AI 芯片 – AI 计算体系

AI 芯片指标

ZOMI

Talk Overview

1. 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

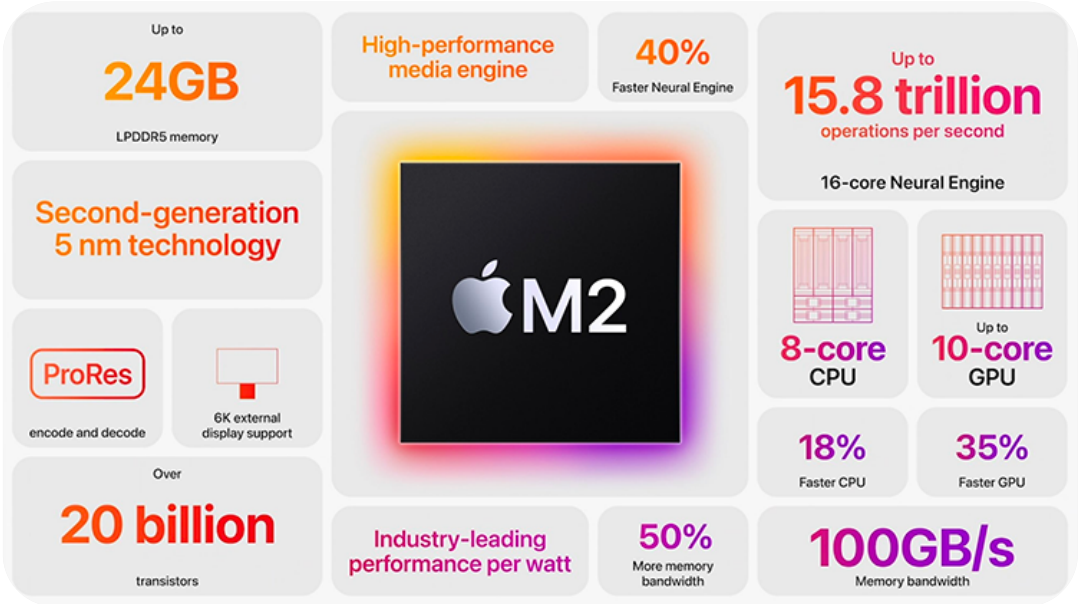
- 通用处理器 CPU
- 从数据看 CPU 计算
- 通用图形处理器 GPU
- AI专用处理器 NPU/TPU
- 计算体系架构的黄金10年

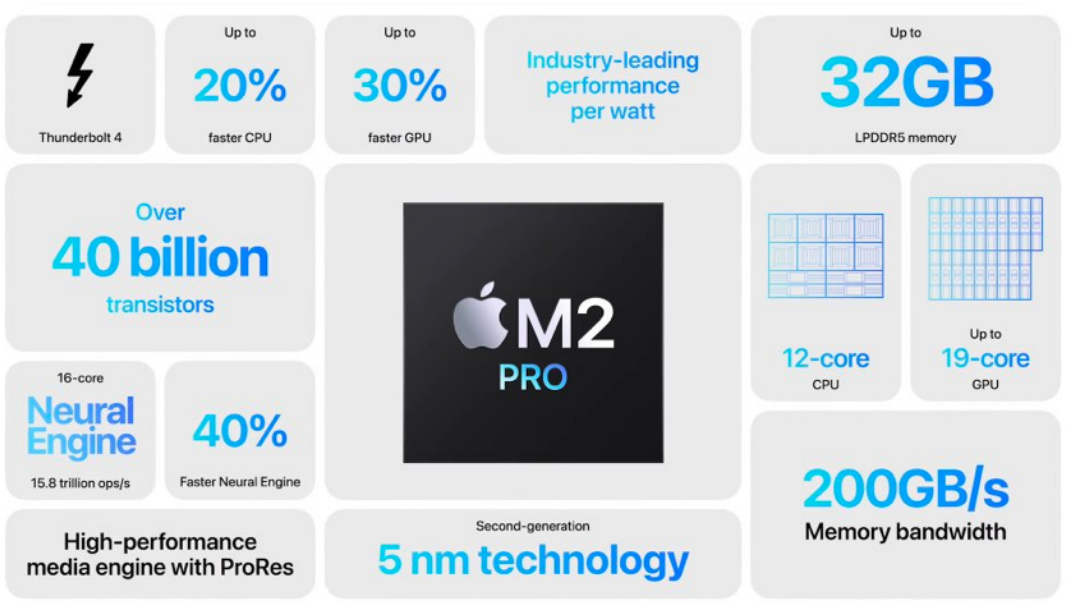
Talk Overview

I. AI 计算体系与矩阵运算

- Key Metrics – AI芯片关键指标
- Bit Width – 比特位数
- Matrix Multiplication – 矩阵运算
- Specialized Hardware – 专用硬件

