2024년 4월 8일

TPU

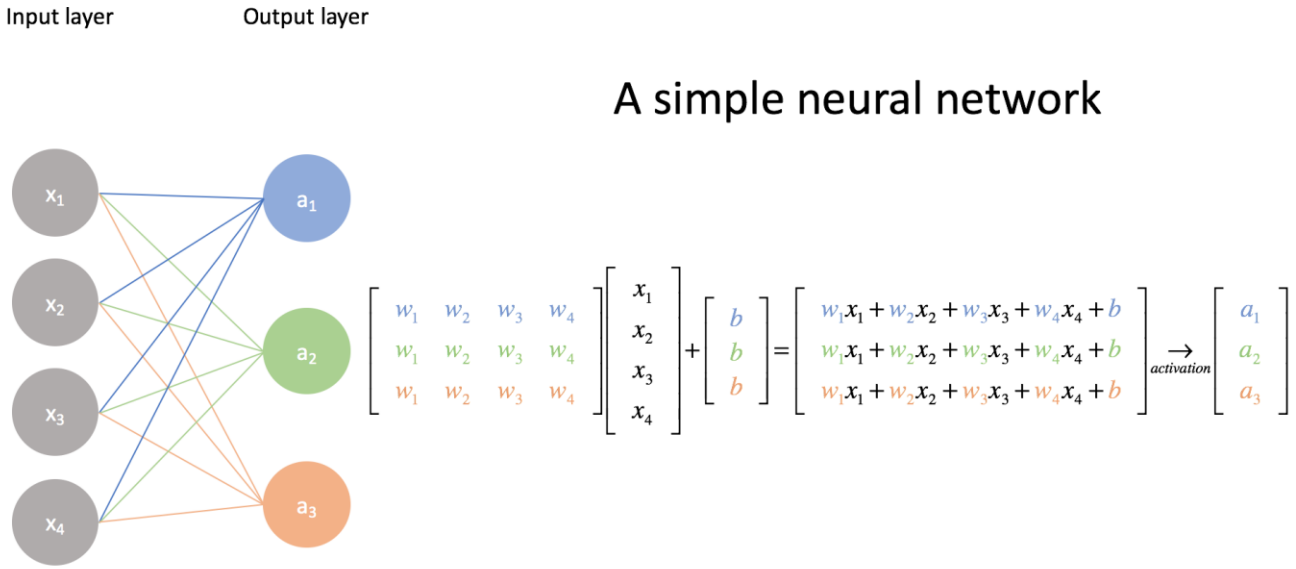
홍익대학교 컴퓨터공학과
C135283 이수현

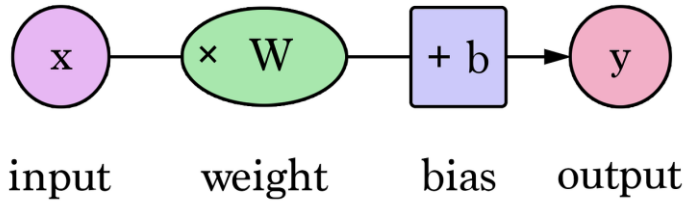
Contents

- Background
  - CPU and GPU
- TPU 특징
  - Systolic Array
  - Bfloat16
  - HBM

1. Background

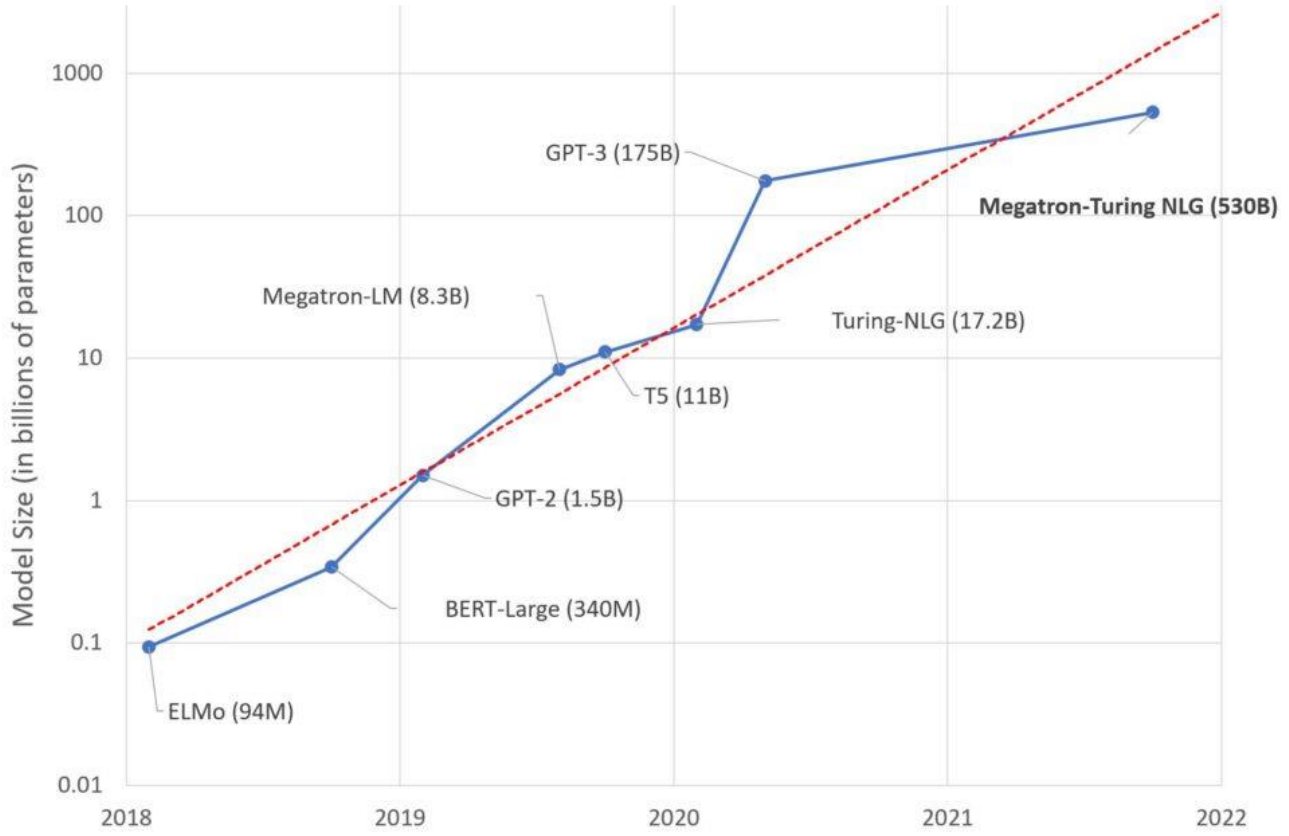
- 인공 신경망의 기본 연산: 행렬곱
  Multiply-Accumulation





