

# AI 芯片 - AI 芯片基础

# 计算时延 Latency



ZOMI



BUILDING A BETTER CONNECTED WORLD

Ascend & MindSpore

[www.hiascend.com](http://www.hiascend.com)  
[www.mindspore.cn](http://www.mindspore.cn)

# Talk Overview

## I. AI 计算体系

- 深度学习计算模式
- 计算体系与矩阵运算

## 2. AI 芯片基础

- 通用处理器 CPU
- 从数据看 CPU 计算
- 通用图形处理器 GPU
- AI专用处理器 NPU/TPU
- 计算体系架构的黄金10年

# 计算强度

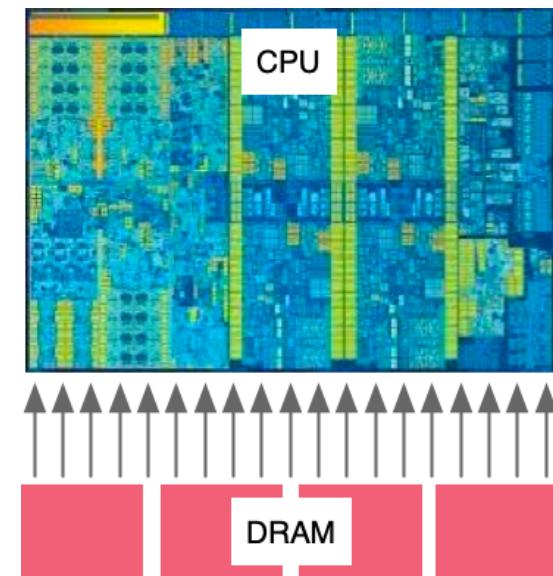
How many operations must I do on some data to make it worth the cost of loading it?

$$\text{Required Compute Intensity} = \frac{\text{FLOPs}}{\text{Data Rate}} = 80$$