算力单位





算力单位

OPS

- OPS(Operations Per Second)，1 TOPS 代表处理器每秒进行一万亿次 10^{12} 计算
- OPS/W 每瓦特运算性能，TOPS/W 评价处理器在1W 功耗下运算能力的性能指标

MACs

- Multiply–Accumulate Operations，乘加累积操作。 1MACs包含一个乘法操作与一个加法操作，~2FLOPs，通常MACs与FLOPs存在一个2倍的关系。

MAC

FLOPs

算力单位

FLOPs

- Floating Point Operations，浮点运算次数，用来衡量模型计算复杂度，常用作神经网络模型速度的间接衡量标准。对于卷积层而言，FLOPs的计算公式如下：

$$FLOPs = 2 \cdot H \cdot W \cdot C_{in} \cdot K \cdot K \cdot C_{out}$$

MAC

- Memory Access Cost，内存占用量，用来评价模型在运行时的内存占用情况。 $1x1$ 卷积FLOPs为 $2 \cdot H \cdot W \cdot C_{in} \cdot C_{out}$，其对应MAC为：

$$H \cdot W \cdot (C_{in} + C_{out}) + (C_{in} * C_{out})$$

OPS

MACs

AI 芯片关键指标

Key Metrics

Key Metrics I

1. **精度 Accuracy**
   - 计算精度 (FP32/FP16 etc.)
   - 模型结果精度 (ImageNet 78%)
2. **吞吐量 Throughput**
   - 高维张量处理 (high dimension tensor)
   - 实时性能 (30 fps or 20 tokens)
3. **时延 Latency**
   - 交互应用程序 (TTA)
4. **能耗 Energy**
5. **系统价格 System Cost**
6. **易用性 Flexibility**

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

Table 3: Variations on the Transformer architecture. Unlisted values are identical to those of the base model. All metrics are on the English-to-German translation development set, newstest2013. Listed perplexities are per-wordpiece, according to our byte-pair encoding, and should not be compared to per-word perplexities.

	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train steps	PPL (dev)	BLEU (dev)	params $\times 10^6$
base	6	512	2048	8	64	64	0.1	0.1	100K	4.92	25.8	65
(A)				1	512	512				5.29	24.9	
				4	128	128				5.00	25.5	
				16	32	32				4.91	25.8	
				32	16	16				5.01	25.4	
(B)					16					5.16	25.1	58
					32					5.01	25.4	60
(C)	2									6.11	23.7	36
	4									5.19	25.3	50
	8									4.88	25.5	80
		256			32	32				5.75	24.5	28
		1024			128	128				4.66	26.0	168
			1024							5.12	25.4	53
			4096							4.75	26.2	90
(D)							0.0			5.77	24.6	
							0.2			4.95	25.5	
								0.0		4.67	25.3	
								0.2		5.47	25.7	
(E)	positional embedding instead of sinusoids									4.92	25.7	
big	6	1024	4096	16			0.3		300K	**4.33**	**26.4**	213

