rebellions_

ATOM™:

GenAI를 위한 최적의 솔루션

Jul 11, 2024

서론

생성형 AI(Generative AI, GenAI)가 다양한 산업에 변화를 가져오면서, 막대한 연산 처리에 특화된 하드웨어 개발은 선택이 아닌 필수가 되었습니다. 이에 따라 AI 워크로드에 특화된 AI 가속기 또는 AI 칩이 핵심 기술로 부상하고 있지만, 효과적인 AI 칩을 설계하는 데에는 많은 도전이 따릅니다.

빠른 처리 속도와 처리량(throughput)은 AI 애플리케이션의 성능과 직접적으로 연관됩니다. 대규모 연산을 처리하는 대표적인 방법 중 하나는 배치 처리(batch processing)로, 다수의 태스크를 그룹화하여 연속적으로 실행합니다. 그러나 이 방식에서는 처리량이 높아지는 대가로 처리 시간이 늘어나게 됩니다.

전체 시스템 수준의 성능을 끌어올리기 위해서는 메모리와 연산 태스크 간의 균형이 매우 중요하나 종종 간과되는 핵심 요소입니다. 그 중요성에 비해 종종 간과되는 유연성(flexibility)도 핵심 요소로, 시스템 수준에서 메모리와 연산 태스크 간의 균형을 말합니다. 예를 들어, 텍스트 기반 대형 언어 모델(LLM) 추론은 방대한 파라미터 처리를 위해 빈번한 메모리 액세스가 요구되는 메모리 집약적인 태스크입니다. 반면, 텍스트-비디오 애플리케이션은 실시간 그래픽 처리와 데이터 처리를 수반하기 때문에 연산 집약적인 태스크로 볼 수 있습니다. 결론적으로, 최적의 AI 칩은 다양한 애플리케이션 지원을 위해 지연 시간, 처리량 및 유연성 간의 균형점을 찾아야 합니다.

연산 효율 극대화

리벨리온은 ATOM™을 설계하는 데 있어 높은 연산 효율을 최우선 과제로 삼으면서 최적의 균형점을 찾고자 했습니다. 다양한 기능을 수행할 수 있도록 재구성이 용이한 CGRA(Coarse-Grained Reconfigurable Array) 아키텍처를 채택하여 유연성을 극대화했습니다. 또한, 유휴 자원을 최소화하고 태스크를 지속적으로 처리하여 효율성을 높이고 처리 시간을 줄였습니다.

유연한 아키텍처에 더해, ATOM™의 동기화 메커니즘은 효율적인 병렬 처리를 지원하기 위해 필요한 자원을 정확히 활성화합니다. 이를 통해 처리 준비에 소요되는 시간과 노력이 줄어들어 레이턴시가 감소합니다. 또한, 다층 메모리 계층 구조는 데이터 의존성을 줄이면서 대역폭을 크게 증가시키고, 실시간 동기화는 제어 의존성을 줄입니다. ATOM™의 모든 요소들은 자원 활용률을 최적화하여 성능과 효율성을 크게 향상시킵니다.

ATOM™: AI 추론을 위한 시스템 온 칩(SoC)

리벨리온의 ATOM™은 AI 추론을 위해 설계된 AI 가속기로, 삼성의 첨단 5nm 공정을 기반으로 제작되었습니다. ATOM™은 FP16 연산에서 32 TFLOPS, INT8 연산에서 128 TOPS를 제공하며, 8개의 뉴럴 엔

진과 64 MB 온칩 SRAM으로 성능을 강화합니다. 세밀하게 설계된 메모리 아키텍처는 성능과 효율성을 극대화합니다.