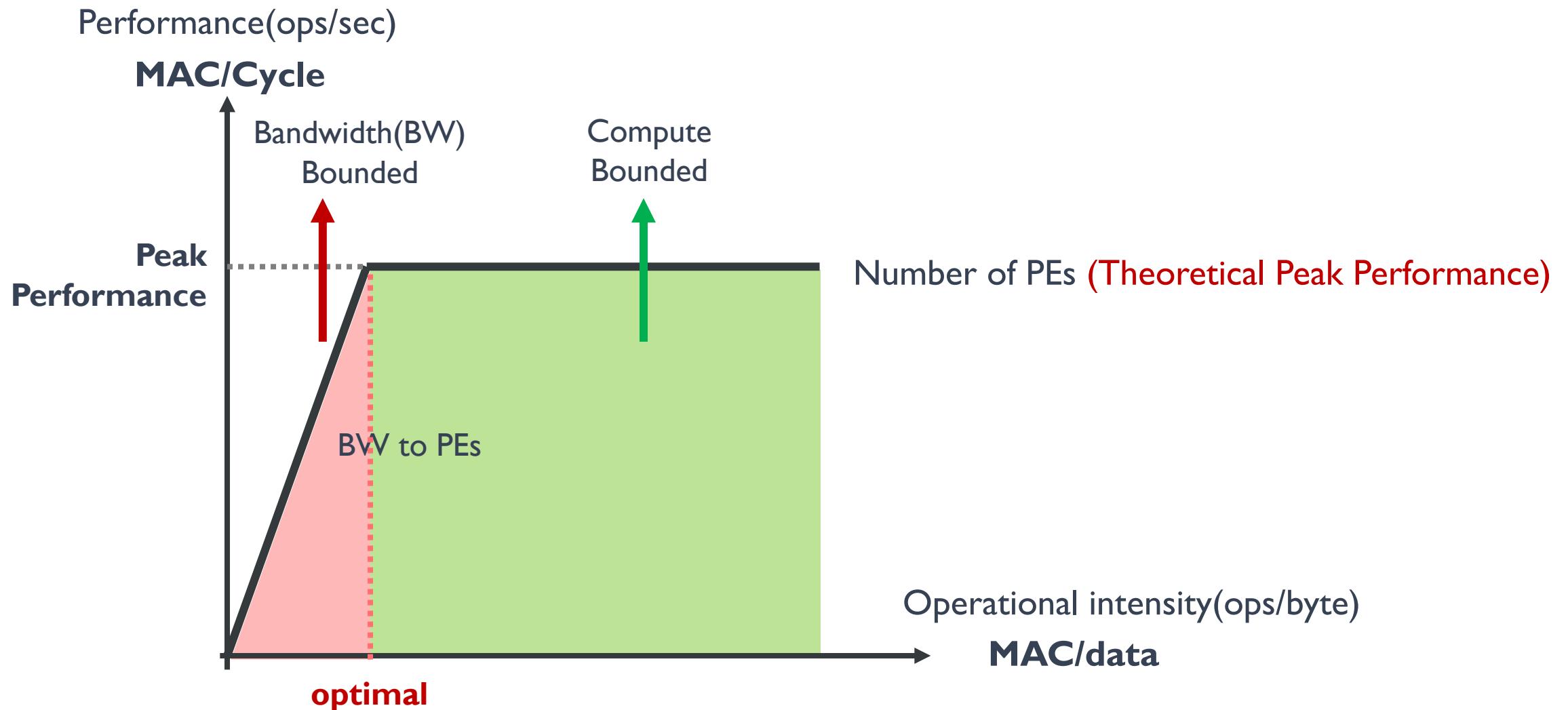
2000 GFLOPs FP64



$$\begin{aligned} & 200 \text{ GBytes / sec} \\ & = 25 \text{ Giga-FP64 / sec} \\ & (\text{FP64} = 8 \text{ bytes}) \end{aligned}$$



So for every number load from memory, Need to do 80 Operations on it to break even.



**更应该关注  
内存、带宽 >> 时延**

# DAXPY 计算 DEMO

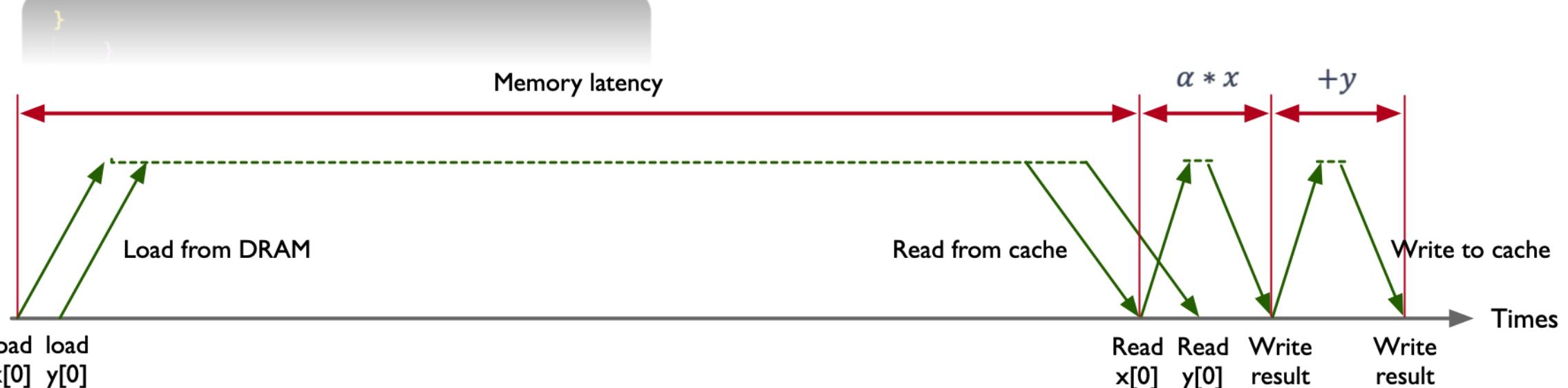
- 2FLOPs : multiply & add
- 2 Memory Loads:  $x[i]$  &  $y[i]$  (per element)
- Single Operation: FMA(fused multiply-add)

```
void demo(double alpha, double *x, double *y)
{
    int n = 2000;
    for(int i = 0; i < n; ++i)
    {
        y[i] = alpha * x[i] + y[i];
    }
}
```

