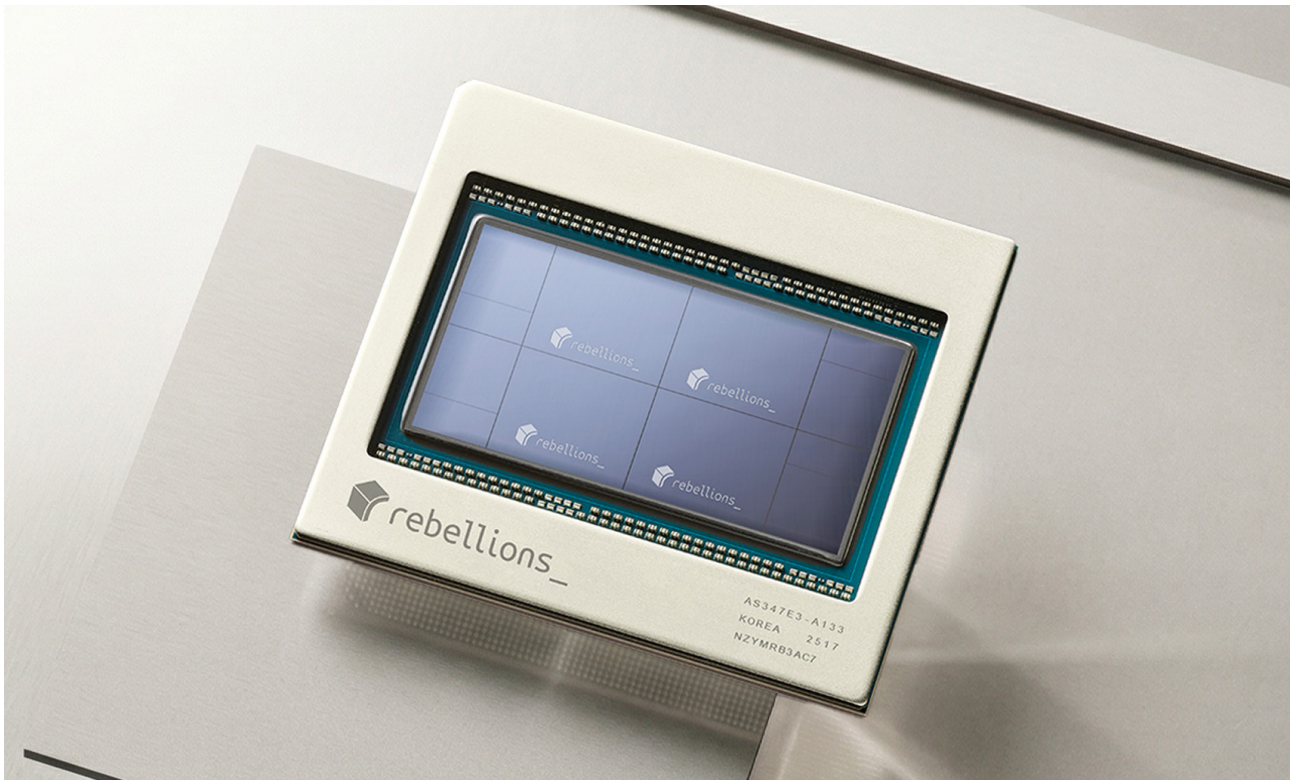
대규모 AI 서빙을 위한 페타스케일 SoC: REBEL-Quad

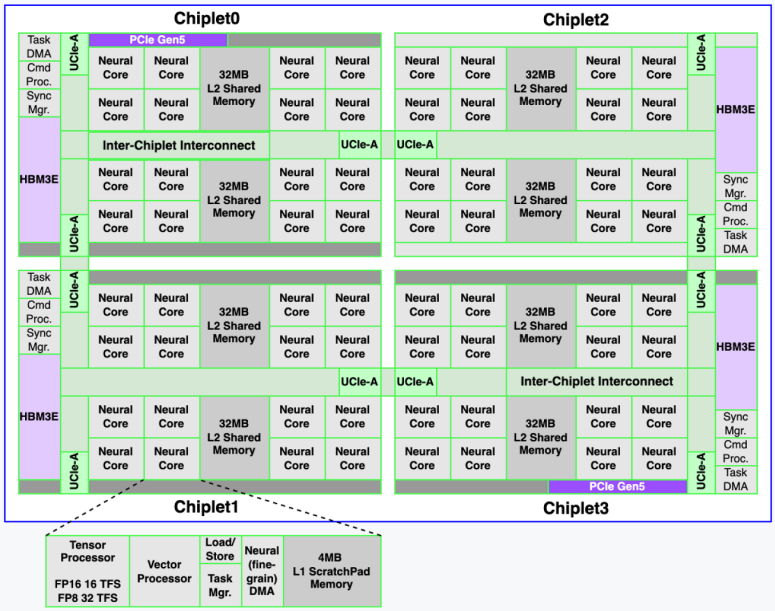
May 02, 2025



REBEL-Quad는 초대형 LLM(대규모 언어모델) 서빙을 위한 **UCIe-Advanced 칩렛 아키텍처 기반 AI SoC**로, 하이퍼스케일러, AI 데이터센터, 엔터프라이즈 환경에서 요구되는 극한의 연산 및 메모리 대역폭 수요를 충족하도록 설계되었습니다.

[대규모 AI 서빙을 위한 페타스케일 SoC]

REBEL-Quad는 하드웨어와 소프트웨어가 완전히 통합된 스택을 기반으로, 연산 집약적인 프리필 **(Prefill) 단계**와 메모리 집약적인 디코딩**(Decoding) 단계** 모두에서 최대 자원 활용도와 탁월한 전력 대비 성능(Performance-per-Watt)을 제공합니다. 또한 칩렛 기반 설계를 통해 저지연성이나 일관성(coherence)을 희생하지 않으면서 초고도 확장성을 구현합니다.



[Figure 1. 네 개의 동형 칩렛으로 구성된 REBEL-Quad 블록 다이어그램]

LLM 추론을 위한 아키텍처적 접근

REBEL-Quad는 대규모 AI 추론에서 발생하는 에너지 효율 및 확장성 문제를 근본적으로 해결하기 위한 구조적 솔루션을 제시합니다.

통합 혼합 정밀도 연산 엔진
Unified Mixed-Precision Compute Engine

- FP8 / FP16 / FP32 연산을 단일 코어에서 통합 지원
- **2.8배 높은 연산 밀도(Compute Density)** 달성

예측형 DMA와 온칩 메시
Predictive DMA & On-Chip Mesh

- 소프트웨어 최적화형 DMA로 **2.7 TB/s 유효 메모리 대역폭 확보**
- 차세대 메쉬 패브릭 기반으로 **3.3배 빠른 코어 간 통신** 구현

REBEL-Quad의 탁월한 전력 효율은 자체 설계된 IP 코어에서 비롯됩니다. 고대역폭 온칩 메쉬 인터커넥트(On-Chip Mesh)는 모든 코어 간 통신을 실시간으로 연결하며, 데이터 흐름의 병목을 제거합니다.

