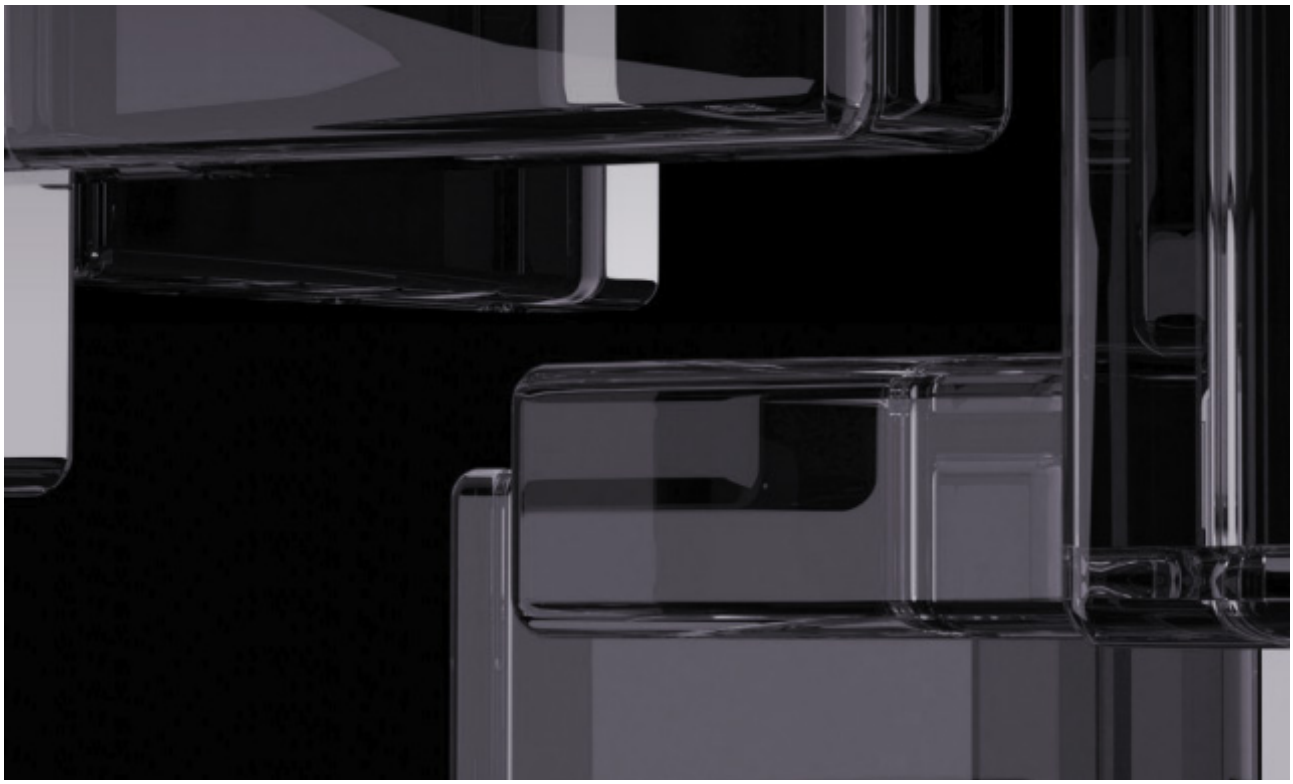
rebellions_

Rebellions Scalable Design

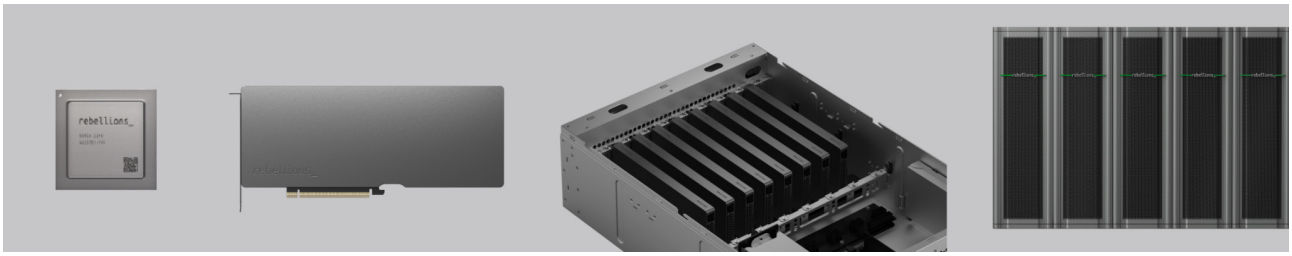
Nov 15, 2024



확장성과 모듈성을 중심으로 한 아키텍처에 뿌리를 두고 있는 리벨리온의 엔지니어링 철학은 Rebellions Scalable Design (RSD)로 구현됩니다. RSD는 현재와 미래의 모든 리벨리온 제품의 핵심 기반으로, 다양한 환경에서 안정적인 확장성과 일관된 성능을 가능하게 하는 견고한 토대를 제공합니다.

RSD는 선형적 확장성(linear scalability)을 실현하도록 설계되었습니다. 시스템 규모가 커질수록 성능 역시 비례적으로 향상되며, 효율의 손실 없이 확장이 가능합니다. 리벨리온은 이러한 강점을 기반으로 소규모부터 하이퍼스케일러(hyperscaler) 수준의 대규모 추론 환경까지, 모든 규모의 인퍼런스 작업에 최적화된 고효율 솔루션을 제공합니다.

대형 언어 모델(LLM)을 포함한 다양한 AI 모델을 완벽히 지원하며, 자체 **소프트웨어 스택**을 통해 성능과 호환성을 극대화합니다. 소규모 **CNN 기반 애플리케이션**부터 고난도 **트랜스포머 기반 워크로드**까지, RSD는 **일관되고 강력한 성능**을 낼 수 있습니다.



핵심 기술

