하고 있다.

3. 범용(General Purpose)에서 AI 특화(Special Purpose)로

병렬 연산 성능이 뛰어나 CPU를 대체하고 있는 GPU도 당초 범용 가속기로 설계되었기 때문에 저정밀, 대규모 행렬 연산 위주의 AI 워크로드를 처리하기에 비효율적 측면이 있다.

이런 문제를 해결하기 위해 AI 워크로드의 특성을 반영한 AI 특화 가속기들이 속속 등장하였는데, 대표적으로 구글의 Tensor Processing Unit(TPU), 하바나 랩스(현, 인텔)의 Gaudi, 세레브라스의 Wafer-Scale Engine(WSE), NVIDIA의 GPU 내 행렬 연산 전용 유닛인 Tensor Core 등이 그것이다. 이들은 행렬 연산기(Multiply-Accumulate Array)와 AI 전용 함수, 낮은 정밀도 데이터 연산을 지원함으로써 AI 워크로드의 병렬 처리에 뛰어난 특성을 보인다. 이에 기존의 GPU도 낮은 정밀도, AI 특화 연산을 집중적으로 지원하기 시작하며 GPU와 AI 가속기의 경계가 줄어들고 있다.

4. 평면 패키징에서 공간 패키징으로

HPC 프로세서에 점점 더 많은 연산 유닛과 메모리 대역폭이 요구되면서 기존의 평면적(2D) 단일 칩

설계는 I/O 핀 수, 배선 밀도, 발열 제약 등의 물리적 제약으로 인해 성능 향상의 한계에 직면하게 되었다. 이를 극복하기 위한 해법으로 칩렛(Chiplet), 2.5D, 3D 패키징 등 패키지 수준의 공간 통합 기술이 부상하고 있다.

칩렛은 초기의 모노리식(Monolithic) 구조와 달리, CPU, GPU, I/O, 메모리 컨트롤러 등을 기능별로 분리한 뒤 각각 최적 공정에서 제작해 하나의 패키지로 통합하는 방식이다. 이를 통해 성능, 전력 효율, 비용을 균형 있게 확보할 수 있으며 현대 HPC 프로세서의 기본 구조로 자리 잡았다.

또한, 칩 간 연결 효율을 높이기 위해 2.5D 패키징과 3D 적층 패키징이 함께 활용되고 있다. 2.5D 패키징은 대형 실리콘 인터포저(TSMC의 CoWoS: Chip on Wafer on Substrate)나 임베디드 브리지(인텔의 EMIB)와 같은 고밀도 연결 기술을 활용하여 여러 칩을 단일 패키지 내에 근접 배치하는 기술로서 신호 경로를 줄이고 고대역폭 · 저지연 데이터 전송을 가능하게 한다. 3D 패키징은 Through-Silicon Via(TSV)를 이용해 다이(Die)를 수직으로 적층함으로써 칩 간 신호 경로를 더 단축하고 공간 효율을 극대화한다.

이러한 패키징의 진화는 프로세서-메모리 간 성능 격차로 인한 병목 현상(메모리 월)을 완화하는 핵심 기술로도 작용하고 있다. 특히 DRAM(Dynamic Random Access Memory) 다이를 TSV로 수직 적층하는 3D 패키징 기술로 만들어진 HBM(High-Bandwidth Memory)을, 연산 칩과 2.5D 패키징(인터포저)으로 통합하는 구조는 대용량 AI 워크로드에서 필수 요소로 자리 잡았다. 예를 들어, NVIDIA 최신 GPU인 B200/B300 GPU는 HBM3e를 통해 최대 약 8TB/s의 초고대역폭을 제공하고 있으며, AMD Instinct MI300A는 CPU, GPU, HBM3를 하나의 패키지에 통합하여 약 5.2TB/s의 대역폭을 구현하고 있다.

결과적으로 칩렛, 2.5D, 3D 적층으로 대표되는 공간 패키징 기술은 HPC 프로세서를 고성능 · 고집적 · 고효율 구조로 이끄는 핵심 동력으로 자리 잡고 있다.

5. 성능 중심에서 전력당 성능 효율 중심으로

공정 미세화로 트랜지스터 밀도가 높아지면서 발열과 전력 밀도 상승이 HPC 프로세서 설계의 큰 제약이 되고 있다. 이에 따라 단순히 클록을 높여 성능을 끌어올리던 과거의 방식에서 벗어나 소비전력 대비 성능, 즉 전력당 성능 효율(Power Efficiency)을 얼마나 높일 수 있는지가 핵심 설계 기준으로 자리 잡고 있다[2,3].

이러한 변화는 단일 대형 코어 중심의 구조보다 다수의 저전력 코어를 병렬로 배치하는 구조로 전환시키고 있으며, 연산 특성에 따라 CPU, GPU, AI 가속기를 조합하는 이기종 컴퓨팅 구조를 보편화시키고 있다. 이 구조는 각 연산 장치가 최적화된 영역만을 담당하게 하여 시스템 전체의 전력당 성능 효율을 실질적으로 끌어올리는 데 기여한다.

전력 효율 중심 설계 전환은 실제 성능 지표에서도 확인된다. 최근 Green500[1] 결과에 따르면, 가속기 중심의 시스템들이 50~70GFLOPS/W로 상위를 차지하는 반면, 전통적인 CPU 중심 시스템은 대부분 8GFLOPS/W 이하로 100위권 밖에 머무른다[4]. 이는 GPU, AI 가속기, HBM, 고속 인터커넥트 기술

1) Green500은 세계 슈퍼컴퓨터의 에너지 효율을 평가하는 공식 지표로, LINPACK(HPL) 벤치마크에서 달성한 연산 성능(GFLOPS)을 소비전력(W)으로 나눈 값(GFLOPS/W)을 기준으로 순위를 매긴다. 이는 HPC 시스템의 전력당 성능 효율(Power Efficiency)을 비교하는 대표적인 국제 표준이다.