1. Background

- 인공 신경망의 기본 연산 = 행렬곱
- LLMs parameters increased by 10x per year over 5 years
  –〉 증가한 연산량에 따른 더 좋은 연산 장치 필요

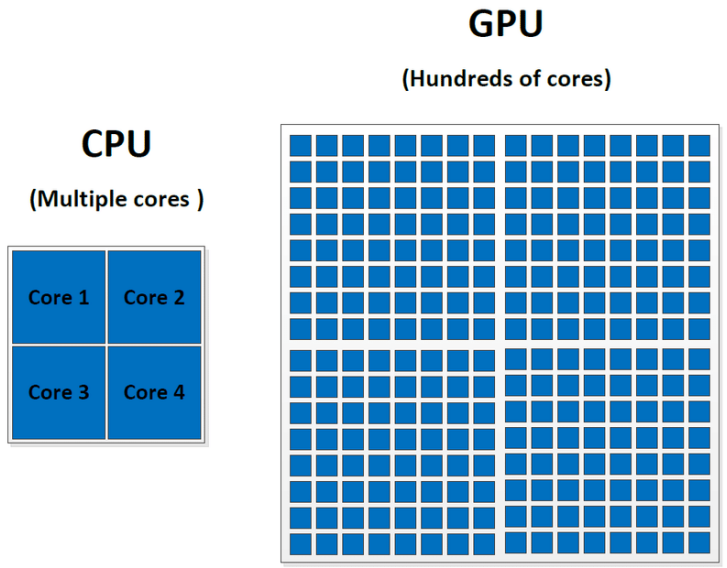


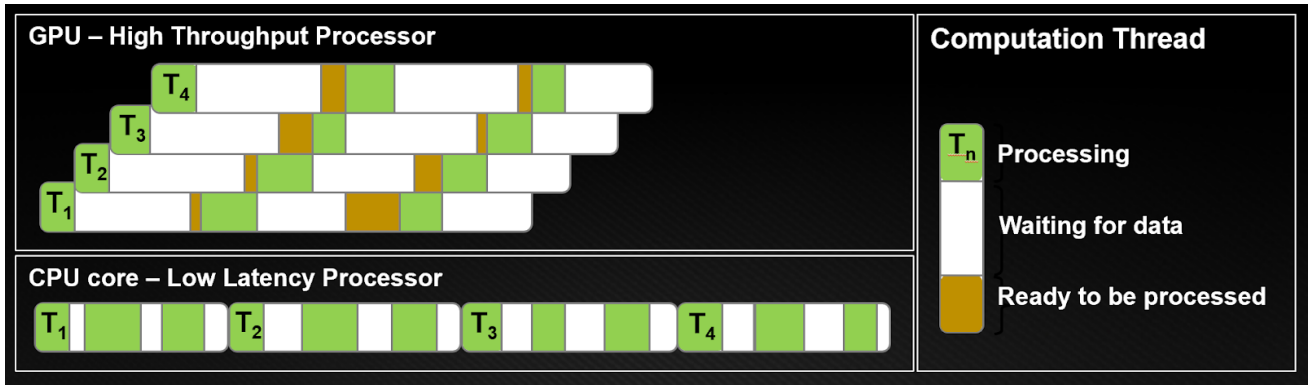
GPT-4(2023.3):
(rumors say)
1.7 trillion (조)
= 1700B

1. Background

- CPU – GPU – NPUNeural Processing Unit

- Why GPU over CPU?
  - CPU의 발전: 반응 시간(latency) 단축 – 순차 처리에 적합
    - 고성능의 Core 개발
  - GPU의 발전: 처리량(throughtput) 확대 – 병렬 처리에 적합
    - 코어 수 증대 (최신 GPU에는 2500~5000개의 ALU)
    - 종류가 같은 다수의 연산들을 한꺼번에 수행 가능 => 행렬 연산에 적합.





1. Background

GPU, CPU – 범용적으로 사용

문제: 폰 노이만 병목 현상

CPU & GPU: 메모리에서 로드 – 계산 – 메모리에 저장

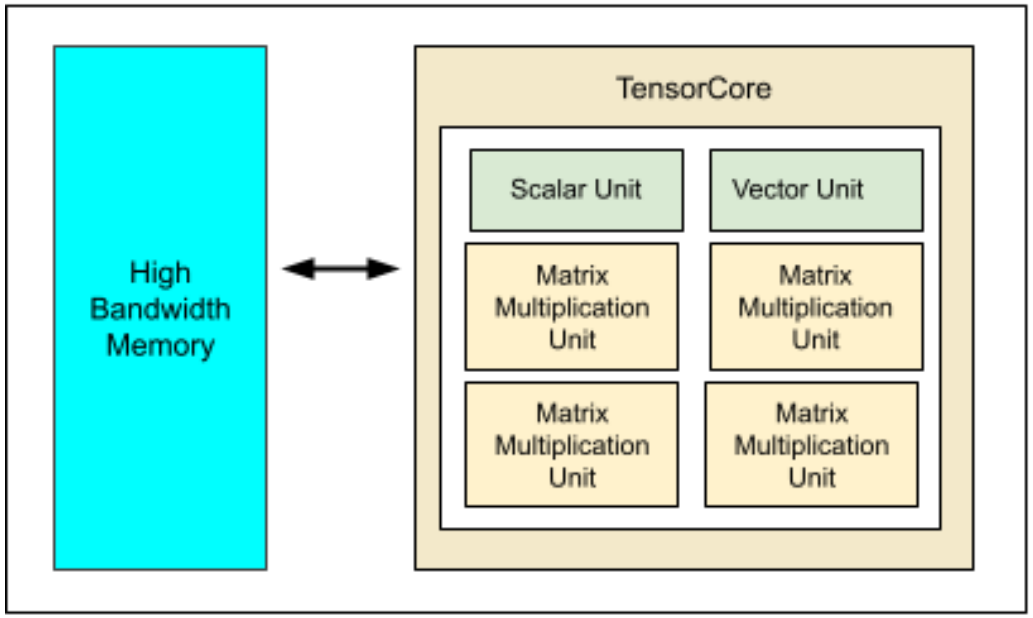
⇒ 메모리 접근 속도가 계산 속도에 비해 느리므로 메모리 액세스 속도가 CPU의 총 처리량을 제한할 수 있다.

➔Google: 신경망에 특화된 **TPU**를 만들자.

대규모 행렬 연산을 고속으로 처리 가능.