AI 모델의 규모가 커질수록 다양한 응용 가능성이 확장되지만, 그 이면에서는 대규모 연산을 여러 AI 프로세서 간 원활하게 분산·동기화하는 고난도의 기술력이 요구됩니다. RSD는 이러한 과제를 해결하기 위해 **텐서 병렬 처리(tensor parallelism)** 구조를 적용했습니다. 이를 통해 대형 모델의 연산을 여러 프로세서에 효율적으로 분배함으로써, 모델 전체가 안정적으로 실행됩니다.

RBLN 컴파일러는 모델을 미세 수준까지 최적화하는 핵심 역할을 수행합니다. 여기에 PCIe Gen5 통합을 통해 고속 입·출력과 카드 간 직접 통신을 구현하여 시스템 전반의 지연(latency)을 최소화하고 데이터 처리량(throughput)을 극대화했습니다. 이처럼 **병렬 처리, 컴파일러 최적화, 고속 인터커넥트 기술**을 결합한 RSD는 가장 까다로운 AI 연산도 정밀하고 빠르게 처리할 수 있는 인프라를 제공합니다.

텐서 병렬 처리

LLM을 AI 프로세서에서 추론(inference)할 때는 여러 기술적 과제가 존재합니다. 프리필(prefill) 단계는 연산 집약적이며, 디코딩(decoding) 단계는 메모리 자원을 크게 요구합니다. 이때 텐서 병렬 처리는 연산 부하를 여러 디바이스에 분산시켜, 각 디바이스의 메모리 점유율과 연산 부하를 효과적으로 줄이는 해결책을 제공합니다. 텐서 병렬 처리를 올바르게 구현하면, KV 캐싱(KV caching)이나 LLM의 대규모