통합이 시스템 수준의 전력 효율을 실질적으로 향상시키고 있음을 보여준다. 결국 전력당 성능 효율은 병렬 처리 성능과 에너지 효율의 최적 조합으로 결정되며 HPC 시스템의 주요 성능 지표로 자리 잡고 있다.

III. 고속 인터커넥트 기술의 발전

HPC 프로세서의 성능은 코어 수나 클록 속도로만 결정되지 않는다. 수천 개의 연산 코어를 탑재하거나 수천 대의 서버가 병렬로 구성된 시스템이라도 데이터가 오고 가는 통로(인터커넥트)가 이를 뒷받침하지 못하면 확장성의 한계에 부딪힌다. 결국 전체 시스템의 성능을 좌우하는 확장성(Scalability)과 확장 효율(Scaling Efficiency)은 인터커넥트 기술에 의해 좌우된다고 할 수 있다.

인터커넥트 기술은 프로세서 내부의 연산 코어뿐만 아니라 CPU, GPU, AI 가속기, 메모리 등 다양한 이기종 구성 요소를 고속으로 연결해 데이터가 병목 없이 전달되도록 하는 핵심 기술이다. 특히 AI 워크로드의 데이터양 증가와 병렬 처리 요구의 확대에 따라 고대역폭 · 저지연 인터커넥트의 중요성은 더 커지고 있다. 인터커넥트는 크게 인터페이스와 프로토콜로 구성되며, 인터페이스는 시스템 구성 요소 간 데이터를 주고받는 물리적 연결 통로를, 프로토콜은 그 통로를 통해 데이터를 교환하는 규칙과 제어 방식을 의미한다.

그림 2는 HPC 프로세서에서 사용 중이거나 최근 표준화가 진행 중인 고속 인터커넥트 기술들의 적용 위치를 보여준다. 본 장에서는 이 그림에 나타난 PCIe, CXL, UCIe, UALink, 그리고 메모리 인터페이스 기술들을 살펴본다.

1. Peripheral Component Interconnect Express(PCIe)

PCIe[16]는 HPC 시스템과 데이터센터에서 CPU와 다양한 주변 장치를 연결하는 범용 인터페이스로 GPU, AI 가속기, 스토리지, 고속 네트워크 어댑터 등 대부분의 고성능 장치를 시스템에 연결한다.

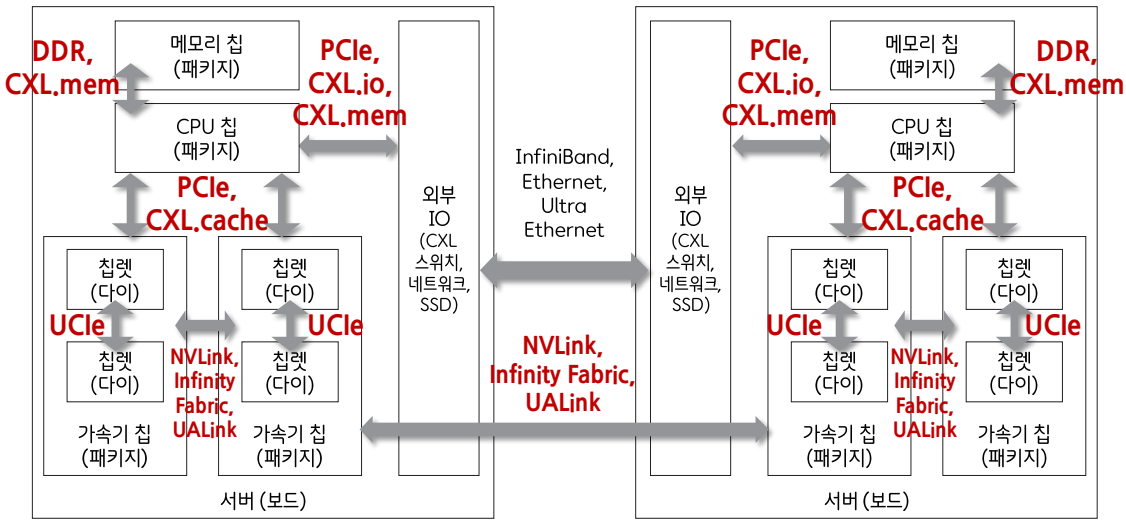

그림 2 HPC 프로세서 중심의 주요 고속 인터커넥트 기술 예시

PCIe는 송 · 수신 쌍으로 구성된 레인을 x1[2], x4, x8, x16 등으로 자유롭게 확장할 수 있어 장치 특성에 맞게 대역폭을 유연하게 조정할 수 있는 장점이 있다.

PCIe 5.0, PCIe 6.0 등 세대가 발전할수록 레인 속도가 증가하여 GPU와 CPU 간 대량의 학습 데이터를 빠르게 교환할 수 있고, NVMe SSD(Solid-State Drive) 스토리지 및 인피니밴드(IB)나 기가비트 이더넷(GbE) 등의 고속 네트워크 어댑터 역시 PCIe 기반으로 확장됨으로써 대규모 병렬 입출력을 지원할 수 있다. 또한, PCIe는 스위치 기반 구조를 사용해 각 장치가 독립 경로를 확보하므로 전통적인 병렬 버스 구조의 대역폭 공유 문제를 효과적으로 해소하고 시스템 확장성도 높다[5,16].

그러나 레인 수 증가 시 전력 소모와 신호 무결성 문제를 관리해야 하며, 멀티 GPU, 대규모 가속기 환경에서는 PCIe만으로 대역폭이 부족해 병목이 발생할 수 있다. 이 때문에 최근 HPC 시스템에서는 PCIe를 기본 I/O 인프라로 유지하면서, GPU 간에는 NVLink, CPU-메모리 간에는 CXL 등 전용 인터커넥트를 병행하는 다층적 연결 구조로 진화하고 있다.