# DAXPY 计算 DEMO

```
void demo(double alpha, double *x, double *y)
{
    int n = 2000;
    for(int i = 0; i < n; ++i)
    {
        y[i] = alpha * x[i] + y[i];
    }
}
```

- 2FLOPs : multiply & add
- 2 Memory Loads:  $x[i]$  &  $y[i]$  (per element)
- Single Operation: FMA(fused multiply-add)



# 光与电的传播速度

Speed of Light = **300,000,000 M/S**

Computer Clock = **3,000,000,000 Hz**

所以在一个时钟周期内光的传播速度为 **100mm (~4 inches)**

# 光与电的传播速度

Speed of Light = **300,000,000 M/S**

Computer Clock = **3,000,000,000 Hz**

Speed of Electricity = **60,000,000 M/S**

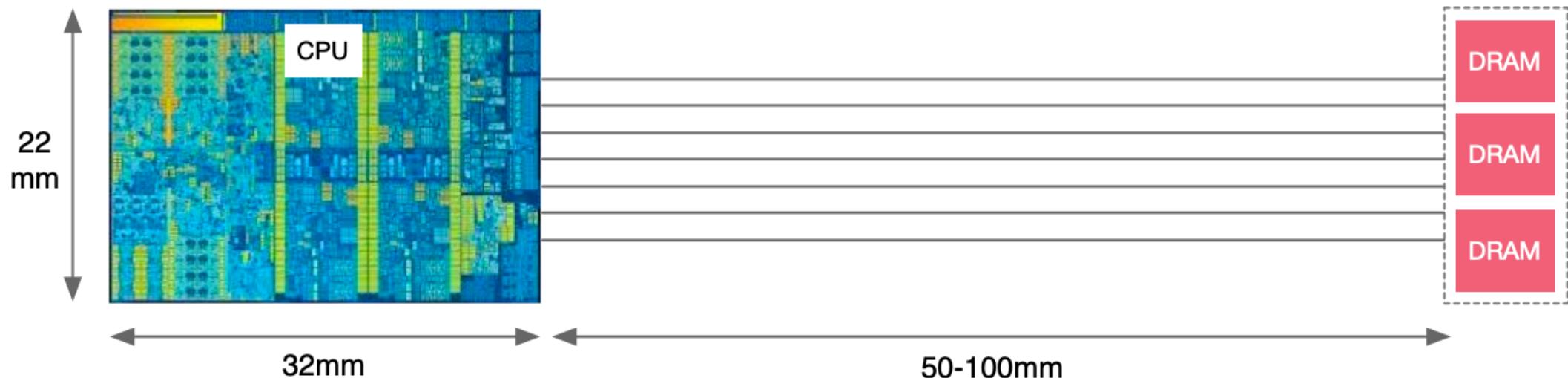
所以在一个时钟周期电流的传播速度为 **20mm (~0.8 inches)**