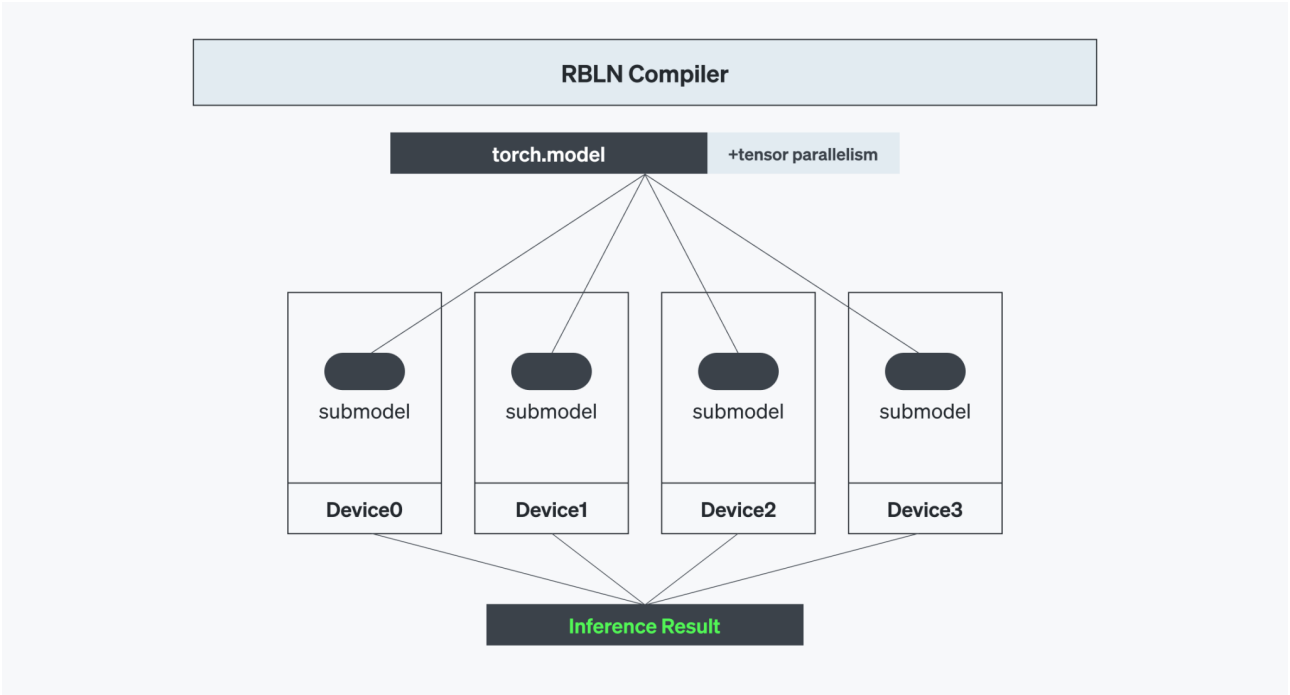
가중치로 인해 발생하는 메모리 대역폭 제약을 완화할 수 있습니다. 이를 통해 하드웨어는 고처리량과 저지연을 유지하면서 복잡한 LLM 추론 작업을 효율적으로 수행할 수 있습니다.

RBLN 컴파일러는 이러한 텐서 병렬 처리를 최대 성능과 활용률(utilization)로 관리하도록 최적화되어 있습니다. 컴파일 시점에서 컴파일러는 모델을 여러 디바이스에 걸쳐 세밀하게 텐서 단위로 분할하며, 각 칩이 전체 연산의 일부만 담당하도록 구성합니다. 이때 생성되는 커맨드 스트림(Command Stream)은 커맨드 프로세서(Command Processor)가 실행할 명령 세트로, 추론 중 칩 간 통신에 필요한 인터디바이스 데이터 이동 정보 또한 포함하고 있습니다.