2. Compute Express Link(CXL)

CXL[17]은 PCIe 물리 계층을 기반으로 하되 그 위에 캐시 일관성(Cache Coherency)과 저지연 데이터 공유 기능을 추가한 차세대 인터커넥트 표준이다. 인텔, AMD, Arm, 삼성, 마이크로소프트 등이 참여하는 CXL 컨소시엄에서 개발된 개방형 규격으로 CPU, GPU, AI 가속기, 메모리 간 효율적인 데이터 공유와 확장성 확보를 목표로 한다. CXL은 다음과 같이 세 가지 프로토콜로 구성된다[6-8].

- **CXL.io:** PCIe와 동일한 트랜잭션 계층을 사용하여 장치의 설정, 제어, 상태 모니터링, 인터럽트 처리 등 기본적인 I/O 기능을 담당한다.
- **CXL.cache:** 가속기가 캐시 일관성을 유지하며 CPU 메모리에 직접 접근하도록 지원한다. 즉, GPU, FPGA, AI 가속기가 CPU DRAM의 데이터를 로컬 캐시로 가져와 수정하더라도 CPU캐시 및 DRAM과 일관성을 유지할 수 있다.
- **CXL.mem:** CXL 장치(CXL 메모리 확장 모듈, CXL 메모리 풀 등)의 메모리를 CPU가 자신의 주소 공간에서 직접 접근(Load/Store 방식)할 수 있도록 지원한다.

이 구조에서 CXL은 PCIe의 대역폭과 호환성을 유지하면서도 CPU, GPU, AI 가속기 간 데이터를 복사하지 않고 공유 메모리에 직접 접근할 수 있게 해준다. 이를 통해 AI 학습 및 추론 과정에서 발생하는 불필요한 데이터 이동을 줄이고 전체 시스템의 처리 효율을 높일 수 있다. CXL.mem을 활용하면 DRAM, NAND 플래시 등으로 대용량 메모리 풀 구성도 가능하다. 또한, CXL 스위치를 사용하면 여러 구성 요소를 유연하게 연결, 분리할 수 있어 필요에 따라 재구성이 가능한 확장형 시스템을 구현할 수 있다[9].

현재는 초기 상용화 단계이지만 주요 반도체, 클라우드 기업들이 CXL 2.0, 3.0을 지원하는 CPU, 메모리 확장 모듈, 서버 보드를 공개하며 기술 검증을 진행 중이다. 특히 메모리 풀링(Pooling)과 리소스 디스어그리게이션(Disaggregation)의 시험 결과가 보고되면서[10,11] CXL은 데이터센터와 HPC 환경에서 메모리 집약적 워크로드를 위한 해결책을 제공할 수 있을 것으로 전망된다.

2) xN은 PCIe 레인이 N개라는 뜻임

3. Universal Chiplet Interconnect Express(UCIe)

칩(패키지)과 외부의 장치를 연결하는 PCIe나 CXL과 달리, UCIe[12,13,18]는 단일 칩(패키지) 내부에서 칩렛(다이) 간 초고속 · 저전력 데이터 전송을 제공하는 다이-투-다이(Die-to-Die) 인터커넥트 표준이다. 인텔, AMD, Arm, TSMC, 삼성, 구글, 마이크로소프트 등 주요 반도체 기업이 참여하는 개방형 컨소시엄에서 빠르게 표준화가 진행되고 있다.

UCIe의 핵심 목적은 서로 다른 기능, 공정, 제조사의 칩렛을 동일 패키지 내에서 호환 가능하게 만드는 것이다. 이를 통해 CPU, GPU, AI 가속기, 메모리 컨트롤러, I/O 다이 등을 칩렛 단위로 조합하여 하나의 SoC처럼 구성할 수 있어 고성능 프로세서 설계의 모듈화와 이기종 통합을 용이하게 한다. 특히 PCIe, CXL 등을 프로토콜 계층에서 캡슐화(Encapsulation)한 형태로 전송할 수 있어 기존 소프트웨어, 하드웨어 생태계를 유지하면서도 패키지 내부 인터커넥트의 성능을 극대화할 수 있다는 장점이 있다.

UCIe는 레인당 수십 Gbps와 레인 확장을 통한 TB/s급 대역폭을 제공하며, 패키지 내부에서 수 pJ/bit 수준의 매우 낮은 전송 에너지를 달성하는 초고효율 인터커넥트이다. 마이크로범프, TSV 기반 근접 연결을 통해 신호 손실과 지연을 최소화해 칩렛 기반 HPC 프로세서가 요구하는 초고대역폭 · 초저지연 데이터 이동을 지원한다.

다만 칩렛 정렬, 전력, 클록 동기화 등 패키징 공정 난이도가 높고 적용 범위가 패키지 내부로 제한된다는 점이 한계로 지적되고 있다. 현재 3.0 규격이 발표되었으며, 향후 2.5D, 3D 패키징과 결합해 HBM, AI 칩렛, I/O 다이 등을 통합하는 온-패키지 고대역폭 구조로 확장될 전망이다. 이러한 진화는 모노리식 구조에서 칩렛 기반 프로세서로의 전환을 더 가속하는 핵심 동력으로 평가된다.

4. Ultra Accelerator Link(UALink)

UALink[14,15,19]는 GPU(또는 AI 가속기) 간 초고속 · 저지연 통신을 위한 개방형 인터커넥트 표준으로 PCIe 기반 연결의 확장성 한계와 NVIDIA의 NVLink, AMD의 Infinity Fabric 등의 벤더 종속성을 해소하기 위해 2024년 AMD, 인텔, 마이크로소프트, 메타, 구글, 시스코, 브로드컴, HPE 등이 공동 제안한 기술이다.

