Rebel100™

에너지 부담 없는
페타스케일 MoE 가속.

- 고효율·저전력으로 프론티어 LLM을 서빙하도록 설계
- 통합 혼합정밀 코어, 지능형 DMA 스케줄링, UCIe 인터커넥트 기반
- 랙 스케일 성능. 모듈형 확장성. 즉시 배포 가능

Rebel100은 세계 최초의 **UCIe-Advanced 기반 AI 가속기 카드**로, 초대규모 환경에서 효율적이고 확장 가능한 추론을 위해 설계되었습니다. 4개의 동일한 연산 칩렛을 고대역폭 UCIe-Advanced 링크로 단일 SoC처럼 결합한 모듈형 아키텍처로 구현되었습니다. 이를 통해 단일 노드부터 멀티 랙 클러스터까지 매끄럽게 확장이 가능합니다.

가속기 카드 사양

Architecture	4-homogeneous-chiplet SoC based on UCIe-Advanced
Compute (Dense)	1,024 TFLOPS (FP16) 2,048 TFLOPS (FP8)
External Memory	HBM3E 144GB 4.8TB/s
Chiplet (UCIe-A) Interconnection	16Gbps 1TB/s per channel

Host Connection
2x PCIe Gen5 x16
Power Consumption
Up to 600W
Software
Native-support of PyTorch 2.x, vLLM and Triton

SoC 아키텍처 핵심 기술

Unified Compute Engine: One Core for All Precisions

피연산자(operand) 단위 유연성을 갖춘 혼합정밀 코어가 FP8과 FP16을 단일 파이프라인에서 지원합니다. 명령어 전환을 제거해 연산 밀도를 ATOM™ 대비 2.8배 향상시킵니다.

Responsive DMA Scheduling: Memory Access Without Waiting

사전 컴파일된 DMA 엔진이 KV 캐시 접근을 명령어 수준 대역폭·QoS 제어와 함께 조율합니다. 그 결과 코어당 대역폭이 3.3배 향상되어 32K 토큰 디코딩도 매끄럽게 처리합니다.

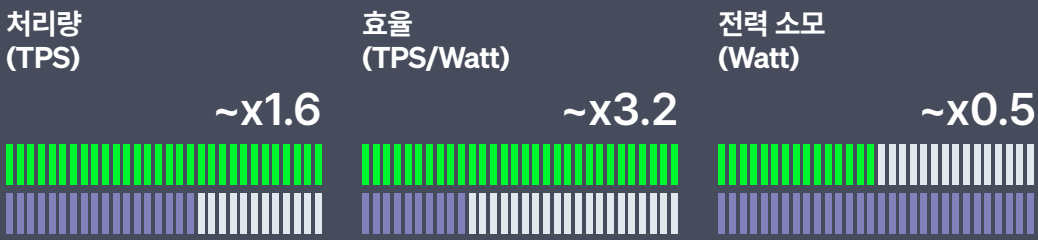
Modular UCIe Interconnect: Multiple Dies as One

4개의 칩렛이 채널당 1TB/s의 UCIe-Advanced 링크로 하나의 시스템으로 통합됩니다. 동일한 인터커넥트가 외부로 확장되어 랙 스케일 분산 아키텍처까지 지원합니다.

Holistic Synchronization: Perfect Orchestration for Peta-Scale Workloads

하드웨어 가속 싱크 패브릭이 칩렛 전반의 실행을 동기화합니다. 희소형(sparse) 모델과 MoE 모델까지 포함해 미세 병렬 처리를 지원하고 전문가 라우팅 과정의 병목을 제거합니다.

Rebel100 vs. H200 (Llama 3.3 - 70B)



벤치마크 조건: Llama 3.3 70B (TP2, FP8), 입력/출력 길이 2048/2048 기준 성능 측정.

Rebel100은 실리콘부터 시스템 규모까지 AI 인프라의 경제성을 새롭게 정의합니다.

Rebel100은 내부 평가에서 대규모 LLM 구동 시 NVIDIA H200급 GPU 대비 더 높은 처리량을 입증했습니다. 이는 컴퓨트와 메모리 병목을 동시에 해소하는 단일 모듈형 시스템 구조 덕분입니다. 현재는 I/O 다이를 포함한 차세대 REBEL 칩렛을 개발 중이며, 조 단위 파라미터 모델과 멀티 노드 엑사스케일 배포까지 확장할 수 있도록 설계되고 있습니다.