2. TPU 특징 (1)

- Matrix Units(MXUs)



Inside a chip of TPU pod
(256 chips in a pod)

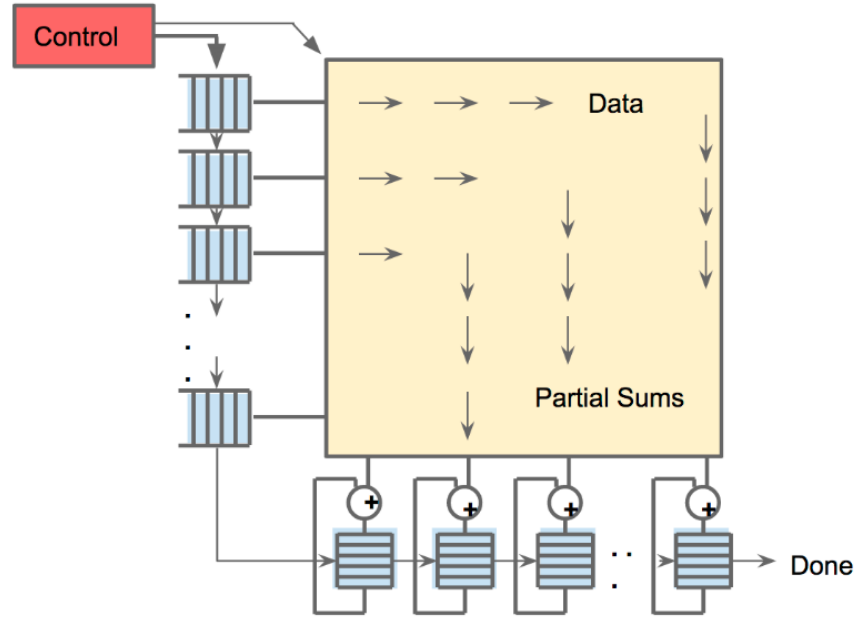
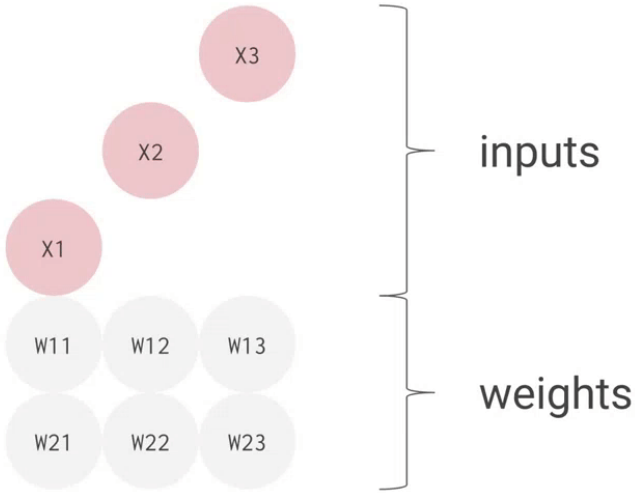


Figure 4. Systolic data flow of the Matrix Multiply Unit.

2. TPU 특징 (1)

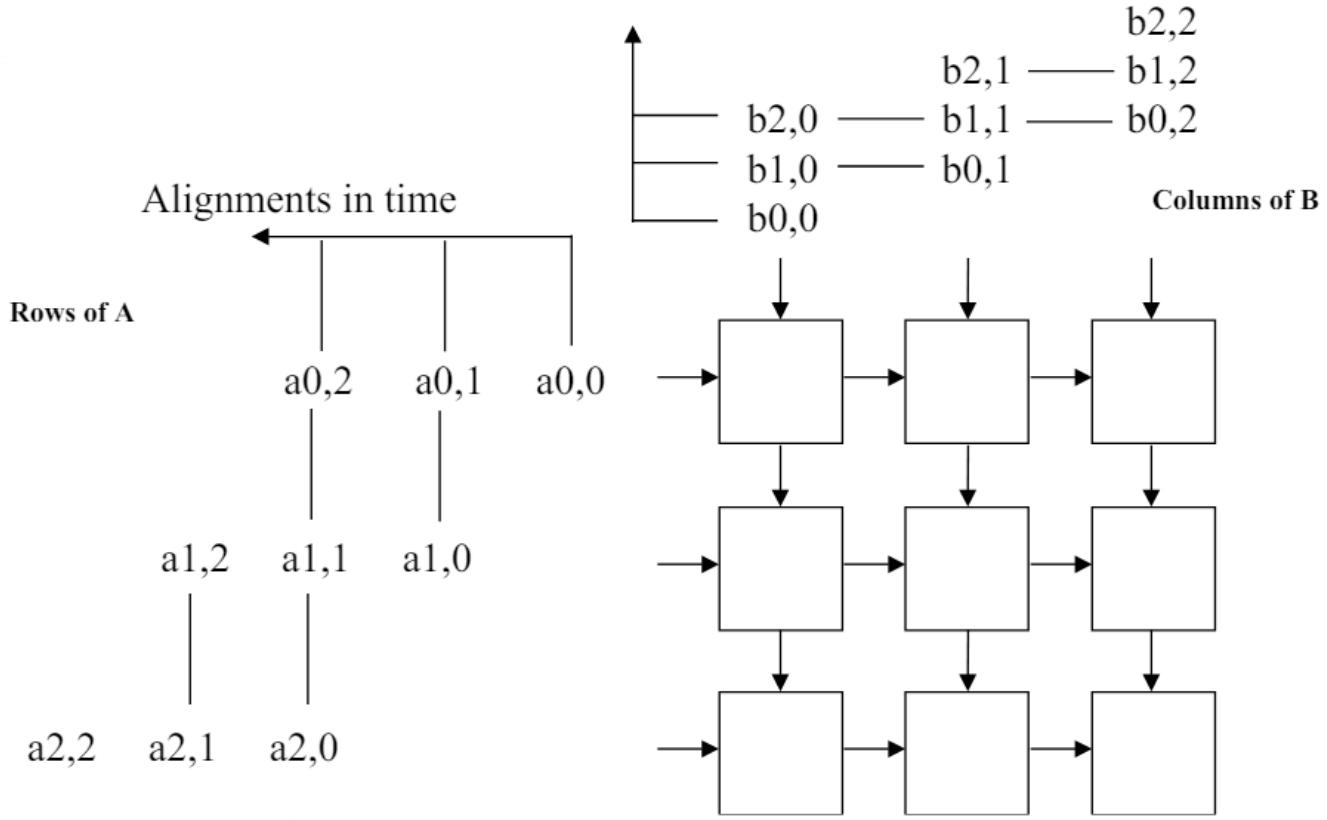
Systolic Array

- Systolic = (혈관 등에서의) 수축
  - 심장이 박동할 때 피가 흐르는 모양새로 데이터가 칩 안에서 움직임.