UALink는 GPU(또는 AI 가속기) 간 직접 데이터 전송을 지원해 CPU 경유에 따른 병목을 줄이며 대규모 AI 학습에서의 통신 지연 문제를 효과적으로 완화할 수 있다. 공개된 1.0 규격에 따르면 최대 1,024개의 가속기를 하나의 풀로 구성할 수 있어 초대형 AI 클러스터 구축에 적합하며, PCIe 대비 높은 링크 속도와 NVLink 수준 이상의 대역폭을 목표로 함으로써 GPU 중심의 HPC 시스템 확장 효율을 높일 수 있다. 이더넷 PHY(물리계층) 기반으로 구현해야 하며, 이로 인해 전력 소모 증가와 라우팅 복잡성, 지연시간 등의 기술적 과제가 뒤따른다. 하드웨어 생태계도 아직 초기 단계여서 확산을 위한 생태계 조성 노력이 요구된다.

향후 생태계의 성숙을 전제로, 이더넷 PHY 기반이라는 특징을 활용하여 스케일 아웃[3] 인터커넥트와 통합 관리 구조로 발전할 가능성이 클 것으로 예상되는바 UALink는 대규모 이종 가속기 패브릭을 구성하기 위한 핵심 기술로 자리 잡을 수 있을 것이다.

3) 스케일 아웃(Scale-Out): 시스템의 처리 능력을 확장하는 방식 중 하나로, 기존 시스템에 더 많은 서버나 컴퓨팅 노드를 병렬로 추가하여 전체 용량을 늘리는 방식

5. 메모리 인터페이스 기술

과학 계산이나 AI 워크로드의 성능은 연산 코어뿐만 아니라 메모리 대역폭과 지연에 크게 좌우된다. 특히 대규모 행렬 연산을 반복 수행하는 AI 학습에서는 연산 속도보다 DRAM 접근 속도가 훨씬 느려 발생하는 메모리 병목이 성능 향상의 핵심 걸림돌이 되고 있다. 이러한 한계를 해소하기 위해 최근 HPC 프로세서들은 단일 메모리 기술에 의존하는 대신 고대역폭 · 근접 메모리(HBM)와 대용량 외부 메모리(DDR5/LPDDR5X)를 결합한 계층형 메모리 구조로 빠르게 전환되고 있다.

HBM3[20]는 TSV 기반 3D 적층 구조를 통해 초고대역폭을 제공하는 근접 메모리로 GPU 및 AI 가속기와 실리콘 인터포저를 통해 직접 연결된다. 핵심 연산 데이터가 연산 유닛 바로 인접한 위치에서 처리되므로 대역폭 병목을 최소화하며 지연을 크게 줄일 수 있다. 반면 DDR5[21]는 폭넓은 용량 확장성과 채널 병렬성이 강점인 범용 메모리 인터페이스로, 대규모 파라미터와 데이터셋을 저장하는 외부 대용량 메모리 계층을 담당한다. LPDDR5X[22]는 본래 모바일용 저전력 메모리이지만 전력 효율과 용량 대비 성능이 우수해 NVIDIA Grace CPU 등 서버급 SoC에도 채택되어 저전력 · 대용량 보조 메모리로 활용되고 있다.

이러한 구성은 HBM을 통해 고대역폭 · 저지연 처리를 보장하고, DDR5/LPDDR5X를 통해 대용량 데이터를 안정적으로 제공하는 이중 계층 메모리 구조를 형성한다. 대표적으로 NVIDIA Grace Hopper는 HBM3와 LPDDR5X를 결합하여 근접 메모리와 보조 메모리 간 역할을 분리하였으며[30,31], AMD MI300A[26-29]와 인텔 Max 시리즈(Sapphire Rapids HBM)[24,25] 역시 CPU와 HBM을 단일 패키지로 통합해 지연과 전력 소모를 동시에 줄이는 방향으로 설계되고 있다.

더 나아가 이러한 계층형 구조는 CXL 기반의 공유 메모리 패브릭과 결합되며 CPU-GPU-가속기 간 메모리 일관성, 공간 확장성, 자원 활용 효율을 동시에 높이는 방향으로 진화하고 있다.

6. 비교 분석

표 1은 앞서 설명한 주요 인터커넥트 및 메모리 인터페이스 기술을 표준화 기관별로 분류하고, 각 표준 규격을 기준으로 버전, 적용 용도, 목표 속도, 지연시간 등을 비교 정리한 것이다.

PCIe는 가장 범용적인 외부 인터페이스지만 지연이 상대적으로 높다는 한계를 가진다. CXL은 PCIe PHY 기반 위에 캐시 일관성 기능을 추가하여 CPU-가속기-메모리 간 공유 메모리 접근과 메모리 확장을 지원한다. UCIe는 패키지 내부의 칩렛 간 연결을 위한 초저지연 · 초고대역폭 인터페이스로 칩렛 기반 이기종 통합의 핵심 기술이다. UALink는 벤더 종속성을 탈피해 이기종 GPU(또는 AI 가속기) 간 직접 통신을 지원하는 개방형 표준으로 NVLink급 성능과 대규모 가속기 확장을 목표로 개발 중이다. 메모리 계층 측면에서 HBM3는 가장 빠른 근접 메모리로 활용되며, DDR5 및 LPDDR5X는 대용량 · 저전력 외부 메모리 역할을 담당한다.