Key Metrics II

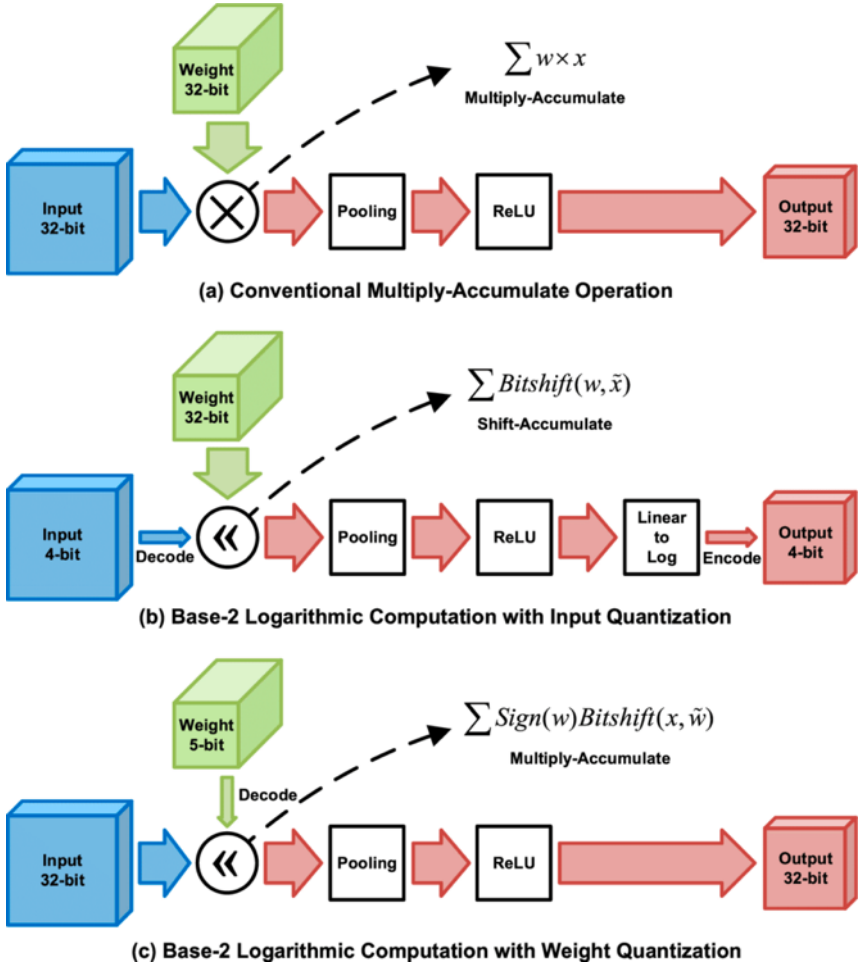
1. 精度 Accuracy
2. 吞吐量 Throughput
3. 时延 Latency
4. **能耗 Energy**
   - IOT 设备有限的电池容量
   - 数据中心液冷等大能耗
5. **系统价格 System Cost**
   - 硬件自身的价格 $$$
   - 系统集成上下游全栈等成本
6. **易用性 Flexibility**
   - 衡量开发效率和开发难度

AI加速器的关键设计点

- 提升吞吐量 Increase Throughput 和降低延时 Reduce Latency
- 低延时 Low Latency 和 Batch Size 之间 Tradeoff

