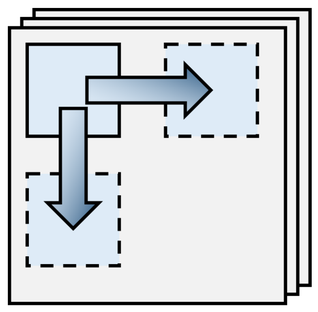
IDEC Chip Design Contest

Weight Only Quantization Model Acceleration with In-Memory Computing Unit

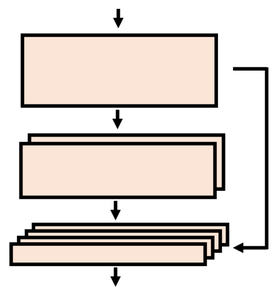
Jaeyong Jang, and Jae-Joon Kim
Seoul National University, Korea

대형 언어 모델(LLM)은 다양한 분야에서 점점 더 널리 활용되고 있다. 그러나 입력값에 활성화 함수가 빈번히 적용된다는 특성상, 입력 자체를 양자화하기는 어렵다. 반면 가중치는 상대적으로 양자화가 용이하며, LLM에서 가중치의 크기가 입력에 비해 훨씬 큰 만큼 가중치만 양자화하는 방식이 일반화되어 왔다. 이전 연구들에서는 양자화가 메모리 전송 효율 측면에서는 이점을 제공하지만, 계산 성능 향상으로까지 이어지지는 않는다는 사실이 밝혀졌다. 본 논문에서는 가중치만 양자화할 때 자주 발생하는 부동소수점 값과 정수 값 간의 혼합 곱 연산(mixed multiplication)의 성능을 높이기 위해 고안된 가속기 아키텍처를 제안한다.