광역 동기화
Holistic Synchronization

- 하드웨어 가속 기반 P2P 및 계층형 통신 구조
- 분산 워크로드 환경에서도 지연 없이 동기화되는 실행 흐름 보장

커스텀 D2D 프로토콜
Custom Die-to-Die Protocol

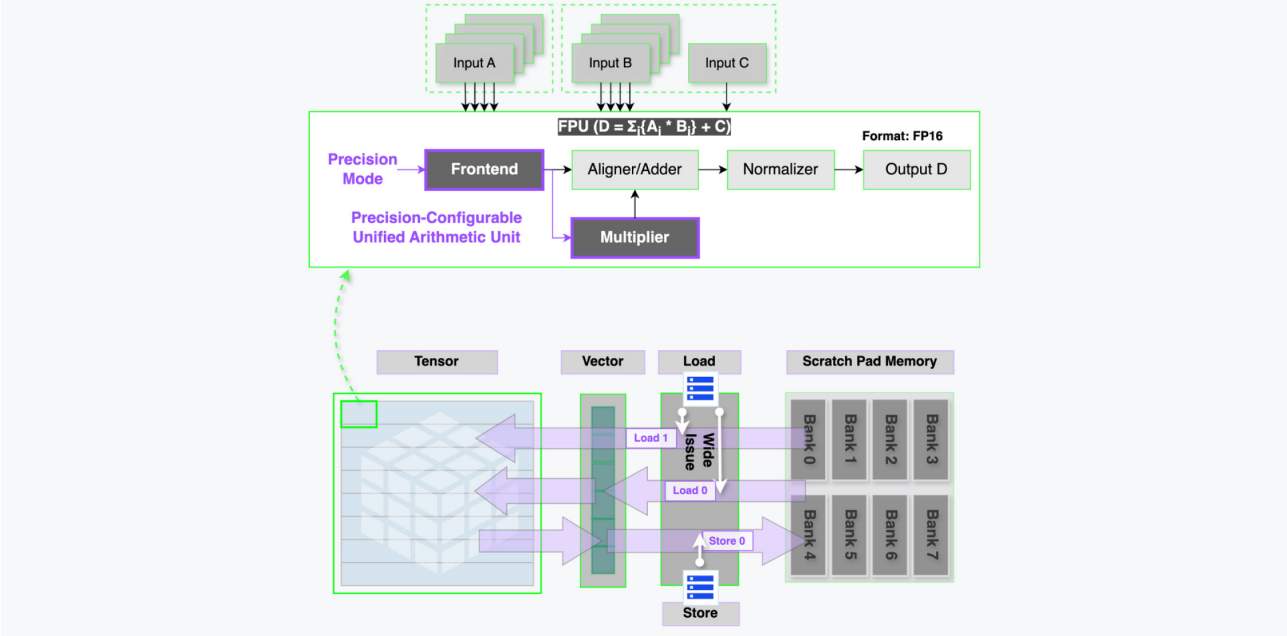
- 채널당 1 TB/s 양방향 대역폭, 칩렛 간 11ns 수준의 초저지연
- 모듈형 확장성을 유지하며 다양한 칩렛 구성 지원

REBEL-Quad의 소프트웨어 스택은 하드웨어와 긴밀하게 결합되어 있으며, **칩렛 간 전용 프로토콜을 기반으로 하이퍼스케일러와 AI 데이터 센터의 초대형 추론 부하를 안정적으로 처리**합니다.

단일 혼합 정밀도 연산 & 칩렛 단위 확장성

기존 NPU는 FP8, FP16, BF16 등 정밀도별로 독립된 연산 블록을 사용해, **면적 비효율과 데이터 흐름 관리의 복잡성**이 발생했습니다. REBEL-Quad는 **연산 단위(operand)별로 정밀도를 설정할 수 있는 통합 산술 엔진**을 도입하여 별도의 기능 블록이 필요 없는 **단일화된 연산 구조**를 구현했습니다.

이를 통해 기존 대비 2.8배 높은 연산 밀도와 하드웨어 수준의 Wide-Issue 실행으로 명령어 종속성 감소를 달성했습니다.



[Figure 2. 통합된 다중/혼합 정밀도 연산 코어]

Wide-Issue 메커니즘은 텐서 및 벡터 코어 간 메모리 대역폭을 균형 있게 배분해, 레지스터와 스크래치 패드 메모리 접근을 동시에 수행합니다. 이는 **FP8 처리량이 중요한 LLM 추론의 프리필 단계**에서 특히 유리합니다. REBEL-Quad는 **4개 칩렛 패키지 내에서 2 PFLOPS(FP8)** 성능을 달성하며, 단일 노드 수준에서 탁월한 **성능 대비 전력 효율**을 확보했습니다.