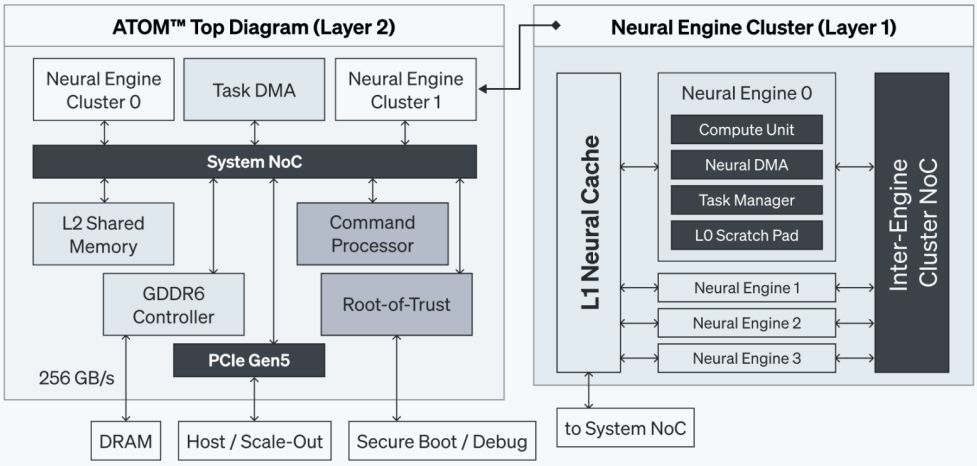
RBLN-CA12

ATOM™은 RBLN-CA12라는 싱글 슬롯, FHFL(Full Height, Full Length) PCIe Gen5 카드로 제공됩니다. 최대 전력(Thermal Design Power, TDP)은 60~130W이며, 256 GB/s 대역폭의 GDDR6 메모리와 PCIe Gen5 ×16 인터페이스를 통해 호스트와 카드, 그리고 카드 간의 통신을 지원합니다. 또한, 멀티 인스턴스(Multi-Instance) 기능을 통해 ATOM™을 16개의 독립적인 하드웨어 분리 인스턴스로 분할하여 강력한 멀티태스킹과 자원 할당을 동적으로 관리할 수 있습니다.

RBLN-CA12	
AI Accelerator	ATOM™
FP16	32 TFLOPS
INT8	128 TOPS
On-chip SRAM	64 MB
External Memory	GDDR6, 256 GB/s, 16 GB
Multi-Instance	Hardware isolation up to 16 independent tasks
Thermal Solution	Passive
Mechanical Form Factor	Full Height, Full Length (FHFL) 266.5 × 111 × 19 mm
Thermal Design Power	60–130 W
Host and Card-to-Card Interface	PCIe Gen5 x16, 64 GB/s
Connectors	One CPU 8-pin power connector (2×4)
Weight	Total: 615 g

[Table 1. RBLN-CA12 사양]

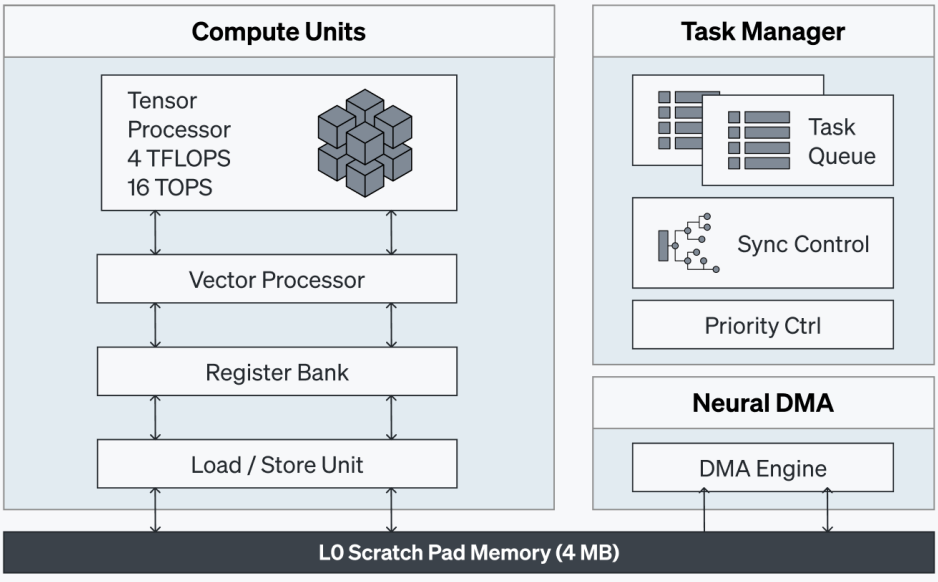
ATOM™ SoC



ATOM™은 여러 핵심 컴포넌트를 하나의 집적회로에 통합한 멀티코어 시스템 온 칩(System-on-Chip)입니다. 그림 1에서 볼 수 있듯, 뉴럴 엔진, 커맨드 프로세서, 온칩 메모리(SRAM), 그리고 GDDR6 메모리를 하나의 칩에 집적했습니다. 다양한 요소를 통합해 칩의 집적도를 높임으로써 공간 활용도와 전력 효율을 최적화합니다.