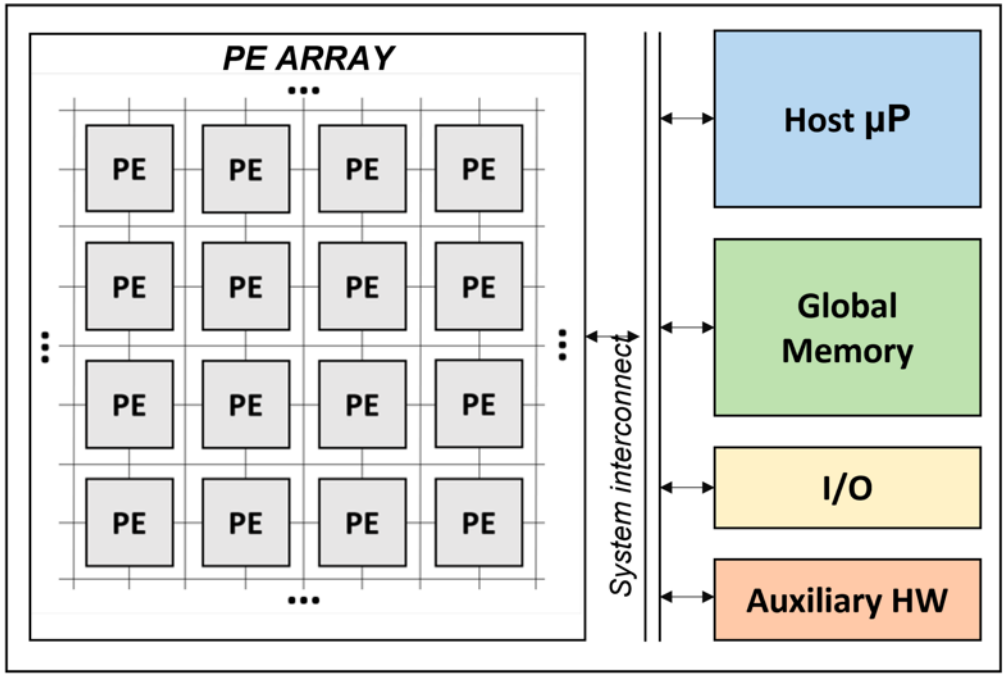
AI加速器的关键设计点：MACs

1. **去掉没有意义的 MACs**
   - 增加对稀疏数据的硬件结构 sparse data
   - 控制流控制和执行 control flow
   - 节省时钟周期 save cycles
2. **降低每次 MAC 的计算时间**
   - 增加时钟频率 clock frequency
   - 减少指令开销 instruction overhead



AI加速器的关键设计点： PE, Processing Elements

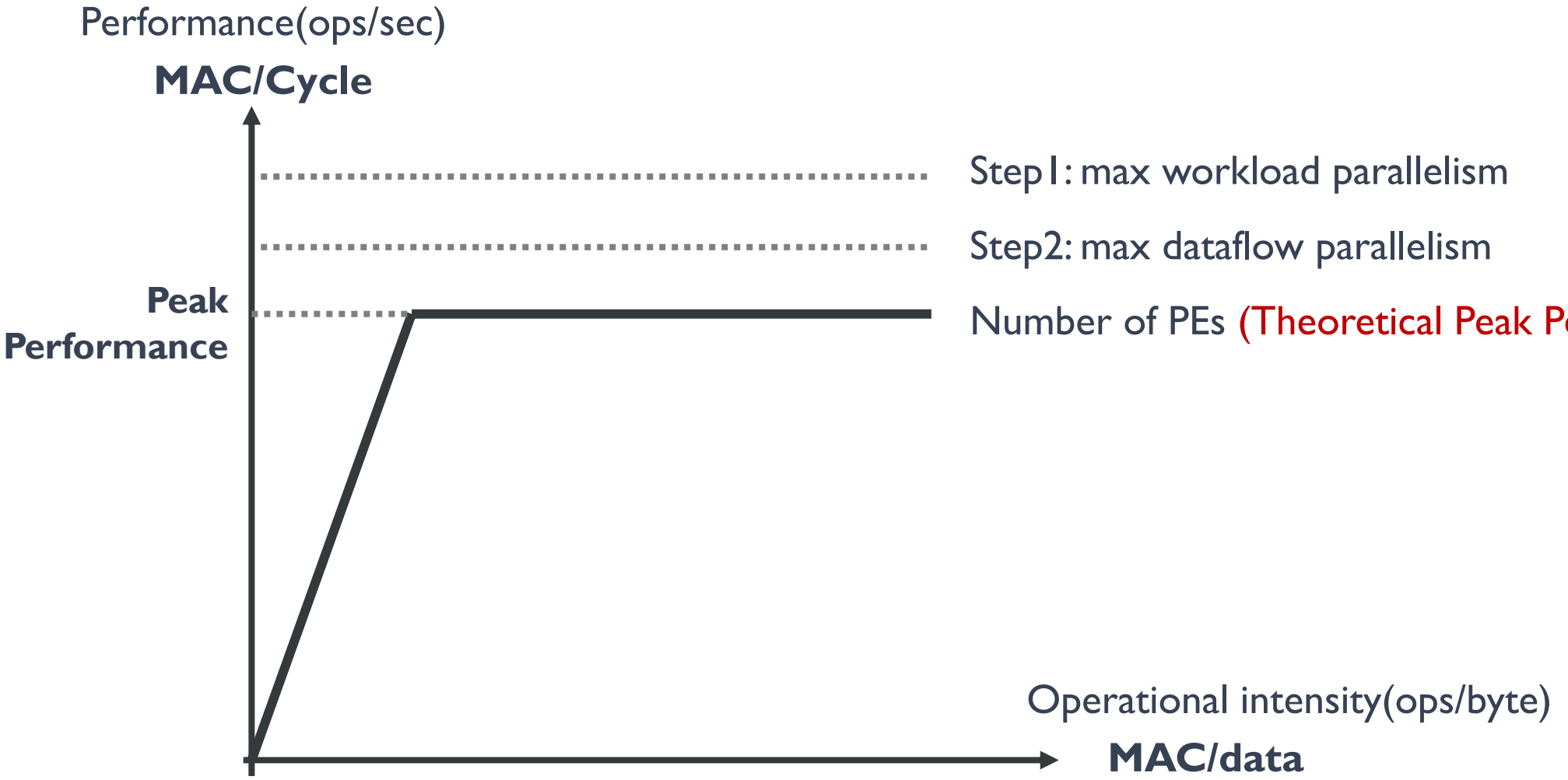
1. **增加 PE 的核心数量**
   - 更多的 MACs 并发 (parallel)
   - 使用更高纳米制程，增加 PE 的面积密度 (area density)
2. **增加 PE 的利用率 utilization**
   - 将计算负载尽可能分配到不同 PE (distribute workload)
   - 均衡 PE 之间计算负载 (balance workloads)
   - 合适的内存带宽有效降低空闲时钟周期 (reduce idle cycles)