예측형 DMA & 고대역폭 메모리 접근

LLM의 디코딩 단계에서는 **KV 캐시 메모리 대역폭**이 병목이 되며, 컨텍스트 윈도우가 길어질수록 처리 효율이 급격히 저하됩니다. 이를 해결하기 위해 REBEL-Quad는 **예측형(Predictive) DMA 엔진**을 적용했습니다. 이 엔진은 소프트웨어에서 구성 가능하며, 다음과 같은 특징을 갖습니다:

- **2.7 TB/s 유효 메모리 대역폭**
- 로컬 및 원격 HBM 동시 접근
- 멀티패스 라우팅으로 대역폭 인터리빙 구현

이 DMA 엔진은 REBEL-Quad의 맞춤형 메쉬 인터커넥트와 긴밀하게 통합되어, **기존 대비 3.3배 높은 코어별 대역폭**을 제공합니다. 또한 **태스크별 QoS 지원**을 통해 워크로드에서 지연 변동을 최소화합니다.

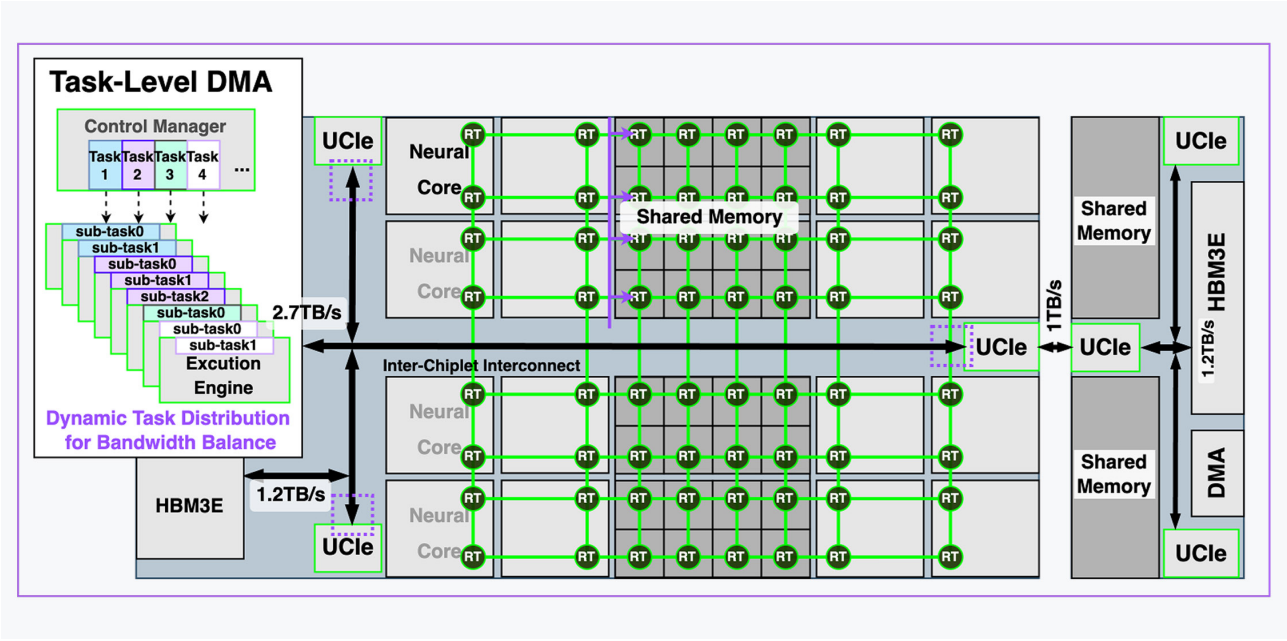
계층형 동기화 및 피어 통신

REBEL-Quad는 복잡한 어텐션 패턴과 장기 의존성을 가진 모델에서도 일관된 성능을 유지하기 위해 **전체 칩 단위 동기화 및 통신 구조**를 제공합니다.