이 구성만으로도 컴포넌트 간 통신을 간소화하고 레이턴시를 크게 줄일 수 있지만, 추가적으로 고대역폭을 제공하는 네트워크온칩(Network-on-Chip, NoC)을 구현했습니다. 또한, 여러 계층 간 동기화를 지원하도록 설계되어 있습니다.

Neural Engine



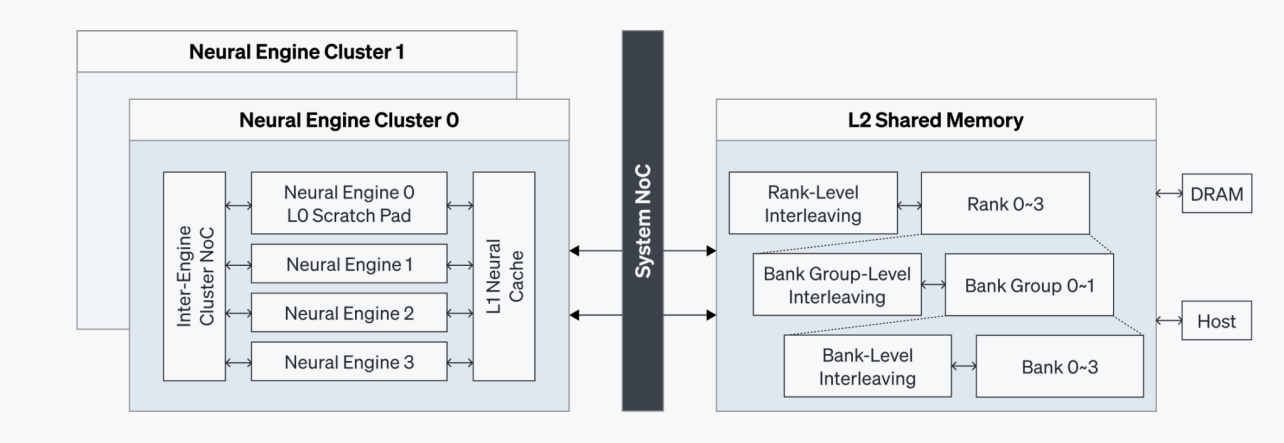
[Figure 2. ATOM™ 뉴럴 엔진]

ATOM™의 뉴럴 엔진은 실제 연산이 이루어지는 핵심 컴포넌트입니다. 뉴럴 엔진 내 연산 유닛들은 이질적인 SIMD(Single Instruction, Multiple Data) 및 MIMD(Multiple Instruction, Multiple Data) 요소를 결합하여 병렬 처리 성능을 극대화하고 명령어 수준에서의 효율적인 의존성 관리를 제공합니다.

4 MB **스크래치 패드(Scratch Pad)** 메모리가 포함된 연산 유닛은 최대 8 TB/s 속도로 SRAM의 중간 데이터에 액세스할 수 있어, 외부 메모리에 대한 의존성을 최소화함으로써 대역폭 한계를 완화하고 처리 속도를 높입니다. 각 뉴럴 엔진에 내장된 태스크 매니저(Task Manager)는 로컬 하드웨어 수준에서 동기화를 가속화하며, 커맨드 프로세서(Command Processor)와 협력하여 최대 활용률을 보장합니다.

ATOM™은 이처럼 뉴럴 엔진의 연산 유닛, 스크래치 패드 메모리, 태스크 매니저를 활용해 높은 활용률과 낮은 레이턴시를 달성합니다.