AI 모델의 급격한 대형화로 인해 이들 계층별 인터커넥트 및 메모리 기술은 시스템 설계에서 그 중요성이 더 커지고 있다. 시스템의 성능과 효율을 극대화하기 위해서는 각 기술이 상호 보완적으로 작동하는 구조적 연계가 필수적이다. 이에 따라 패키지 내부(UCIe), 패키지 간(UALink), 시스템 외부(PCIe/CXL), 그리고 HBM-DDR로 구성되는 다층 메모리 계층이 각기 다른 역할을 수행하면서도 유기적으로 연동되는 통합형 아키텍처로 진화하고 있

표 1 주요 고속 인터커넥트 기술 비교

비교 기준	PCIe[16]	CXL[17]	UCIe[18]	UALink[19]	HBM[20]	DDR[21]	LPDDR[23]
표준화 기관	PCI-SIG (국제 표준)	CXL 컨소시엄 (개방형 표준)	UCIe 컨소시엄 (개방형 표준)	UALink 그룹 (개방형 표준)	JEDEC 표준 (국제 표준)	JEDEC 표준 (국제 표준)	JEDEC 표준 (국제 표준)
버전 이력	1.0(2003년) ~7.0(2025년)	1.0(2019년) ~3.2(2024년)	1.0(2022년) ~3.0(2025년)	1.0(2025년)	HBM (2013년) ~HBM4 (2025년)	DDR (2000년) ~DDR5 (2020년)	LPDDR (2009년) ~LPDDR6 (2025년)
적용 위치	패키지 외부 (CPU-가속기/SSD/ 네트워크)	패키지 외부 (CPU-가속기-메모리)	패키지 내부 (칩렛 간)	패키지 외부 (GPU 간)	패키지 내부 (근접 메모리)	패키지 외부 (DIMM)	패키지 외부 (SoC)
주요 용도	범용 확장성	메모리 공유, 캐시 일관성	칩렛 통합	가속기 간 직접 통신, 확장성	고대역 메모리	대용량 메모리	저전력 메모리
최신 버전 목표 속도* (대역폭)	PCIe 7.0[16]: 128Gbps	CXL3.2 (PCIe 6.0 기반)[17]: 64Gbps	UCIe 3.0[18]: 32Gbps	UALink 1.0[19]: 200Gbps	HBM4[20]: 8.0Gbps	DDR5[21]: 6.4Gbps	LPDDR6[23]: 14.4Gbps
최신 버전 목표 지연시간**	PCIe 7.0[16]: 6.0 대비 비트당 지연 시간 추가 단축 - 시스템 전체: 100~500ns	CXL3.2[17]: -CXL.cache: 수십ns -CXL.mem: 100~200ns -CXL.io: 100~200ns	UCIe 3.0[18]: 수ns~수십ns	UALink 1.0[19]: ~100~150ns (NVLink 수준)	HBM4[20]: <100ns	DDR5[21]: <50~80ns	LPDDR6[23]: 16~19ns

* 표준 규격의 목표치(제품은 더 높을 수 있음), 단일 레인 또는 단일 핀, 단방향 대역폭 기준으로 Gbps 단위로 통일

** 링크 레이어(Link-Layer) 또는 물리 레이어(Physical Layer) 기준

다. 이처럼 계층별로 특화된 기술들의 상호 보완적 구성은 향후 AI 및 기존 HPC 워크로드를 동시에 처리하기 위한 HPC 시스템 설계의 핵심 전략이자 진화 방향이 될 것이다.

IV. 최신 HPC 프로세서 사례 분석

Ⅱ장과 Ⅲ장에서 살펴본 바와 같이 AI 워크로드를 지원하는 최신 HPC 프로세서는 칩렛 기반 구조, 고속 인터커넥트, 고대역폭 메모리 통합을 중심으로 발전하고 있다. 본 장에서는 이러한 기술적 흐름이 실제 제품에서 어떻게 구현되고 있는지를 살펴본다. 대표적 최신 HPC 프로세서인 Intel 4세대 Xeon Max시리즈(미국 아르곤 국립연구소의 Aurora 슈퍼컴퓨터에 탑재), AMD Instinct MI300A(미국 오크리지 국립연구소의 Frontier 슈퍼컴퓨터, 로렌스 리버모어 국립연구소의 El Capitan 슈퍼컴퓨터에 탑재), 그리고 NVIDIA Grace Hopper Superchip(독일 율리히 슈퍼컴퓨팅 센터의 JUPITER 슈퍼컴퓨터에 탑재)을 중심으로, 각 프로세서의 구조, 메모리, 인터커넥트 측면에서 공통점과 차이점을 비교 및 분석한다.

1. 인텔 4세대 Xeon Max 시리즈

인텔의 4세대 Xeon Max 시리즈(이전 코드명: Sapphire Rapids HBM)는 CPU 중심 서버 아키텍처가 AI 시대의 고성능 컴퓨팅 요구사항에 대응하여 진화한 대표적 사례이다[24,25].

이 프로세서는 최대 4개의 CPU 타일(Compute IP)을 EMIB(Embedded Multi-Die Interconnect Bridge)로

연결한 모듈러 패키징을 도입하였으며, 이는 칩렛과 2.5D 패키징 기술이 실제 서버 CPU 제품에 구현된 형태를 보여준다. 기능적으로 CPU 타일에 Advanced Matrix Extensions(AMX)를 탑재하여 기존 벡터 연산(AVX-512)에 대비하여 행렬 연산 처리 성능을 대폭 강화하였다. 메모리 측면(Memory IP)에서는 DDR5과 HBM2e를 탑재해 메모리 병목 현상을 완화한다. 인터커넥트 측면에서는(I/O IP) CPU 간 연결에 인텔 UPI(Ultra Path Interconnect)를 적용하고, 외부 연결은 PCIe 5.0 및 CXL 1.1을 통해 가속기와의 고속 통신 및 확장 메모리 공유를 지원한다.

Xeon Max 시리즈는 CPU 기반 서버 아키텍처의 틀을 유지하면서도 고대역폭 메모리와 AI 가속 기능을 통합한 AI 적응형 HPC 프로세서라 할 수 있다.