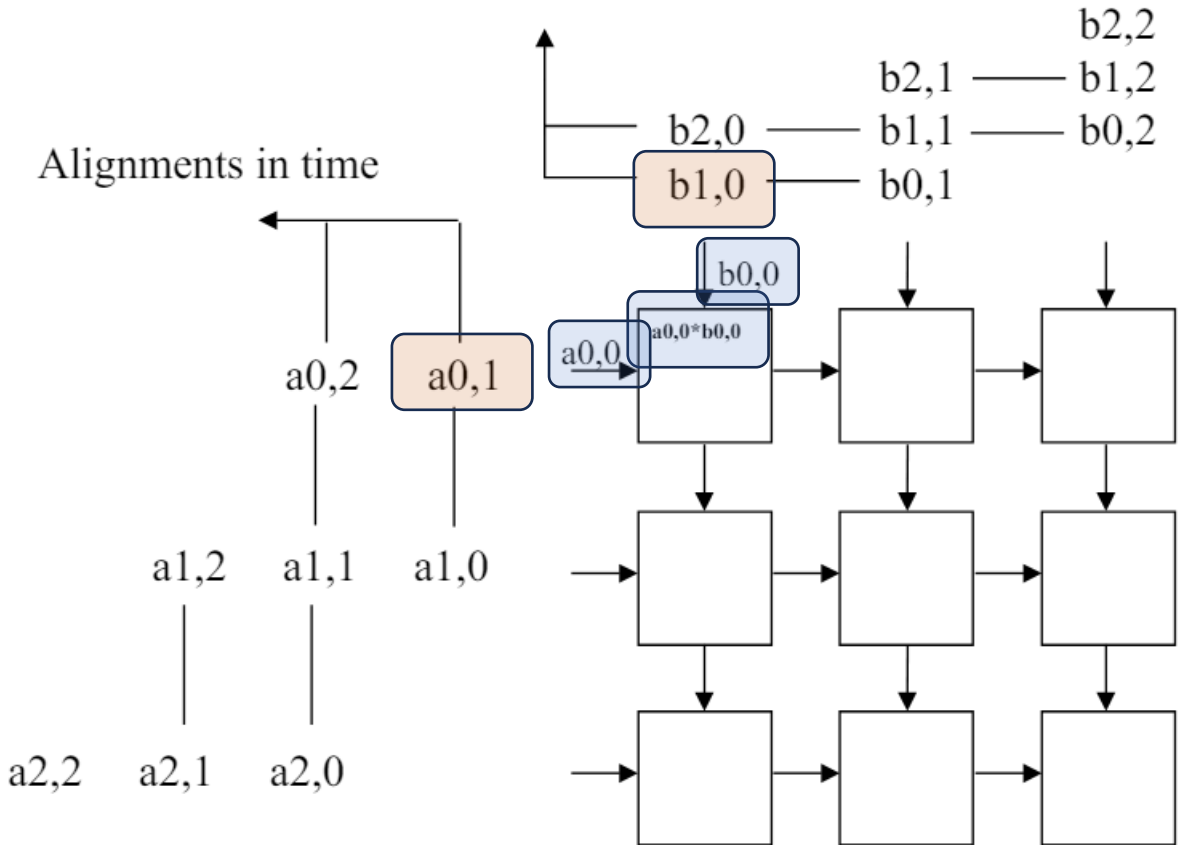
Systolic Array

Time = 0



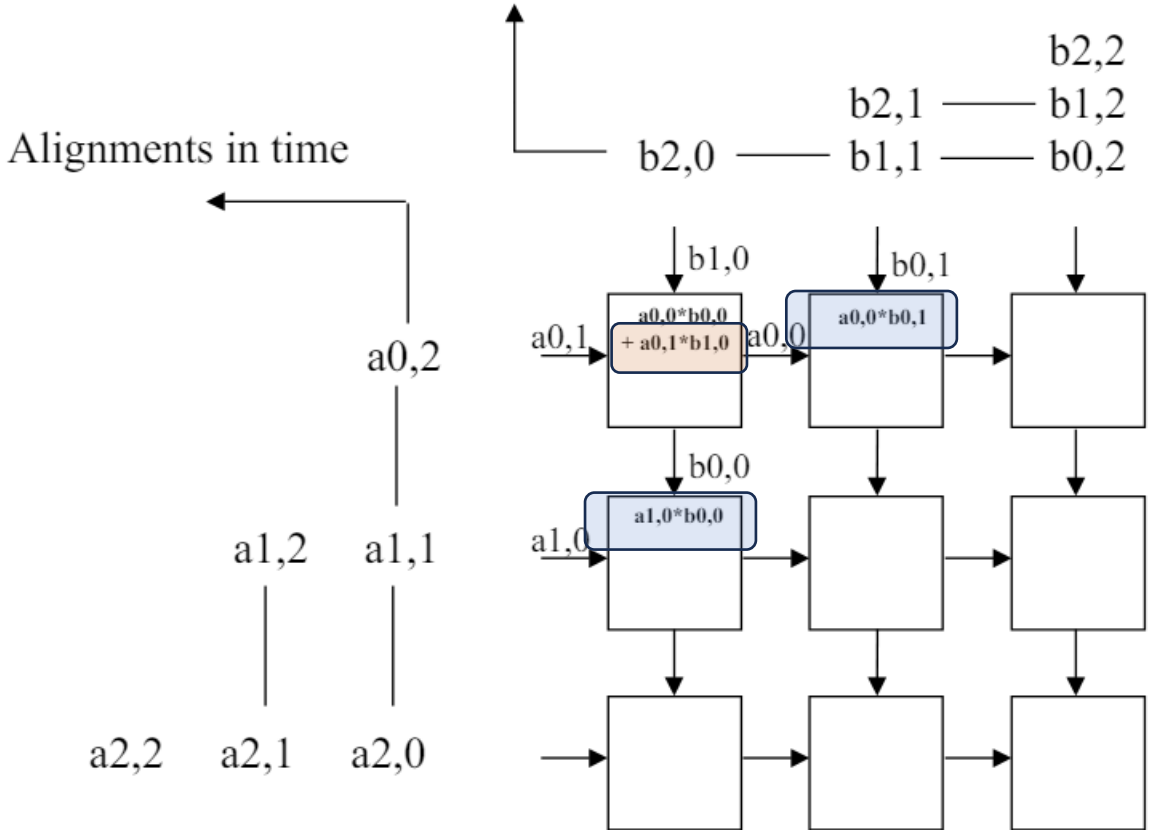
Systolic Array

Time = 1



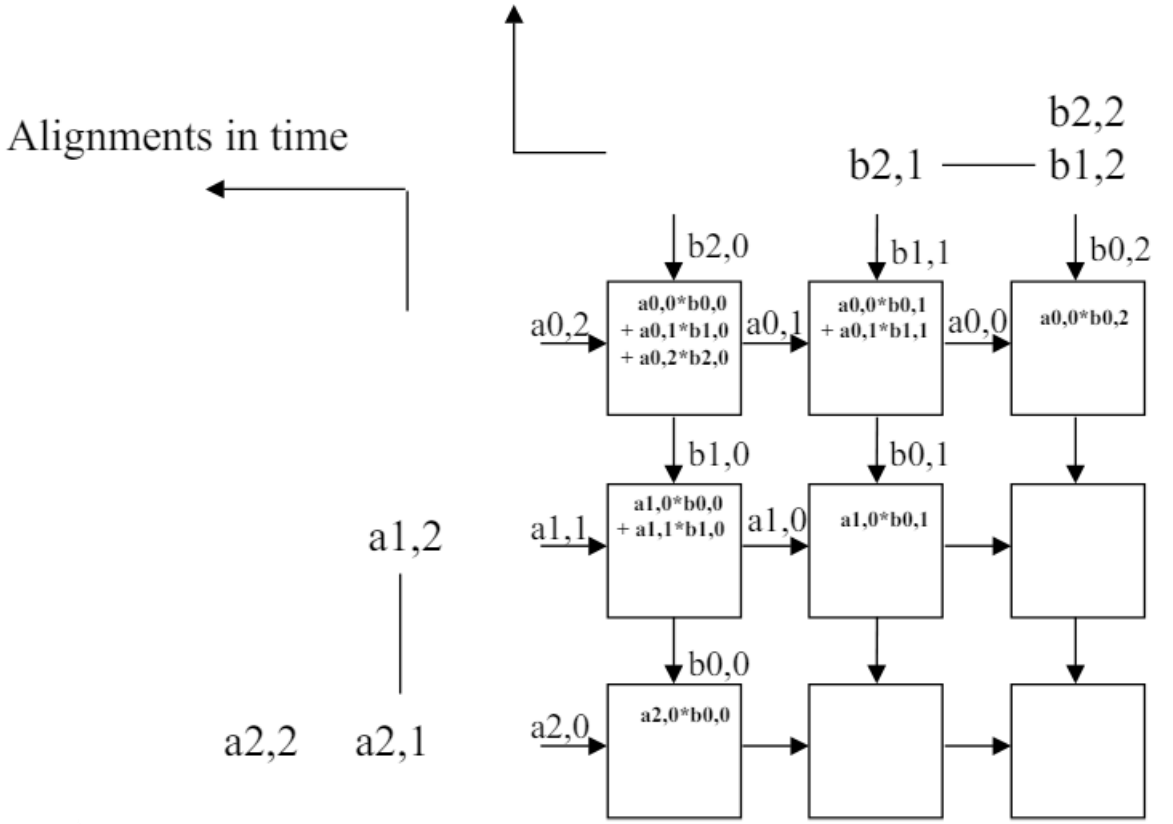
Systolic Array

Time = 2



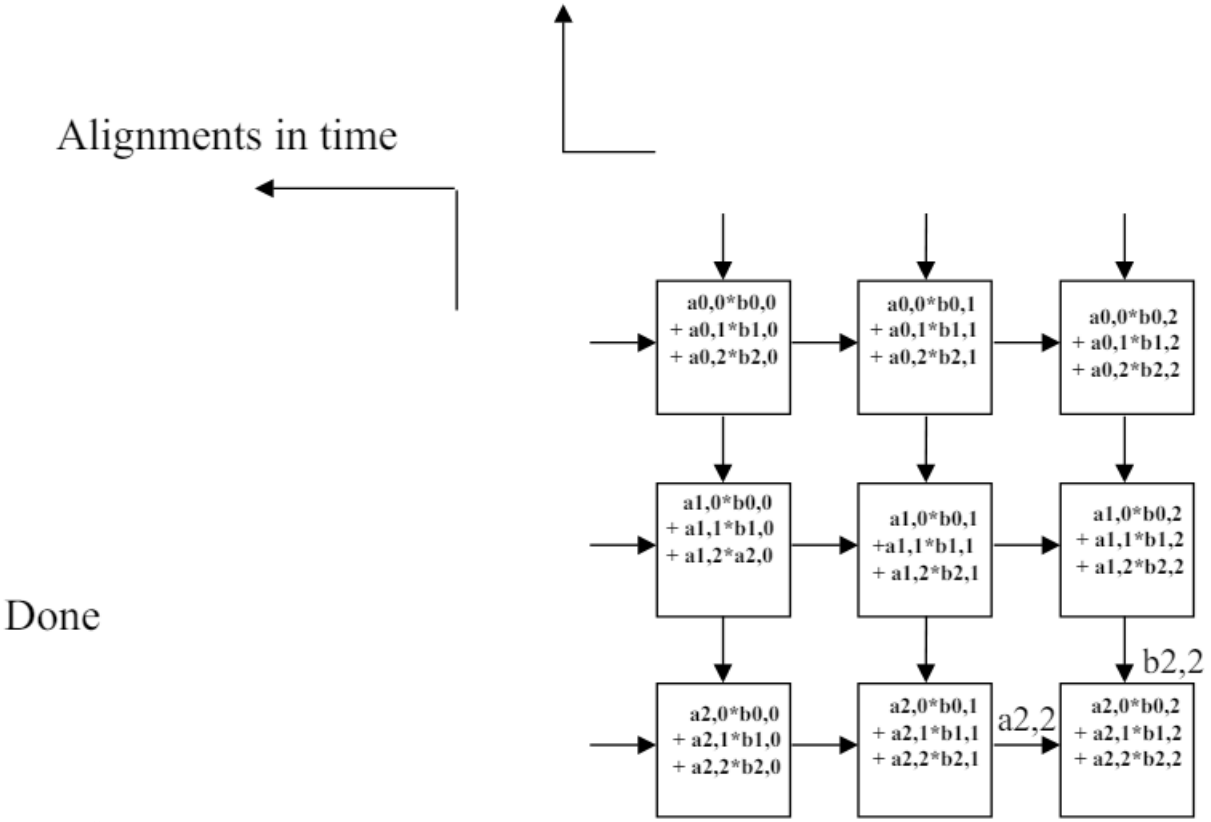
Systolic Array

Time = 3

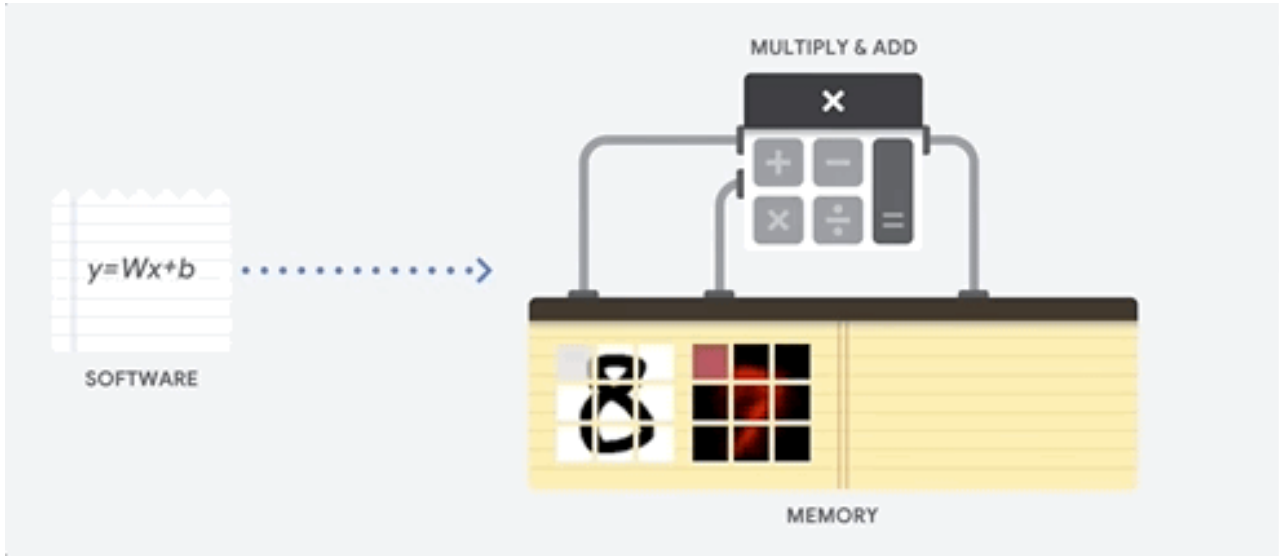


Systolic Array

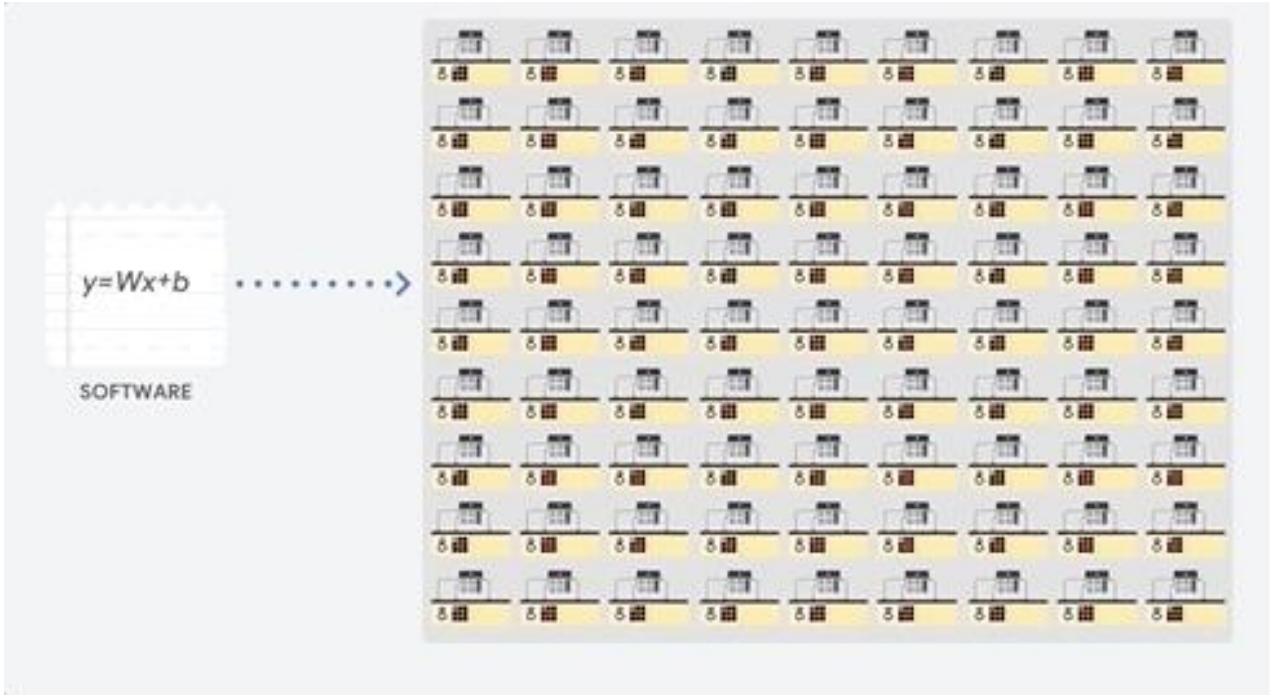
Time = 7



- CPU


MULTIPLY & ADD