계층적 메모리 서브시스템

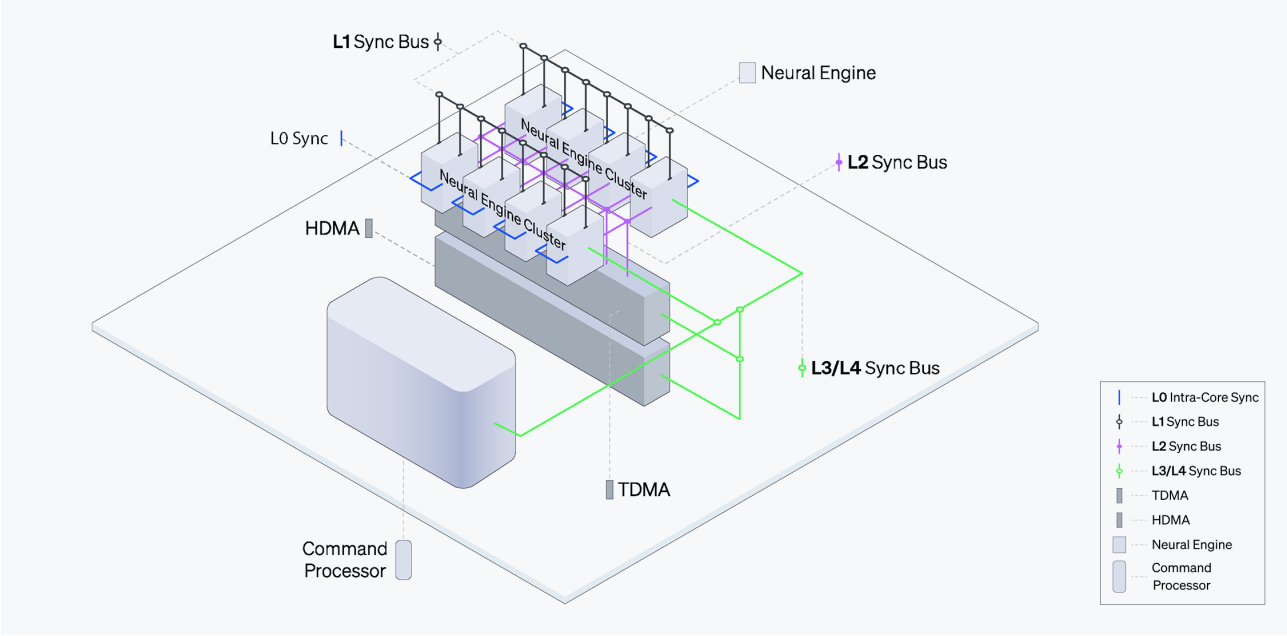


[Figure 3. ATOM™ 계층적 메모리 서브시스템]

ATOM™의 계층적 메모리 아키텍처는 최고 수준의 성능 효율성을 보장하도록 설계되어, 뉴럴 엔진에 충분한 대역폭을 제공하는 동시에 레이턴시를 최소화합니다.

- **GDDR6 메모리**: 16GB의 GDDR6 메모리를 탑재하여 높은 처리량을 유지하면서도 전력 소비를 줄입니다.
- **스크래치 패드(L0)**: 각 뉴럴 엔진에 내장된 4 MB의 스크래치 패드는 즉각적인 로컬 데이터 액세스를 제공합니다.
- **L1 뉴럴 캐시**: 뉴럴 엔진 근처에 위치하여 데이터 접근 속도를 높입니다.
- **L2 공유 메모리**: 64MB SRAM으로, 다층의 인터리빙 기술을 적용하여 병렬성을 지원하고 대역폭을 최적화하며 레이턴시를 축소합니다.

다층 동기화와 병렬 처리



[Figure 4. ATOM™ 동기화 구성]

ATOM™의 동기화 메커니즘은 효과적인 병렬 처리와 칩의 성능 확장을 가능하게 합니다. 동기화는 명령어 및 태스크 수준에서 이루어지며, 커맨드 프로세서와 태스크 매니저, 그리고 전용 로컬 버스를 통한 안정적인 대역폭으로 태스크 흐름을 원활하게 유지합니다.