Question ?

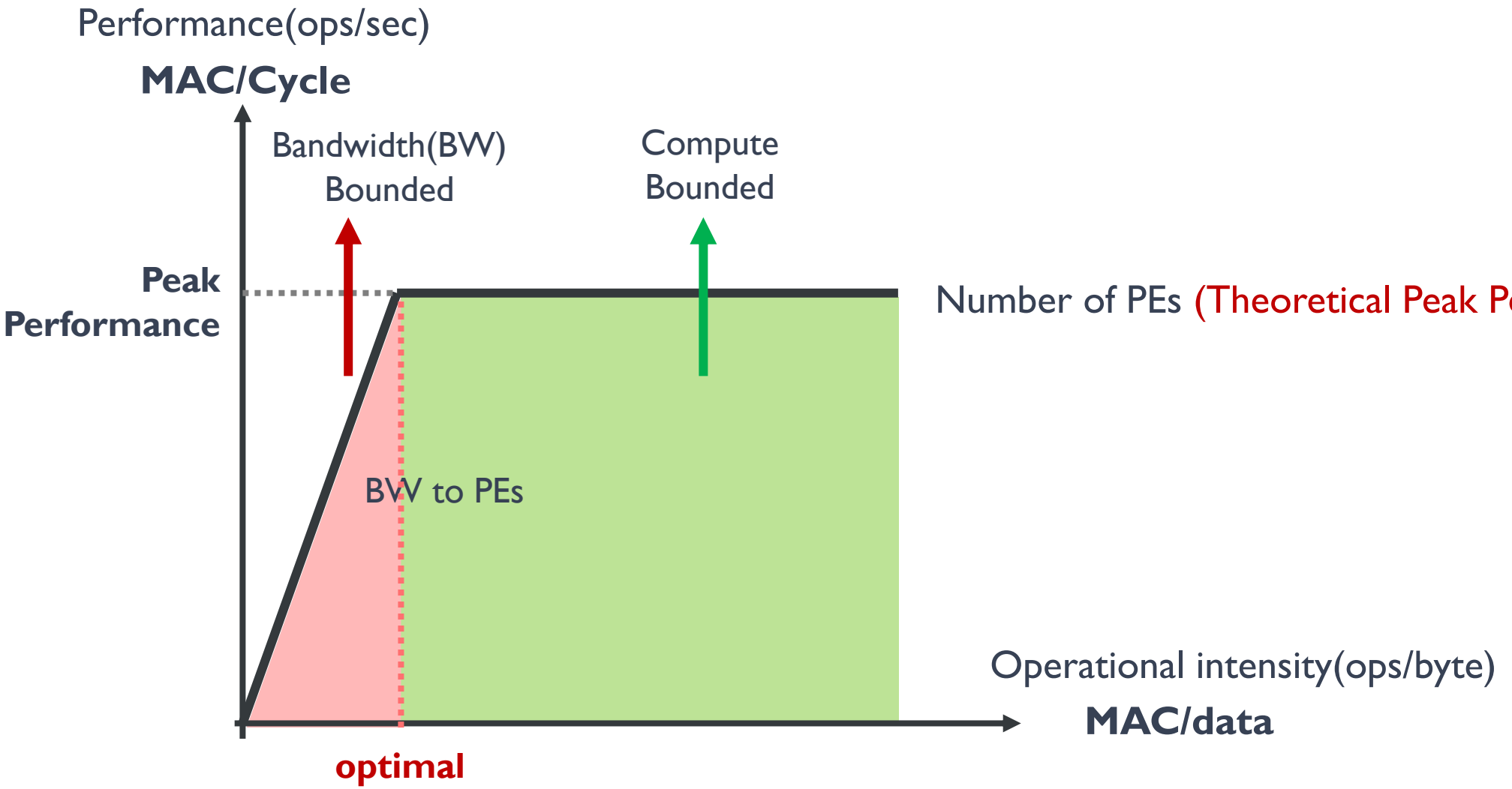
- AI芯片的关键指标里面有吞吐量 Throughput 和时延 Latency，这个主要是由什么产生的吗？
- AI芯片的关键设计点是 MACs 和 PE，感觉这个主要是针对提升单个核心的计算能力吗？