OUTPUT
y=Wx+b
SOFTWARE
MEMORY

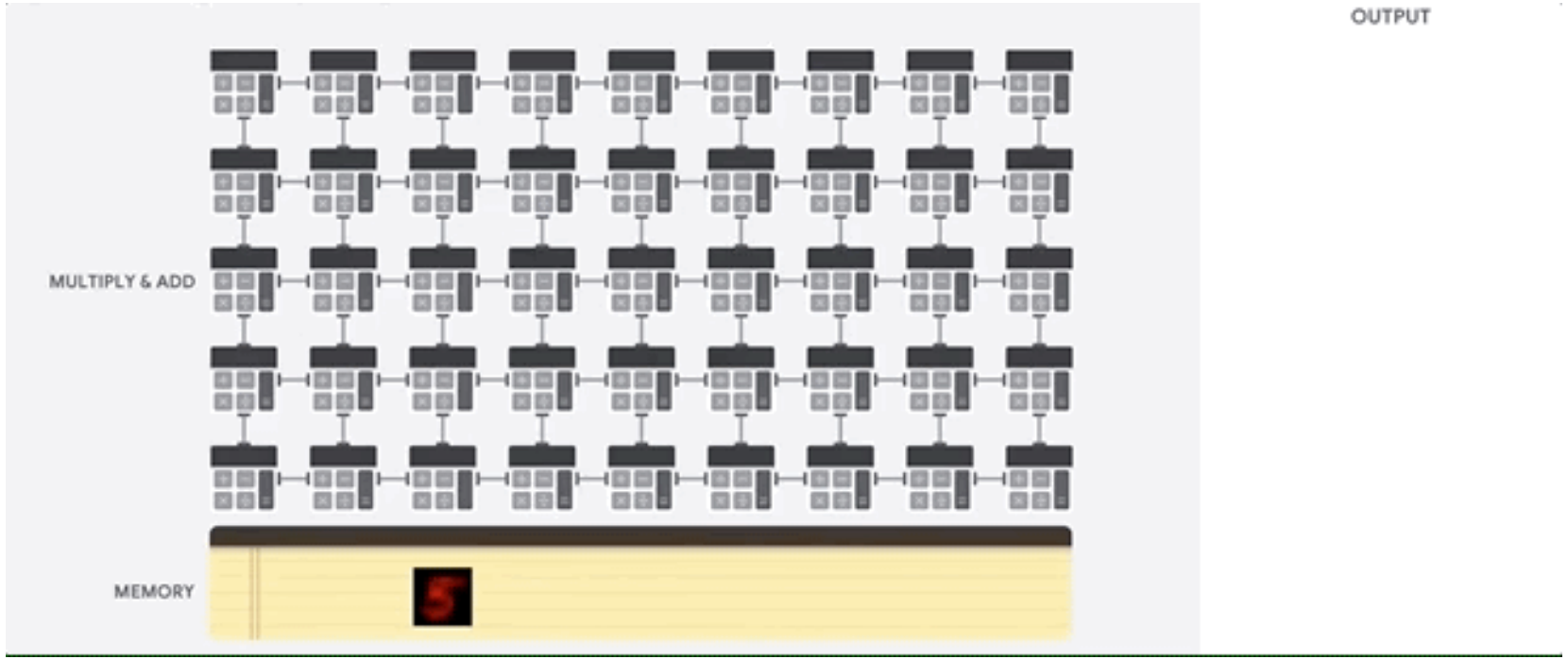
- GPU


OUTPUT
y=Wx+b
SOFTWARE

- TPU

Step 1

메모리의 매개 변수를
MXU로 로드

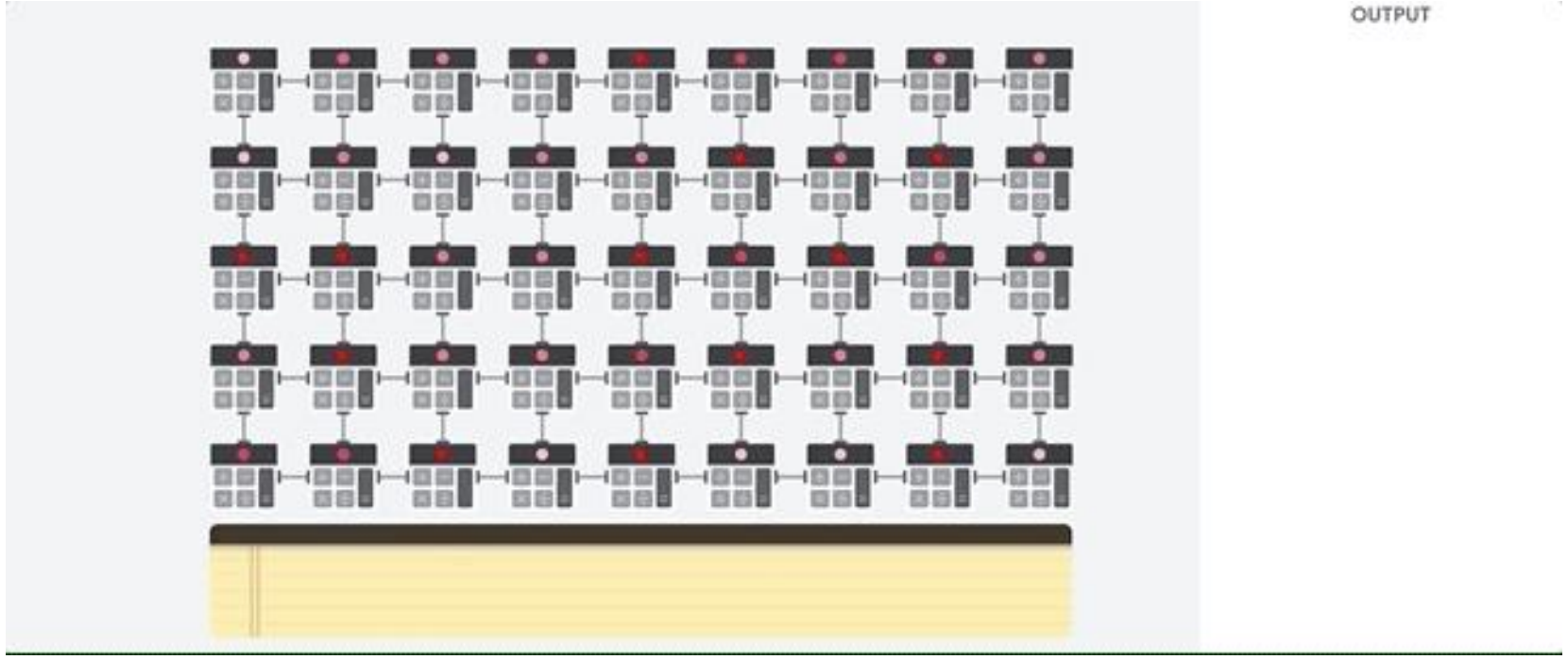


Step 2

메모리에서 데이터를 로드

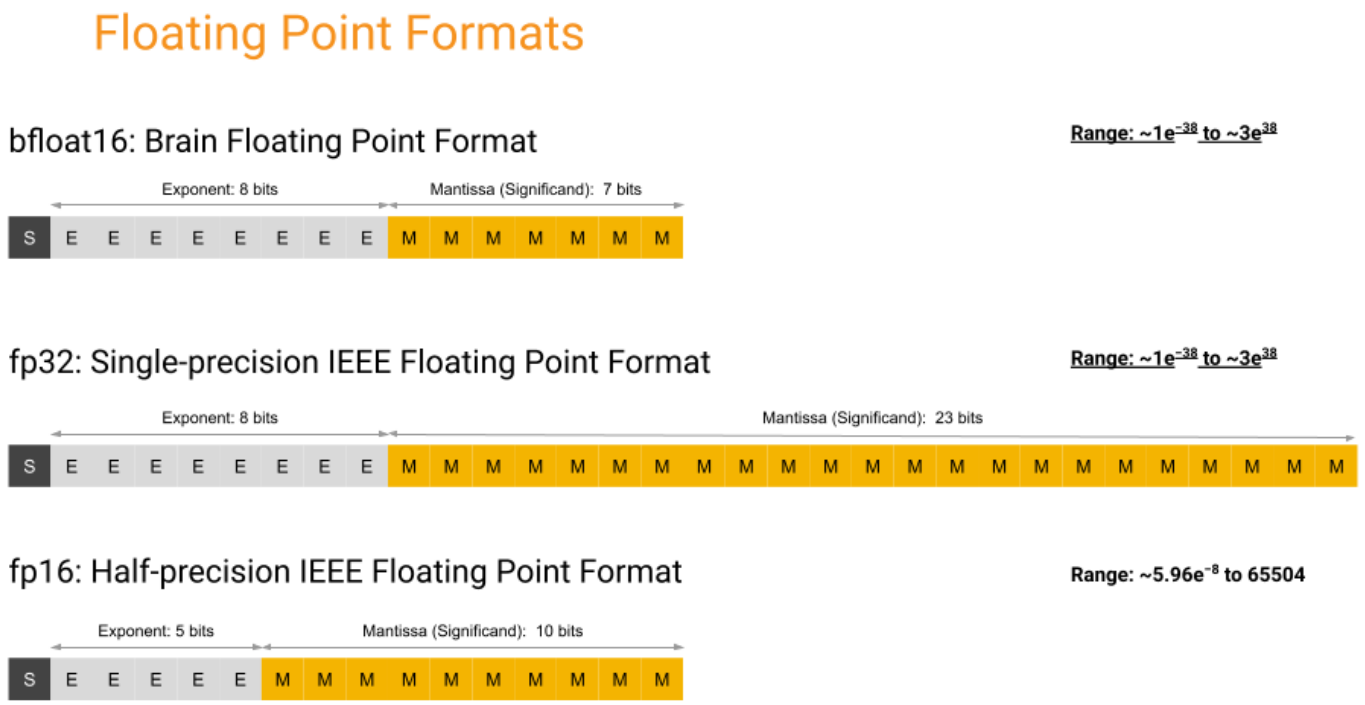
Step 3

데이터와 매개 변수간의
곱셈 값이 다음 누산기로 전달



2. TPU 특징 (2)

- 곱셈: bfloat16, 누적: FP32
  - Low precision: 정확도를 손실하지 않으면서 수렴 시간을 줄이기 위한 일반적인 방법

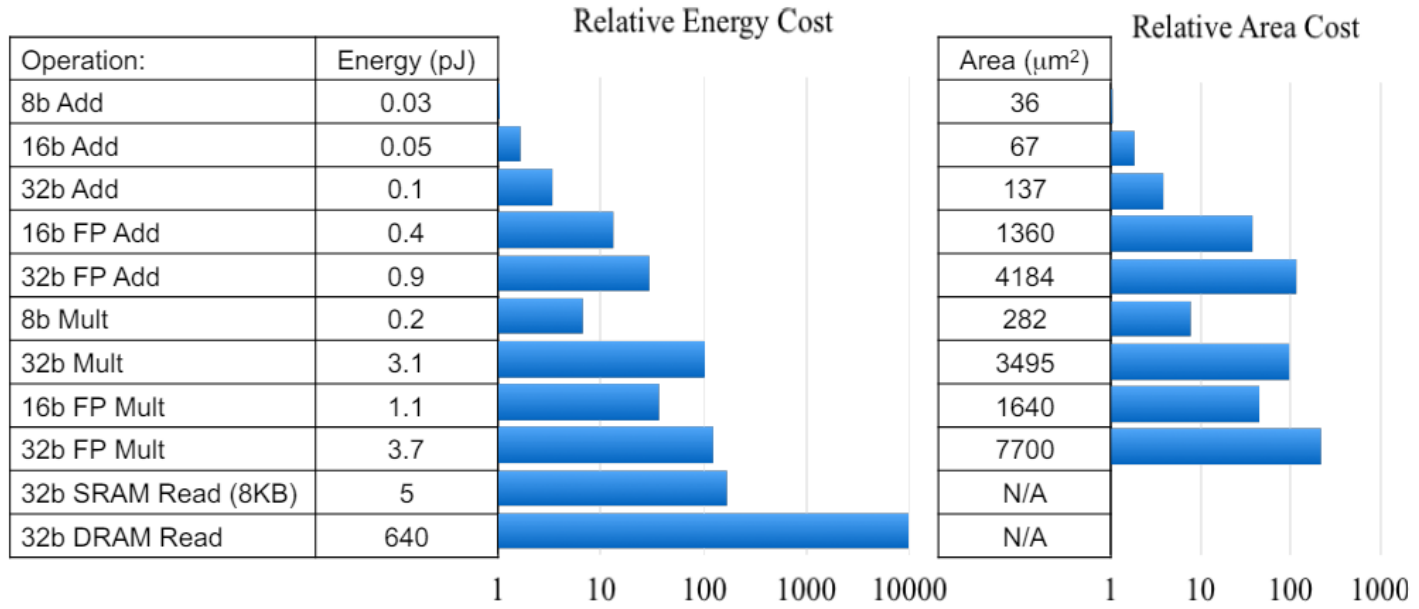