# 光与电的传播速度

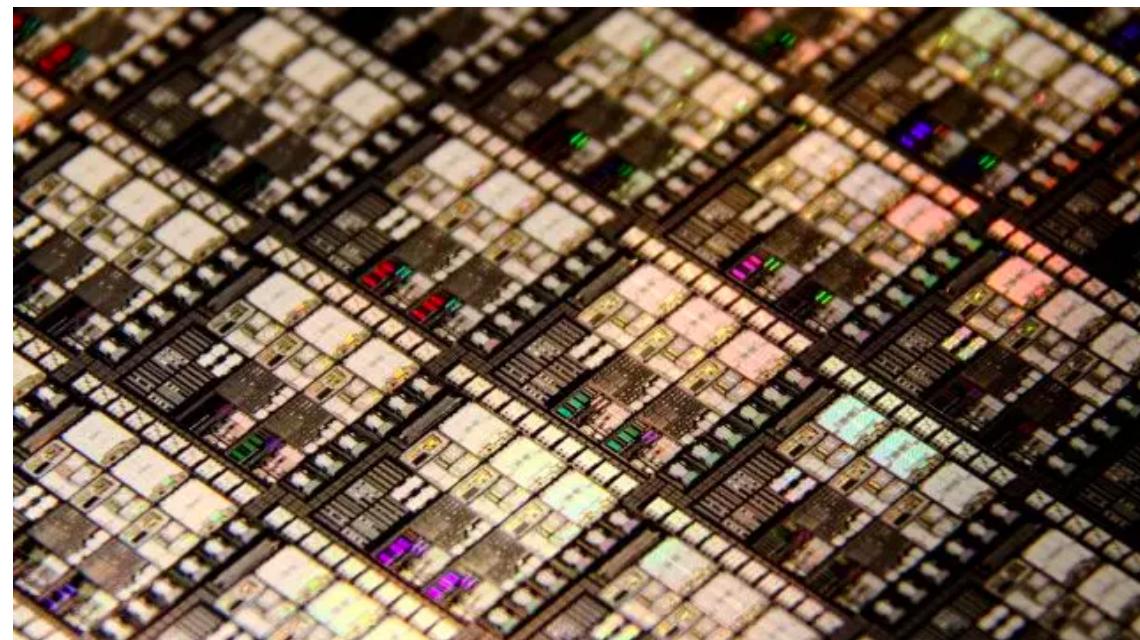
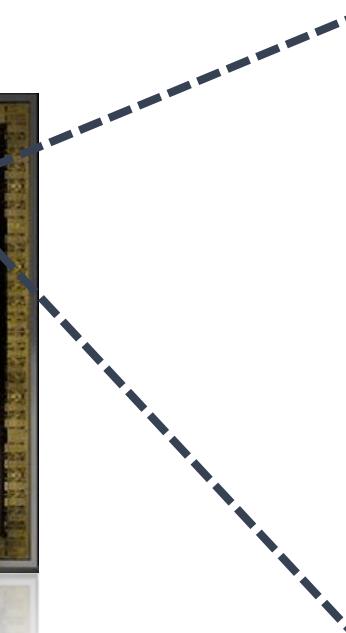
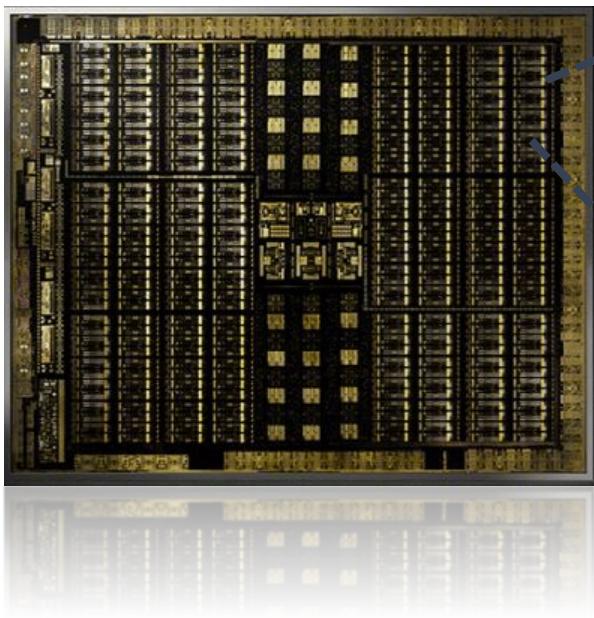
Speed of Light = **300,000,000 M/S**

Computer Clock = **3,000,000,000 Hz**

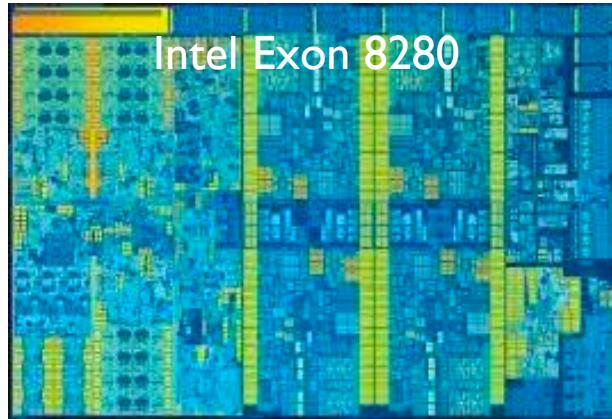
Speed of Electricity = **60,000,000 M/S**