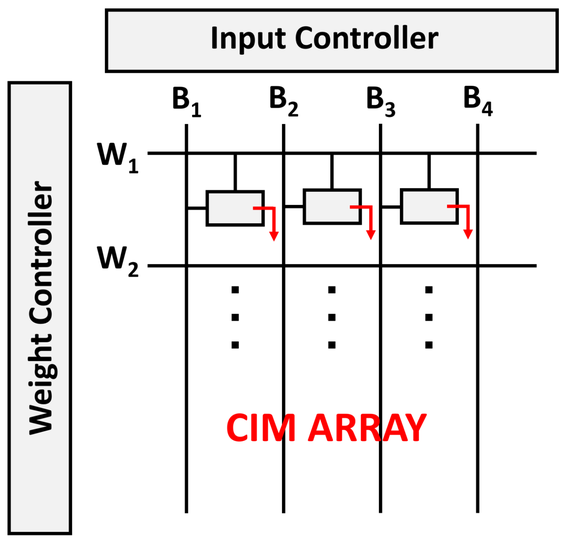
Background



DNN model



Neural Network



Compute-In-Memory (CIM) array with energy efficient CIM macro controller logic

Proposed Hardware Architecture

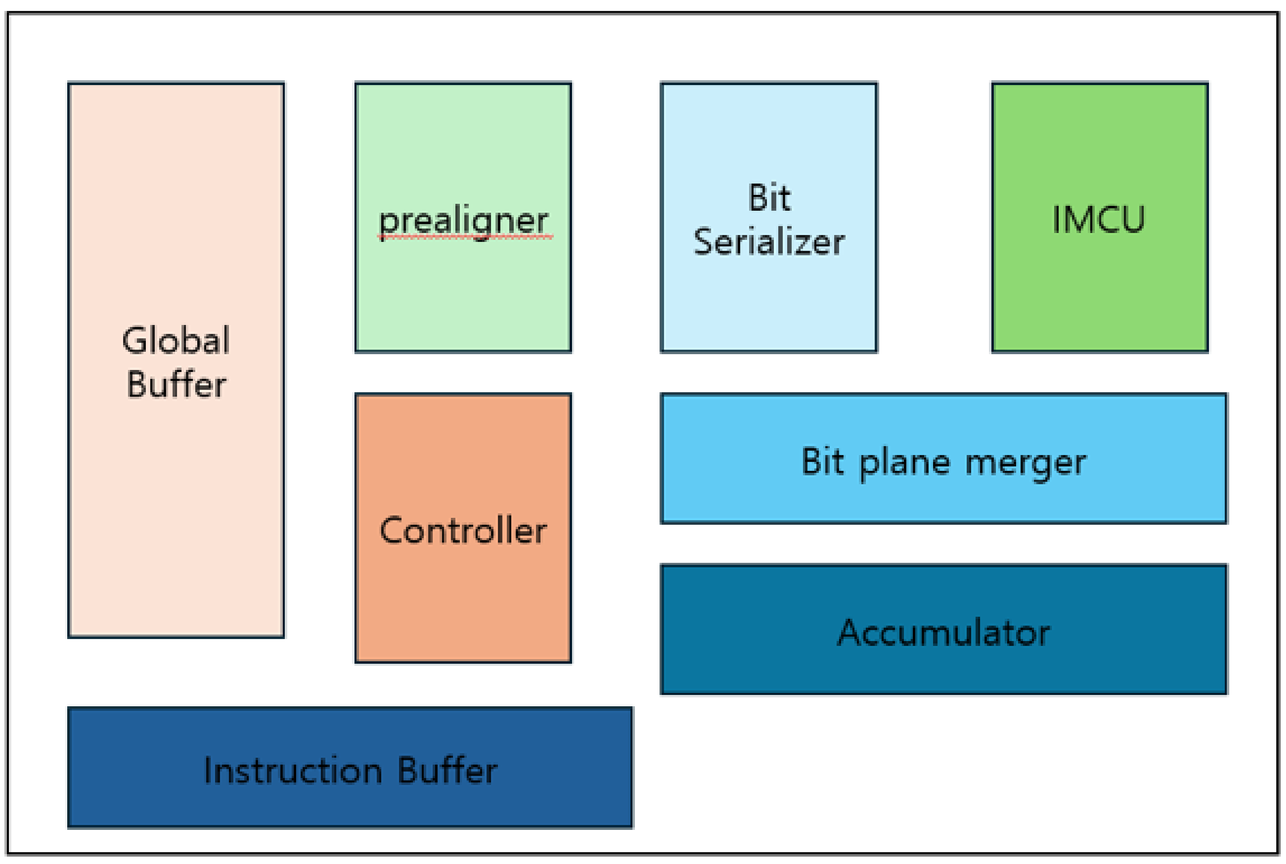


Fig.1 FICIM overall architecture

Key Component

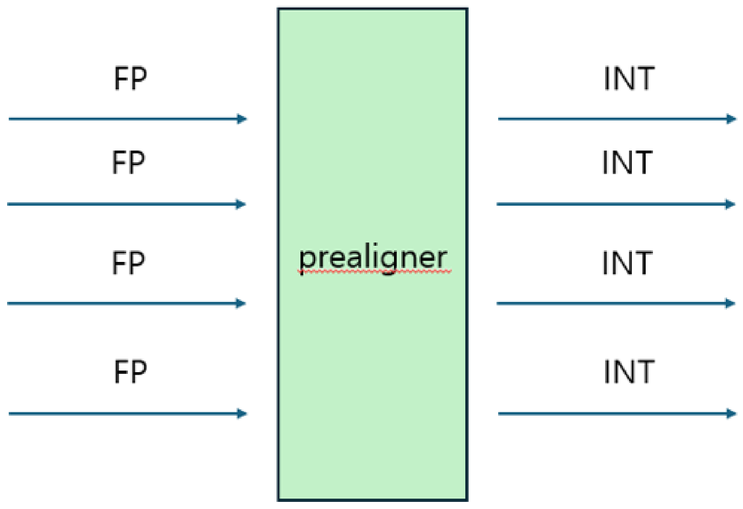
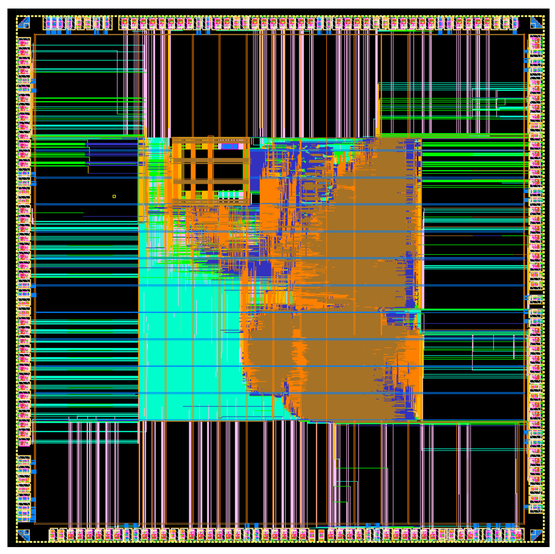
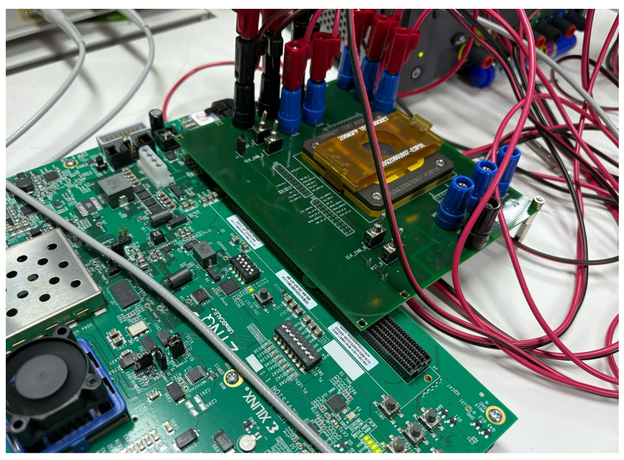


Fig 2. Prealigner

Results





Chip test setting

* 본 연구는 IDEC에서 지원 받았음.