- bfloat16 과 FP32의 다이내믹 레인지는 동일하지만 bfloat16은 메모리 공간의 절반만 차지.

Low Precision

Cost of Operations

Operation:	Energy (pJ)	Area (μm²)
8b Add	0.03	36
16b Add	0.05	67
32b Add	0.1	137
16b FP Add	0.4	1360
32b FP Add	0.9	4184
8b Mult	0.2	282
32b Mult	3.1	3495
16b FP Mult	1.1	1640
32b FP Mult	3.7	7700
32b SRAM Read (8KB)	5	N/A
32b DRAM Read	640	N/A



Energy numbers are from Mark Horowitz "Computing's Energy Problem (and what we can do about it)", ISSCC 2014
Area numbers are from synthesized result using Design Compiler under TSMC 45nm tech node. FP units used DesignWare Library.

2. TPU 특징(3)

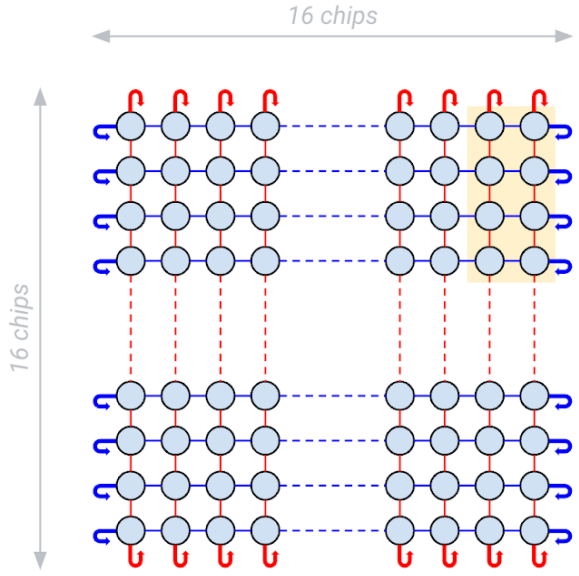
- HBMHigh-Bandwidth Memory
  - Provide larger memory bus width
  - Faster data transfer

- DDR4-3200 bandwidth: 25.6 GB/s

- (2016)TPUv1 bandwidth: 34GB/s
- (2023)TPU v5e bandwidth: 819.2GB/s
- TPU v5e pod bandwidth: more than 400 Tb/s

최신: TPUv5e

- 1 chip – 13B 파라미터
- 256 chips – 2T 파라미터



TPU v5e Chips	Max Model Size* (# parameters)	Reference Models & Sizes
1	13 Billion	LLaMA 2 13B
4	32.5 Billion	LLaMA 32.5B
8	65 Billion	LLaMA 65B
16	175 Billion	GPT-3 175B
32	280 Billion	Gopher 280B
64	540 Billion	PaLM 540B
128	1 Trillion	GLaM 1T
256 (1 Pod)	2 Trillion	Switch Transformer 1.6T