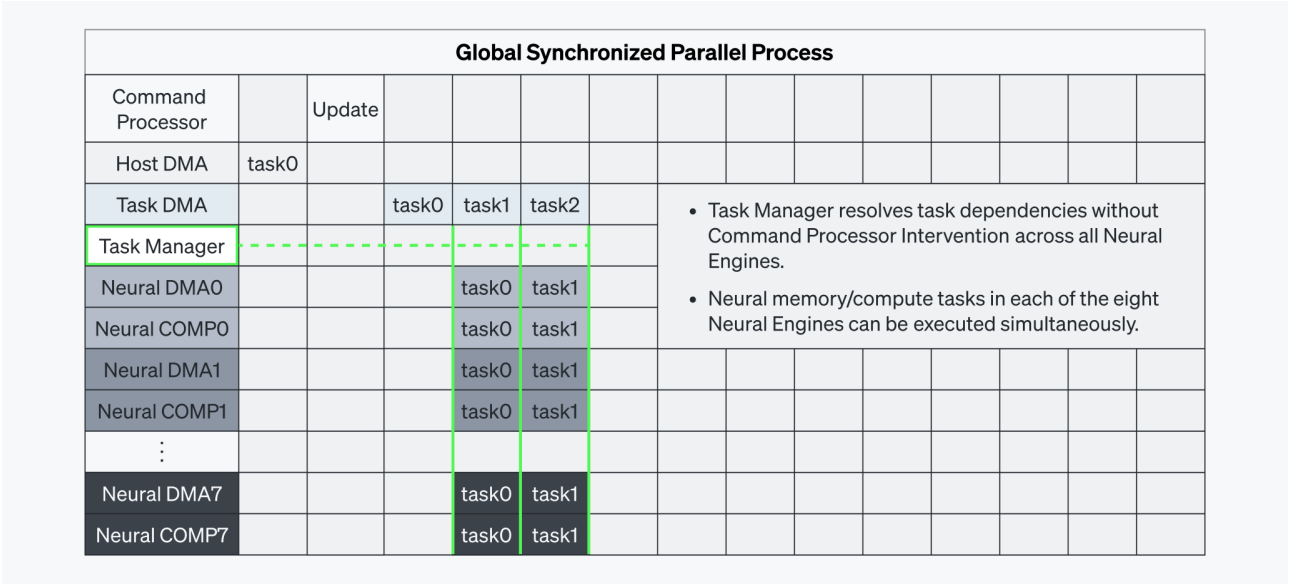
뉴럴 엔진은 L1 동기화 버스(L1 Sync Bus)를 통해 태스크 매니저와 통신하며, 이 태스크 매니저는 L2 동기화 버스(L2 Sync Bus)를 통해 태스크 수준 DMA와도 연결됩니다. 이러한 구성을 통해 시스템은 전체적으로 의존성을 확인하고, 다양한 코어를 동기화하여 고밀도 연산 태스크를 수행할 수 있습니다.

Basic Sequential Process

Command Processor		Update		Update		Update		Update		Update		Update	
Host DMA	task0												
Task DMA			task0						• All dependencies are controlled by the Command Processor				
Neural DMA0					task0								
Neural COMP0							task0						
Neural DMA1													
Neural COMP1											task0		
⋮									task0				
Neural DMA7													
Neural COMP7													task0

[Figure 5-1. 태스크 매니저가 없는 순차적 명령 실행]

커맨드 프로세서가 명령 실행을 단독으로 관리하면, 태스크가 순차적으로 처리되어 레이턴시가 늘어납니다. 태스크 의존성이 해소될 때까지 커맨드 프로세서의 처리를 기다려야 하기 때문입니다(그림 5-1 참고). 이러한 방식은 통신 오버헤드를 발생시킵니다.