2. AMD Instinct MI300A

AMD Instinct MI300A는 CPU, GPU, HBM을 단일 패키지에 통합한 APU(Accelerated Processing Unit)형 HPC 프로세서이다[26-29].

참고문헌 [26-29]에 제시된 MI300A는 Zen 4 기반 CPU 칩렛(CCD) 3개, CDNA3 기반 GPU 칩렛(XCD) 6개, 4개의 I/O 다이(IOD), 그리고 8개의 HBM3 스택(최대 128GB)을 3D 적층 및 고밀도 인터포저 기반의 패키징으로 구성된다. 이 구조는 각 기능 블록을 최적 공정에서 제조하면서도 단일 SoC처럼 동작할 수 있도록 하며, 특히 HBM3를 CPU와 GPU에 근접 배치함으로써 메모리 병목을 효과적으로 완화하고 AI 학습 및 과학 계산 모두에서 성능 향상에 이바지할 수 있다. 칩 간 연결은 AMD의 고속 인터커넥트 기술인 Infinity Fabric을 기반으로 하여 CPU와 GPU 간 메모리 일관성을 유지하면서 대규모 병렬 연산에 필요한 고대역폭 · 저지연 데이터 전송을 지원한다. 또한 PCIe 5.0 및 CXL 2.0도 적용해 외부 가속기 및 메모리 확장 장치와의 연결성과 시스템 확장성도 확보하였다.

이처럼 MI300A는 CPU-GPU-메모리-인터커넥트가 패키지 수준에서 긴밀히 통합됨으로써 데이터 이동 비용을 최소화하고 전력당 성능 효율을 극대화할 수 있어 AI와 전통적인 HPC 응용을 모두 효과적으로 지원할 수 있는 차세대 HPC 인프라 설계 방향을 보여주는 사례로 평가된다.

3. NVIDIA Grace Hopper 슈퍼칩

Grace Hopper 슈퍼칩은 Grace CPU 칩와 Hopper GPU 칩을 단일 패키지에 통합한 HPC 프로세서이다. 인텔과 AMD가 CPU 중심으로 아키텍처를 확장해 온 것과 달리 NVIDIA는 GPU 전문성을 기반으로 CPU와 GPU를 융합한 슈퍼칩을 개발하였다[30,31].

NVLink-C2C는 패키지 내부에 최적화된 칩-투-칩 전용 링크로 양방향 최대 900GB/s의 대역폭과 칩 간 메모리 일관성을 보장한다. 이를 통해 Grace와 Hopper는 서로의 메모리를 하나의 공유 메모리처럼 접근할 수 있어 데이터 복사 없이 자원 공유와 실시간 연산이 가능하다.

메모리는 HBM3(Hopper용)과 LPDDR5X(Grace용)를 결합한 구조로 초고대역폭과 대용량 요구를 동시에 충족한다. 여기에 2.5D CoWoS-L(CoWoS-Local Silicon Interconnect/Organic Interposer) 패키징 기술을 적용하여 프로세서와 메모리를 근접 배치함으로써 신호 경로를 단축하고 전력 효율도 향상시켰다. 또한, NVLink 4.0 및 NVSwitch 기반 확장 구조를 통해 GPU 중심의 최대 256개 병렬 시스템 구성이 가능하며, Grace CPU는 CXL 2.0을 지원하여 이기종 컴퓨팅 생태계 전반에 대한 확장성과 호환성을 강화하였다.

이러한 Grace Hopper 슈퍼칩은 CPU와 GPU의 경계를 제거한 통합형 아키텍처를 통해 전통적인 HPC 워크로드와 AI 워크로드를 아우르는 차세대 HPC 플랫폼의 설계 방향을 제시한다고 할 수 있다.

4. 비교 분석

표 2는 앞서 설명한 세 가지 제품을 아키텍처, 인터커넥트, 메모리 등 주요 기술별로 비교 요약한 표이다.

인텔 Xeon Max는 EMIB 기반으로 CPU 타일과 온-패키지 HBM을 통합해 CPU 중심의 HPC 워크로드에 최적화된 구성을 제공한다. 칩렛 기반의 AMD Instinct MI300A는 Infinity Fabric을 통해 CPU 칩렛, GPU 칩렛, HBM을 단일 패키지에 융합했으며, NVIDIA Grace Hopper 슈퍼칩은 NVLink-C2C를 기반으로 Grace CPU, Hopper GPU, HBM3를 단일 패키지에 통합한 구조를 갖춘다. 세 가지 제품 모두 AI 시대의 대규모 워크로드 요구를 충족시키지만 가속기 간 연결에 자사 고유의 인터커넥트 규격(Infinity Fabric, NVLink)을 사용하고 있어 칩 제조사 간의 호환성을 높일 개방형 칩-투-칩 인터커넥트 표준의 필요성을 시사한다.

사례들을 통해 최신 HPC 프로세서들은 칩렛 기반 고효율 설계, CPU-GPU-HBM의 패키지 내 통합, 그리고 고속 인터커넥트 기술의 내재화와 개방화를 병행하며 통합적으로 진화해 나갈 것으로 예상된다.