计算性能仿真



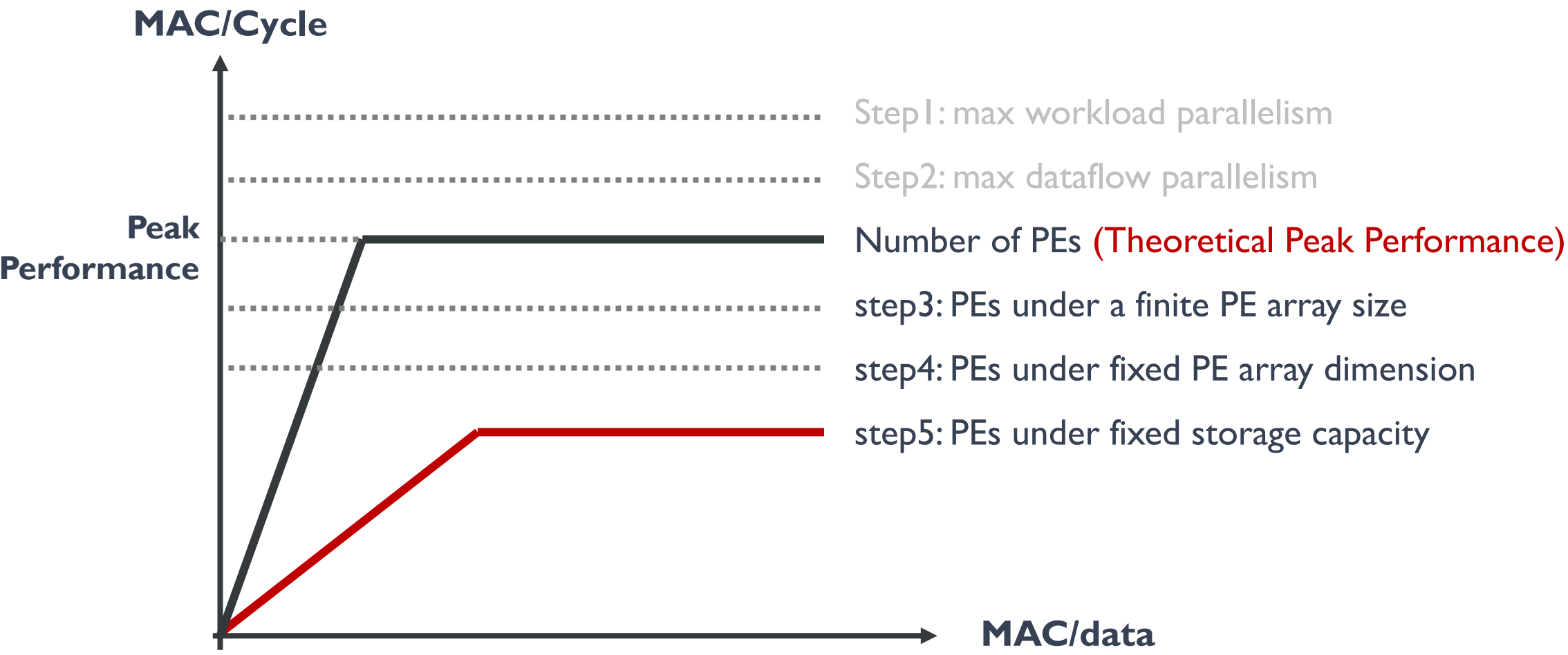
https://arxiv.org/abs/1807.07928

计算性能仿真



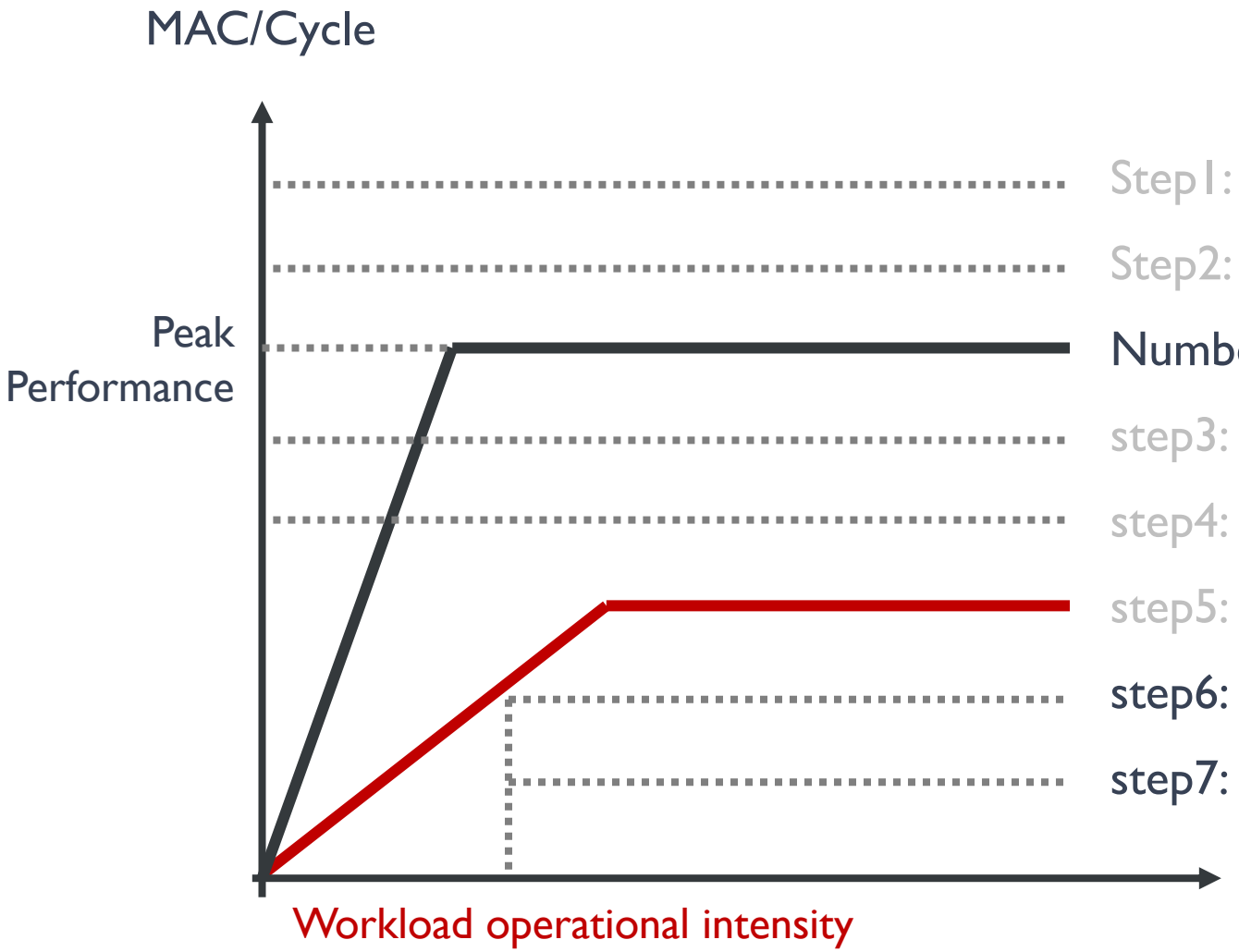
https://arxiv.org/abs/1807.07928

计算性能仿真



https://arxiv.org/abs/1807.07928

计算性能仿真



Step1: max workload parallelism

Step2: max dataflow parallelism

Number of PEs (Theoretical Peak Performance)