[Figure 5-2. 태스크 매니저를 포함한 병렬 실행]

명령 실행의 최적화를 위해, 리벨리온은 태스크 매니저를 도입하여 하드웨어 수준에서 로컬 의존성을 자동으로 해소하도록 했습니다. 그림 5-2에서 볼 수 있듯, 커맨드 프로세서가 의존성을 해결하지 않고도, 각 뉴럴 엔진 내 DMA/COMP 태스크가 병렬적으로 실행될 수 있습니다. 뉴럴 엔진의 태스크 매니저가 의존성을 해결함으로써 작업이 동시적으로 처리됩니다. 이러한 과정은 그림 4와 같이 전용 L1/L2 데이터 경로를 통해 이루어집니다. 결과적으로, 모든 뉴럴 엔진에서 태스크가 효율적으로 조정되어 병렬 실행이 원활하게 이루어지며 레이턴시가 최소화됩니다.

벤치마크 결과

ATOM™과 NVIDIA A100의 성능을 비교하기 위해 대표적인 AI 모델인 T5-3B(자연어 처리)와 SDXL-Turbo(텍스트-이미지 생성)를 활용해 테스트를 진행했습니다. 이러한 비교를 통해 ATOM™이 최신 AI 워크로드를 얼마나 효과적으로 처리할 수 있는지 검증했습니다.

언어 모델 벤치마크: T5-3B

구글이 개발한 T5(Text-to-Text Transfer Transformer)는 Transformer 아키텍처를 활용한 혁신적인 대형 언어 모델(LLM)입니다. T5 모델은 6천만에서 110억 개의 파라미터 규모를 갖추고 있습니다.

이번 비교에서는 언어 번역, 텍스트 요약, 질문 응답, 텍스트 생성과 같은 워크로드에 적합한 30억 파라미터 모델을 배치 크기 1로 실행했습니다.

- **성능:** 초당 생성된 토큰 수로 측정.
- **전력 소비:** 와트(W) 단위로 측정.
- **전력 효율성:** 성능 대비 소비 전력으로 계산.

테스트 결과, ATOM™은 A100 대비 최대 44% 높은 전력 효율성을 달성함으로써 복잡한 언어 처리 워크로드에서의 강력한 성능과 효율성을 입증했습니다.

	Input	Output	Performance (Token/s)	Average Power (W)	Average Power Efficiency (Token/J)
ATOM™	349	512	45.0	56.1	0.80
A100	349	512	44.3	177.5	0.25

* 두 테스트는 모두 FP16 정밀도로 진행되었습니다.
* ATOM™의 결과는 예측치에 기반합니다. A100의 결과는 Hugging Face transformers 라이브러리에 기반합니다.

텍스트-이미지 모델 벤치마크: SDXL-Turbo

Stability AI가 개발한 SDXL-Turbo는 고해상도 이미지 생성을 전문으로 하며, 기존 Stable Diffusion 모델 대비 추론 속도가 크게 향상되었습니다.

ATOM™은 A100보다 훨씬 적은 전력을 소비하면서도 높은 성능을 보였습니다. 즉, 더 적은 자원으로도 우수한 결과를 달성하며, 운영 비용을 크게 절감하고 서비스 배포의 지속 가능성 또한 향상시킬 수 있습니다.

	Performance (img/s)	Power (W)	Power Efficiency (Performance/Power)
ATOM™	3.74	60.3	0.062
A100	7.36	192.7	0.038

* 이미지 크기 512×512, Diffusion step: 1
* ATOM™의 결과는 예측치입니다. A100의 결과는 Hugging Face diffusers 라이브러리에 기반합니다.

결론

업계를 불문하고 AI 의존성이 높아지는 가운데, 지속 가능한 확장성을 갖춘 최적의 AI 칩을 찾는 것은 많은 기업들이 당면한 과제입니다. ATOM™은 유연성, 전력 효율성, 높은 성능 간의 균형을 이루면서도 빠른 처리 시간을 확보할 수 있도록 설계되었습니다. 혁신적인 뉴럴 엔진과 다층 메모리 아키텍처, 강력한 동기화 기능을 통해 지연 시간과 전력 효율성을 최적화하였으며, 높은 연산 활용률을 구현합니다. ATOM™은 AI 서비스의 운영 비용을 획기적으로 줄이고 수익성을 향상시킬 수 있는, 지속 가능한 AI 서비스를 위한 최적의 AI 칩입니다.