V. 결론 및 전망

본고는 대규모 AI 모델 확산에 대응한 HPC 프로세서의 구조적, 기술적 진화를 아키텍처, 인터커넥트, 메모리 관점에서 고찰하고 Intel Xeon Max,

표 2 주요 HPC 프로세서 기반 주요 기술 비교

구분	인텔 4세대 Xeon Max 시리즈 (Sapphire Rapids HBM)[24,25]	AMD Instinct MI300A(APU) [26-29]	NVIDIA Grace Hopper Superchip[30,31]
출시년도	2023년 1월	2023년~2024년	2024년
반도체 공정	인텔 7nm	TSMC 5nm (CCD), 6nm (IOD)	TSMC 4nm(N4)/4N 커스텀 공정
아키텍처	CPU 타일 칩렛(AMX 포함)*4개 + HBM2e*4~6개	CPU(Zen 4) 칩렛(CCD)*3개 + GPU(CNDA3) 칩렛(XCD) *6개 + IOD*4개 + HBM3*8개	CPU(Grace) 칩 + GPU(Hopper) 칩 + HBM3e*채널수
칩 패키징, 온 패키지 인터커넥트	-타일 내부: 2D mesh -타일 간: EMIB	-3D 적층+고밀도 인터포저 -CPU-GPU:Infinity Fabric	-2.5D CoWoS-L + 인터포저 -CPU-GPU: NVLink-C2C
칩/가속기 간 인터커넥트	-CPU-CPU: UPI -CPU-외부: PCIe 5.0/CXL 1.1	-APU-APU:Infinity Fabric Link -APU-외부:PCIe5.0/CXL 2.0	-GPU-GPU: NVLink 4.0/ NVSwitch -CPU-외부: PCIe5.0/CXL 2.0
메모리	DDR5 + HBM2e(온패키지) -DDR5: 8채널, 소켓당 최대 4TB -HBM2e(온패키지): 최대 64GB	HBM3(온패키지): 최대 128GB	HBM3e(온패키지) + LPDDR5X -HBM3e(Hopper): 96~144GB -LPDDR5X(Grace): 최대 480GB -Unified Memory 구조
특징 요약	CPU 중심의 Host 프로세서에 HBM을 통합하여 메모리 병목 현상을 해결하고, 외부 GPU 가속기를 관리함	CPU-GPU-HBM 통합 APU로 통합된 자원 접근성을 극대화하는 HPC 노드용 프로세서임	CPU-GPU 통합형 Superchip으로 ARM기반 Grace CPU가 호스트 역할을 수행하여 x86 의존성을 제거하고 NVLink 통신을 최적화함

AMD MI300A, NVIDIA Grace Hopper 등 최신 사례를 분석하였다.

분석 결과, HPC 프로세서는 단순한 범용 CPU 중심을 넘어 AI 워크로드를 전제한 통합형 인프라 플랫폼으로 빠르게 재편되고 있음을 확인할 수 있었다. 특히 HBM 통합, GPU 및 AI 가속기의 패키지 수준 통합, 칩 간 고속 인터커넥트 기술이 HPC 프로세서의 구조적 혁신을 견인하고 있다. 향후 UCIe, CXL, UALink 같은 차세대 인터커넥트 기술을 통해 CPU-GPU-메모리 간 경계는 더욱 모호해지고 이기종 연산 자원의 통합은 심화될 것으로 예상된다.

이러한 CPU-GPU-HBM의 단일 플랫폼 통합 흐름은 소프트웨어 생태계에도 새로운 방향성을 요구하고 있다. 하드웨어별 최적화의 한계를 극복하고 이기종 자원을 효율적으로 활용하기 위해 추상화 계층과 프로그래밍 모델의 중요성이 주목받고 있다. 이에 따라 OpenCL, SYCL과 같은 범용 병렬 프로그래밍 모델이 제시되고는 있지만, 현재 산업은 CUDA, ROCm/HIP 등 하드웨어 특화 플랫폼과 더불어 XLA(Accelerated Linear Algebra), TensorRT, TVM(Tensor Virtual Machine) 같은 도메인 특화 컴파일러 및 자동 최적화 기술 중심으로 발전하고 있다.

결론적으로, HPC 프로세서의 발전은 단순 성능 경쟁을 넘어 HW-SW 공동 설계(Co-Design)를 중심으로 하는 통합 생태계로 확장되고 있다. 향후 HPC 인프라의 경쟁력은 이기종 하드웨어 구성과 이를 효율적으로 운용할 수 있는 소프트웨어 스택의 성숙도에 의해 결정될 것이며, 이는 초거대규모 AI 모델과 엑사스케일(Exascale)급 과학 계산을 위한 핵심 기반이 될 것으로 전망된다.

용어해설

반도체(Semiconductor) 전류의 흐름을 조절할 수 있는 성질을 가진 전자재료(소재)로, 모든 전자기기의 핵심 구성 요소

프로세서(Processor) 입력된 데이터를 계산하고 처리하는 기능을 담당하는 반도체 칩

CPU(Central Processing Unit) 컴퓨터의 중앙 연산 장치로, 프로그램의 명령을 순차적으로 실행하는 범용 프로세서

가속기(Accelerator) CPU가 수행하기엔 비효율적인 특정 연산(예: 행렬 곱, 신경망 연산 등)을 빠르게 처리하도록 설계된 특수목적 프로세서

SoC(System on Chip) CPU, 메모리 컨트롤러, 그래픽, I/O 등 시스템 구성 요소를 하나의 칩에 통합한 형태

다이(Die) 반도체 웨이퍼에서 잘라낸, 회로가 직접 구현된 실리콘 조각으로, 칩의 핵심 연산·저장 기능을 수행하는 실제 동작 단위

패키지(Package) 다이를 외부 환경으로부터 보호하고, 전기적 신호를 보드나 다른 부품과 주고받을 수 있도록 다이를 기판 위에 실장·연결하는 구조체

칩(Chip) 특정 기능을 수행하도록 제작된 다이에 패키지를 씌워 외부와 연결 가능한 완성품으로 메모리 칩·그래픽 칩·AI 칩 등 다양한 형태가 존재

참고문헌

[1] Gartner, “Definition of High-Performance Computing (HPC),” Gartner IT Glossary.