# 处理器内部



# DAXPY 计算 DEMO



Memory Bandwidth: **131** GB/sec

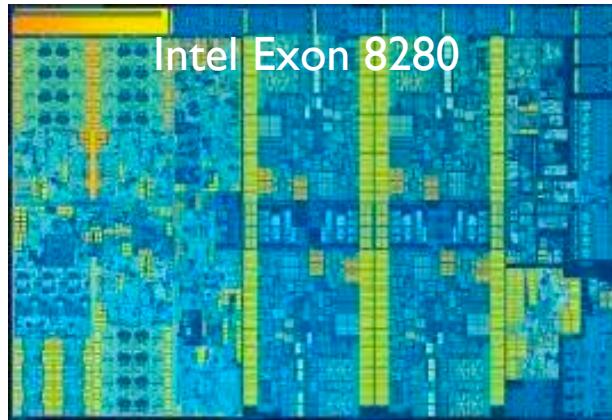
Memory latency: **89** ns

**11,659** bytes can be moved in **89** ns

AXY demo move **16** bytes per **89** ns latency

Memory efficiency = **0.14%**

# DAXPY 计算 DEMO



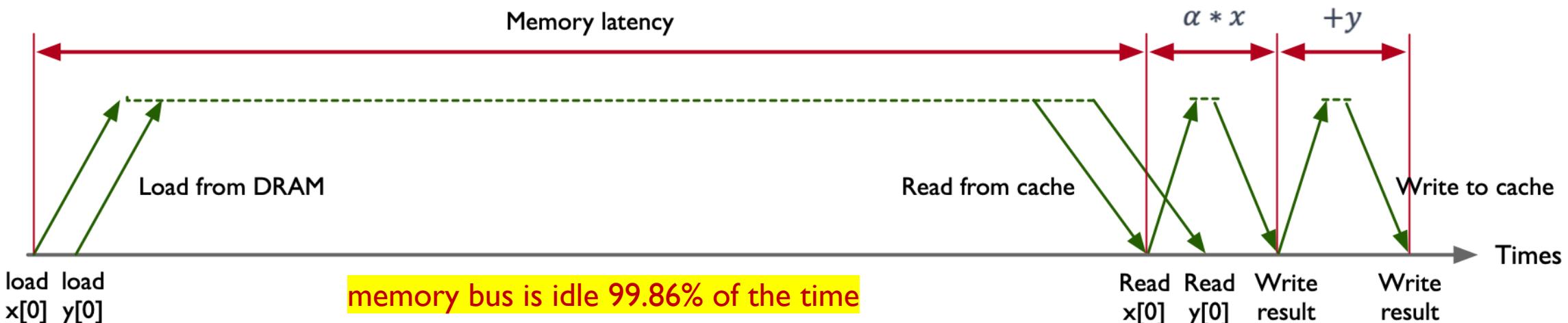
Memory Bandwidth: 131 GB/sec

Memory latency: 89 ns

11,659 bytes can be moved in 89 ns

A XY demo move 16 bytes per 89 ns latency

Memory efficiency = 0.14%



# 不同芯片产品的计算性能

	<b>AMD Rome 7742</b>	<b>Intel Xeon 8280</b>	<b>NVIDIA A100</b>
Memory B/W(GB/sec)	204	131	1555
DRAM Latency(ns)	122	89	404
Peak bytes per latency	24,888	11,659	628,220
Memory Efficiency	0.064%	0.14%	0.0025%

# 引用

1. <https://www.youtube.com/watch?v=3jHi8E5C-18>
2. <https://www.youtube.com/watch?v=-P28LKWTzrl>
3. <https://www.youtube.com/watch?v=3I10o0DYJXg>





BUILDING A BETTER CONNECTED WORLD

THANK YOU

Copyright©2014 Huawei Technologies Co., Ltd. All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.