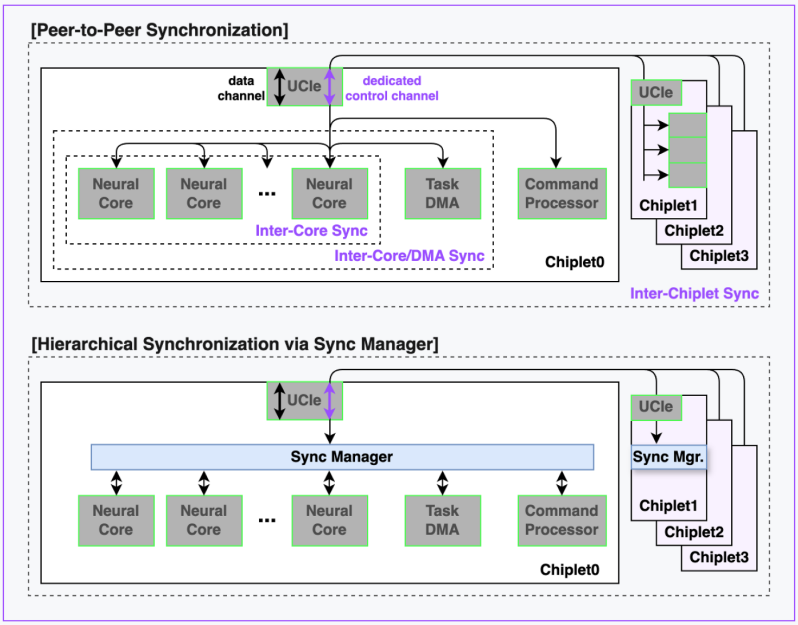
주요 구성은 다음과 같습니다:

- 메쉬 네트워크 전반의 **제어 신호 전용 가상 채널**
- 실행 흐름을 제어하는 **중앙 집중형 동기화 매니저**
- 코어, DMA, 동기화 유닛 간 **하드웨어 가속형 P2P 통신**

이 **계층형 통신 프로토콜**은 칩렛 내부(intra-chiplet)와 칩렛 간(inter-chiplet) 의존성 해결을 모두 지원하며, 프리필과 디코딩이 병행되는 상황에서도 최대 자원 활용률을 유지합니다. 그 결과, 기존 아키텍처의 동기화 병목을 제거하고 소프트웨어 오버헤드를 최소화하여, **높은 연산 밀도, 최대 활용률, 최고 수준의 전력 효율**을 달성합니다.



[Figure 3. 뉴럴 코어와 DMA 엔진을 활용한 풀칩 데이터 전송 구조]



[Figure 4. 칩렛별 중앙 동기화 매니저를 이용한 풀칩 P2P 및 계층적 동기화 방식]

모듈형 확장을 위한 다이-투-다이 프로토콜

REBEL-Quad의 모듈형 확장은 **UCIe-Advanced 기반의 커스텀 다이-투-다이(Die-to-Die) 프로토콜**로 구현됩니다.

- 채널당 **1 TB/s 양방향 대역폭**, 11ns의 칩렛 간 전체 경로 지연
- 칩 간 **Load-Store 메모리 접근** 지원
- 향후 **Scale-Up / Scale-Out 구조 확장**에 대비한 유연성 확보

이 인터커넥트는 다중 칩 시스템을 **가상 단일 시스템(virtually monolithic unit)** 으로 통합하면서, 모듈형 확장성을 유지해 미래 시스템 설계에도 대응합니다. 각 칩렛은 **3개의 UCIe 채널**로 연결되며, **토폴로지 기반 다이 회전(Die Rotation)** 을 통해 수평 메쉬 연속성을 확보합니다.

또한, 고신뢰성 운용을 위한 **스위치 네트워크 및 실시간 디버그 메커니즘**이 탑재되어 대규모 AI 추론 환경에서도 안정적이고 무오류(Zero-Error) 실행을 보장합니다. 향후에는 **I/O 및 메모리 확장 칩렛(Expander Chiplet)** 을 통해 재설계 없이 시스템 구성을 확장할 수 있도록 계획되어 있습니다.

REBEL-Quad는 차세대 LLM 서빙에 요구되는 **성능, 효율성, 확장성**을 모두 충족합니다. 칩렛 기반의 모듈형 구조를 통해 **유연한 업그레이드와 장기 확장성**을 제공하며, **하이퍼스케일러와 엔터프라이즈 AI 시스템을 위한 최적의 기반**으로 자리잡을 것입니다.