[2] C. Porter, “Energy Efficiency in High-Performance Computing: Balancing Speed and Sustainability,” NVIDIA Technical blog, 2023. 11. 14. https://developer.nvidia.com/blog/energy-efficiency-in-high-performance-computing-balancing-speed-and-sustainability/

[3] ED Sperling, “The Rising Price of Power in Chips,” SemiConductorEngineering. 2024. 3. 14. https://semiengineering.com/the-rising-price-of-power-in-chips/

[4] Green500. https://www.top500.org/lists/green500/2025/06/

[5] S. N. Nag, “Technical Analysis of PCIe to PCIe 6: A Next-Generation Interface Evolution,” World J. Eng. Technol., vol. 11, no. 3, 2023, pp. 504-525.

[6] D. D. Sharma, “CXL 3.0 specification White Paper,” Compute Express Link, 2022. 8. 1. https://computeexpresslink.org/wp-content/uploads/2023/12/CXL_3.0_white-paper_FINAL.pdf

[7] 김선영 외, “CXL 인터커넥트 기술 연구개발 동향,” 전자통신동향분석 38권 5호, 2023. 10, pp. 23-33.

[8] 안후영 외, “CXL 메모리 및 활용 소프트웨어 기술 동향,” 전자통신동향분석 39권 1호, 2024. 2, pp. 62-73.

[9] Yole Group, “CXL technology unlocks memory performance,” 2023. 10. 10. https://www.yolegroup.com/press-release/cxl-technology-unlocks-memory-performance/

[10] J. Wahlgren et al., “Evaluating Emerging CXL-enabled Memory Pooling for HPC Systems,” arXiv preprint, 2022. doi: 10.48550/arXiv.2211.02682

[11] H. K. Lee and J. M. Choi, “CXL-Based Memory Disaggregation for HPC and AI Workloads,” in Proc. Int. Conf. High Perform. Comput. Netw. Storage Anal. (Denver, CO, USA), Nov. 2023.

[12] D. A. Sharma et al., “Universal Chiplet Interconnect ExpressTM An Open Standard for Chiplets,” Hot Chips 2023 Tutorial, 2023. 8.

[13] Synopsys, “UCIe 3.0 Is Here: Synopsys IP Solutions Are Ready,” 2025. 8. 11. https://www.design-reuse.com/blog/56187-ucie-3-0-is-here-synopsys-ip-solutions-are-ready/

[14] Business Wire, “Ultra Accelerator Link Consortium Incorporates; Announces Membership Opportunity,” 2024. 10. 29. https://www.businesswire.com/news/home/20241029998800/en/Ultra-Accelerator-Link-Consortium-Incorporates-Announces-Membership-Opportunity

[15] M. Cooney, “UALink releases inaugural GPU interconnect specification,” Network World, 2025. 4. 9. https://www.networkworld.com/article/3957541/ualink-releases-inaugural-gpu-interconnect-specification.html

[16] PCI-SIG, “PCI Express Base Specification Revision 7.0,” 2025. https://pcisig.com/

[17] CXL Consortium, “Compute Express Link (CXL) Specification 3.2,” 2024. https://computeexpresslink.org/

[18] UCIe Consortium, “Universal Chiplet Interconnect Express (UCIe) Specification 3.0,” 2025. https://www.uciexpress.org/

[19] UALink Consortium, “Ultra Accelerator Link (UALink) Specification 1.0,” 2025. https://ualinkconsortium.org/

[20] JEDEC, “High Bandwidth Memory 4 (HBM4) Standard (JESD238-5),” 2025. https://www.jedec.org/

[21] JEDEC, “DDR5 SDRAM Standard (JESD79-5C),” 2023. https://www.jedec.org/

[22] JEDEC, “LPDDR5X SDRAM Standard (JESD209-5B),” 2022. https://www.jedec.org/

[23] JEDEC, “LPDDR6 SDRAM Standard (JESD209-6),” 2025. https://www.jedec.org/

[24] Intel Corporation, “Intel® Xeon® Scalable Processor Max Series (formerly Sapphire Rapids HBM) Technical Overview,” 2022. 8. 4. https://www.intel.com/content/www/us/en/developer/articles/technical/xeon-scalable-processor-max-series.html

[25] P. Kennedy, “Intel Xeon Max CPU is the Sapphire Rapids HBM Line,” ServeTheHome, 2022. 11. 9. https://www.servethehome.com/intel-xeon-max-cpu-is-the-sapphire-rapids-hbm-line/

[26] AMD, “AMD Instinct MI300A APU Data Sheet,” 2023. https://www.itcreations.com/user-manuals/2145gh-tnmr/amd-instinct-mi300a-apu-data-sheet.pdf?srsltid=AfmBOoqdMjwPxWb9EAucjTNSvtUJAeDOIHYZFaNhVZpUFF85sx0whfU1

[27] AMD, “Introducing AMD CDNA3 Architecture,” AMD white paper, 2025. https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/white-papers/amd-cdna-3-white-paper.pdf

[28] S. K. Moore, “AMD’s Next GPU Is a 3D-Integrated Superchip,” IEEE Spectrum, 2023. 12. 6. https://spectrum.ieee.org/amd-mi300

[29] Quantum Zeitgeist, “MI300A APU Benchmarks Reveal Efficient Inter-Chip Communication for HPC Applications,” 2025. 8. 18. https://quantumzeitgeist.com/mi300a-apu-benchmarks-reveal-efficient-inter-chip-communication-for-hpc-applications/

[30] NVIDIA, “NVIDIA Grace CPU Superchip White Paper,” 2024. https://resources.nvidia.com/en-us-data-center-overview-mc/en-us-data-center-overview/nvidia-grace-cpu-superchip-whitepaper?utm_source=chatgpt.com&xs=1123821

[31] NVIDIA, “NVIDIA GH200 Grace Hopper Superchip Architecture Overview,” 2023. https://resources.nvidia.com/en-us-data-center-overview-mc/en-us-data-center-overview/nvidia-grace-cpu-superchip-whitepaper?utm_source=chatgpt.com&xs=1123821