step3: PEs under a finite PE array size

step4: PEs under fixed PE array dimension

step5: PEs under fixed storage capacity

step6: lower PE utilization due to insufficient average BW

step7: lower PE utilization due to insufficient BW

https://arxiv.org/abs/1807.07928

计算性能仿真

Step	Constraint	Type	New Performance Bound	Reason for Performance Loss
1	Layer Size and Shape	Workload	Max workload parallelism	Finite workload size
2	Dataflow loop nest	Architectural	Max dataflow parallelism	Restricted dataflow mapping space by defined by loop nest
3	Number of PEs	Architectural	Max PE parallelism	Additional restriction to mapping space due to shape fragmentation
4	Physical dimensions of PEs array	Architectural	Number of active PEs	Additional restriction to mapping space due to shape fragmentation for each dimension
5	Fixed Storage Capacity	Architectural	Number of active PEs	Additional restriction to mapping space due to storage of intermediate data (de- pends on dataflow)
6	Fixed Data Bandwidth	Microarchitectural	Max data bandwidth to active PEs	Insufficient average bandwidth to active PEs
7	Varying Data Access Patterns	Microarchitectural	Actual measured performance	Insufficient instant bandwidth to active PEs

https://arxiv.org/abs/1807.07928

AI芯片的设计目标 I

- **降低功耗**
  - 减少 each MAC 功耗
  - 避免无效 MACs 计算
  - 减少耗能的数据格式搬运 >> 数据重用
- **芯片功耗的降低受到散热的影响**
- **MACs 并行计算的吞吐提升会增加功耗**

Operation	Energy(pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Multiply	0.2
32b Multiply	3.1
16b FP Multiply	1.1
32b FP Multiply	3.7
32b SRAM Read	5
32b DRAM Read	640

Key Metrics 与计算体系思考 I

1. **精度 Accuracy**
   - 能够处理各类型的无规则数据 >> 异构平台
   - 能够应对复杂网络模型结构 >> 计算冗余性
2. **吞吐量 Throughout**
   - 除了峰值算力，看 PE 的平均利用率 >> 负载均衡
   - SOTA网络模型的运行时间 >> MLPerf
3. **时延 Latency**
   - 通信时延对 MACs 的影响 >> 优化带宽
   - Batch Size 大小与内存大小 >> 多级缓存设计

Key Metrics 与计算体系思考 II

4. 能耗 Energy

- 执行SOTA网络模型时候 Ops/W >> 部署场景
- 内存读写功耗 (e.g., DRAM) >> 降低能耗

5. 系统价格 System Cost

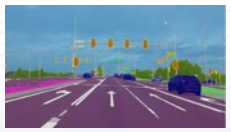
- 片内多级缓存 Cache 大小 >> 内存设计
- PE 数量、芯片大小、纳米制程 >> 电路设计

6. 易用性 Flexibility

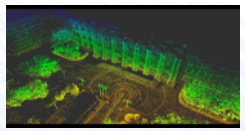
- 对主流AI框架支持度 (PyTorch) >> 软件栈

AI 芯片主要指标

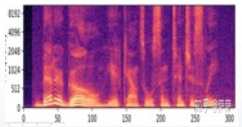
行业应用


自动驾驶


姿态识别


三维重建

自然语言处理


音频信号识别

应用使能

ModelArts
华为云服务

MindX SDK
昇腾行业SDK

操作系统

鸿蒙

Android OS　　Harmony OS　　Linux OS

AI开源框架

MindSpore
全场景深度学习框架

异构驱动

CANN
统一异构驱动架构，释放硬件澎湃算力

华为硬件

端(IOT)　　边(基站)　　云

AI 芯片主要指标

- AI 芯片不仅仅作为一款硬件，更是对客户、应用提供全栈的解决方案，包括SDK、集群管理、AI框架、AI编译器、AI驱动和通信、硬件IC产品化，最后才是AI芯片。而 AI 芯片中最重要的指标有：

1. **精度 Accuracy**：决定是否能够解决具体的 AI 业务和深度学习模型；
2. **功耗 Energy**：决定 AI 芯片对应产品形态具体部署在端侧、边侧还是云测；
3. **时延和吞吐 Latency & throughput**：决定 AI 芯片主要性能，是否计算得足够快；

Reference

1. https://en.wikipedia.org/wiki/Apple_M1
2. https://www.sciencedirect.com/science/article/pii/B9780128134530750011 8
3. Base-Reconfigurable Segmented Logarithmic Quantization and Hardware Design for Deep Neural Networks
4. Eyeriss v2: A Flexible Accelerator for Emerging Deep Neural Networks on Mobile Devices