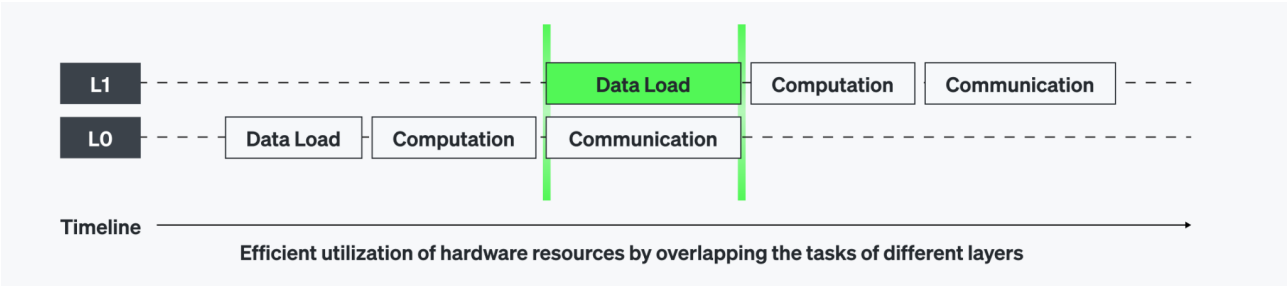
컴파일러 수준 최적화

RBLN 컴파일러는 AI 워크로드의 확장 복잡성을 효율적으로 관리하도록 설계된 고성능 AI 추론을 위한 핵심 도구입니다. 텐서 병렬 처리를 효과적으로 지원하여 모델을 여러 디바이스에 매끄럽게 분산하고, 최적의 자원 활용과 고속 실행 성능을 보장합니다. 또한 멀티 디바이스 간 통신 최적화, 자동 연산 분할(automatic splitting), 레이어 파이프라이닝(layer pipelining) 등 기능을 통해 확장성을 한층 강화하였습니다.

1. **자동 멀티 디바이스 분할 (Automatic Multi-Device Splitting)**
   RBLN 컴파일러는 연산의 분할과 재결합 과정을 모두 자동으로 수행합니다. 사용자 개입 없이 멀티 디바이스 간 연산이 자동 처리되므로, 사용자 입장에서는 복잡한 텐서 병렬 처리가 단순하고 직관적인 프로세스로 전환됩니다.



2. **디바이스 간 통신 최적화 (Optimization of Inter-Device Communication)**
   RBLN 컴파일러는 LLM 실행 중 발생하는 디바이스 간 통신(inter-device communication)을 고도화하여 브로드캐스트(broadcast), 리듀스(reduce), 부분 합(partial sum) 등과 같은 집합 통신 패턴(collective communication pattern)을 효율적으로 처리합니다.
3. **디바이스 간 통신 최적화 (Optimization of Inter-Device Communication)**
   RBLN 컴파일러는 각 디바이스 내부에서 레이어 파이프라이닝(layer pipelining)을 적용하여 디바이스 간 통신이 중단 없이(seamless) 진행되도록 합니다. 이 기법을 통해 모든 연산이 병렬로 처리되며, 유휴 시간(idle time)을 최소화하여 하드웨어 활용률을 극대화하고 통신 오버헤드를 줄일 수 있습니다.



PCIe Gen5

리벨리온의 **RSD**는 **PCIe Gen5 ×16 인터페이스**를 채택하여, 호스트 연결과 카드 간 직접 통신 모두에서 양방향 전이중(full-duplex) 64GB/s 대역폭을 제공합니다. 카드 간 효율적인 통신은 텐서 병렬 처리에서 특히 중요하며, 모든 뉴럴 엔진(Neural Engine) 간 통신에서 고처리량과 저지연을 보장함으로써 확장 가능한 속도로 탁월한 추론 성능을 구현합니다. 리벨리온은 PCIe의 성능을 최대한 끌어내기 위해 펌웨어를 전용으로 최적화하고, PCIe 스위치와 CPU 역시 상호운용성을 극대화하도록 설계하여 시스템 전반의 효율성을 한층 향상시켰습니다.

시스템 솔루션

경량 워크로드(workstation)부터 SLM, LLM과 같은 고부하 AI 작업에 이르기까지, RSD는 탁월한 TPS/Watt 효율을 제공하는 강력하고 경제적인 시스템 솔루션입니다. 이 효율성은 RSD의 진정한 선형 확장성과 랙 수준의 성능 최적화 덕분에 구현되며, vLLM과 LiteLLM의 통합을 통해 완성됩니다. 이 기술 조합을 통해 RSD는 연산 수요가 증가해도 안정적으로 확장되며, 에너지·비용당 처리량을 극대화하여 모든 규모의 AI 인프라에 이상적인 솔루션을 제공합니다.

선형 확장성

선형 확장성은 연산 자원을 추가할수록 성능이 비례적으로 증가하여, 하드웨어 투입이 처리량 향상으로 직접 연결되도록 보장합니다. 이러한 특성은 대규모 AI 모델과 데이터 집약적 애플리케이션의 급격히 증

가하는 연산 요구를 효율성과 속도 저하 없이 대응하기 위해 필수적입니다.

RSD는 카드, 서버, 랙 시스템 등 다양한 배포 환경에서 이 선형 확장성을 완벽히 구현합니다. 노드 간 데이터 동기화(data synchronization)의 최적화, **노드 간 통신 오버헤드**의 최소화, **메모리 대역폭 관리 효율화**를 통한 병목 방지를 통해 디바이스 수가 증가할 때도 일관된 낮은 지연시간을 유지하면서 처리량을 비례적으로 확장할 수 있습니다.

또한 저지연 성능을 유지하기 위해 부하 분산(load balancing)과 **동적 워크로드 분배(dynamic workload distribution)** 기술을 활용하여 모든 디바이스가 지연 없이 최대 효율로 동작하도록 합니다. RSD는 이러한 복잡성을 첨단 하드웨어 아키텍처와 지능형 소프트웨어 프레임워크를 통해 해결하며, 연산 수요가 증가해도 확장 가능한 고속 처리 성능을 유지할 수 있습니다.

랙 수준 최적화 (Rack-level Optimizations)

랙 단위 AI 추론 환경에서는 서버 간 원활한 통신과 워크로드 분배를 위한 효율적인 라우팅 프로토콜이 핵심입니다.

RSD는 vLLM과 통합된 라우터 서버(router server)를 도입하여 랙 수준에서 최적의 성능과 확장성을 제공합니다. 이 라우터 서버는 여러 vLLM 인스턴스를 하나의 통합 시스템으로 묶는 프레임워크 역할을 수행하며, 서버 간 워크로드를 지능적으로 분배하여 랙 전체의 효율을 극대화합니다. 이를 통해 모든 서버가 균형 잡힌 부하 상태에서 최대 처리량을 유지하고, 과부하를 방지하며 안정적인 저지연 성능을 제공합니다.

사용자 입장에서는 LLM 모델에 API 엔드포인트를 통해 쉽게 접근할 수 있으며, 우수한 확장성과 사용 편의성을 동시에 누릴 수 있습니다. 결과적으로, vLLM과 라우터 서버의 결합은 개별 서버를 정교하게 조율된 고성능 AI 시스템으로 전환시키는 강력한 시너지를 만듭니다.

Llama3-8B

Llama3-8B 벤치마크 결과는 RSD의 전력 효율(energy efficiency)이 경쟁 제품 대비 얼마나 우수한지를 명확히 보여줍니다. 단일 **ATOM™ 서버**가 2,500W 전력에서 **1800 TPS**를 기록하며, 224장의 ATOM™ 카드로 구성된 랙 시스템에서는 **25,200 TPS**까지 선형적으로 확장됩니다. 규모가 커질수록 그 효율성은 더욱 두드러지며, **TPS/Watt** 및 TPS/$ 기준으로 **에너지 효율과 비용 효율 모두 8배 이상의 향상**을 달성했습니다.

Server			Rack (assumes rack power = 35 kW)	
Target model: Llama3-8B, FP16, Input/output = 2K/2K	rebellions_ ATOM™	Peer A	rebellions_ ATOM™	Peer A
TPS (max TPS at best batch)	1.8 k	constraints	At rack power of 35 kW	
			224 cards	72 cards**
TDP (watt)	2.5 k	a single node cannot run Llama3-8B	25.2 k	3 k**
			Rack power fixed to 35 kW	
TPS/Watt	0.72		energy 0.72 (8X ↑ vs. Peer A)	0.09**
TPS/$	0.012		cost 0.008 (8X ↑ vs. Peer A)	0.001**

* Highly likely to be subjected to semiconductor export regulations; strongly advised to check regulatory risks before proceeding.
** Estimated based on the public information researched.

결론

RSD는 복잡하고 자원 집약적인 AI 워크로드의 요구를 완벽히 충족하는 리벨리온의 첨단 AI 인프라 비전을 구현합니다. 모듈형·확장형 아키텍처를 기반으로 한 RSD는 단일 워크스테이션부터 대규모 랙 시스템까지 일관된 선형 확장성을 제공합니다.여기에 고급 텐서 병렬 처리, PCIe Gen5 통합, 그리고 RBLN 컴파일러의 정교한 최적화 기술이 더해져, RSD는 어떤 규모에서도 효율적이고 고성능의 AI 추론 환경을 보장합니다.