

04835370

人工智能芯片设计导论

Introduction to Artificial Intelligence Chips

Fall 2023

---

燕博南

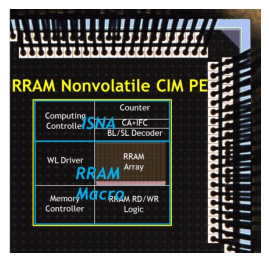
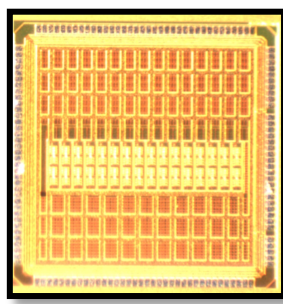
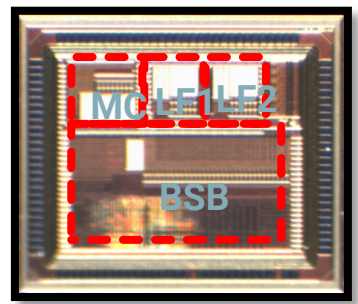
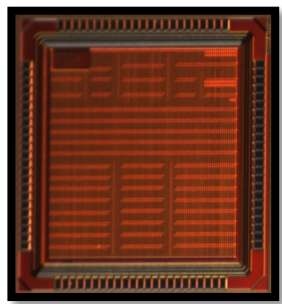
# 课程教授：燕博南， Ph.D.



- Education:
  - PhD Duke University, 2020
- Research:
  - In-Memory Computing Circuits & Systems
  - Domain-Specific Accelerators
  - Emerging CPU Microarchitecture
- Start-up:
  - Pimchip Technology
- Bilibili: Dr燕同学



## What I am working on



## Where I worked

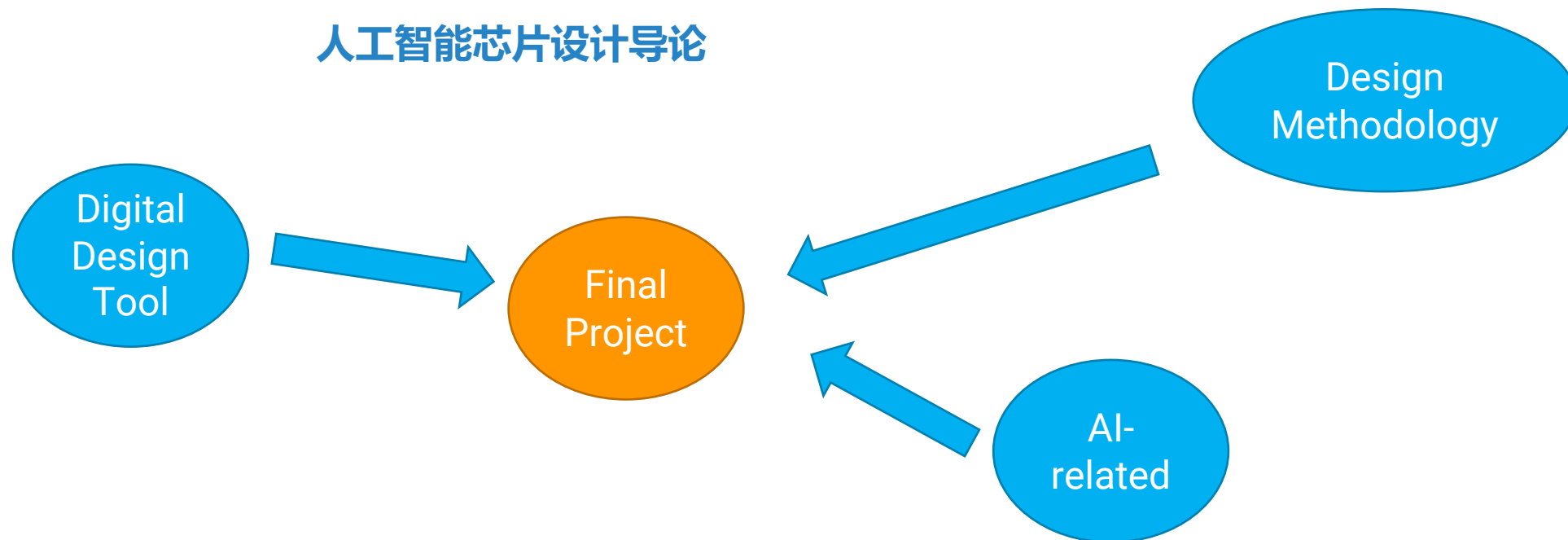
# Duke



## University of Pittsburgh

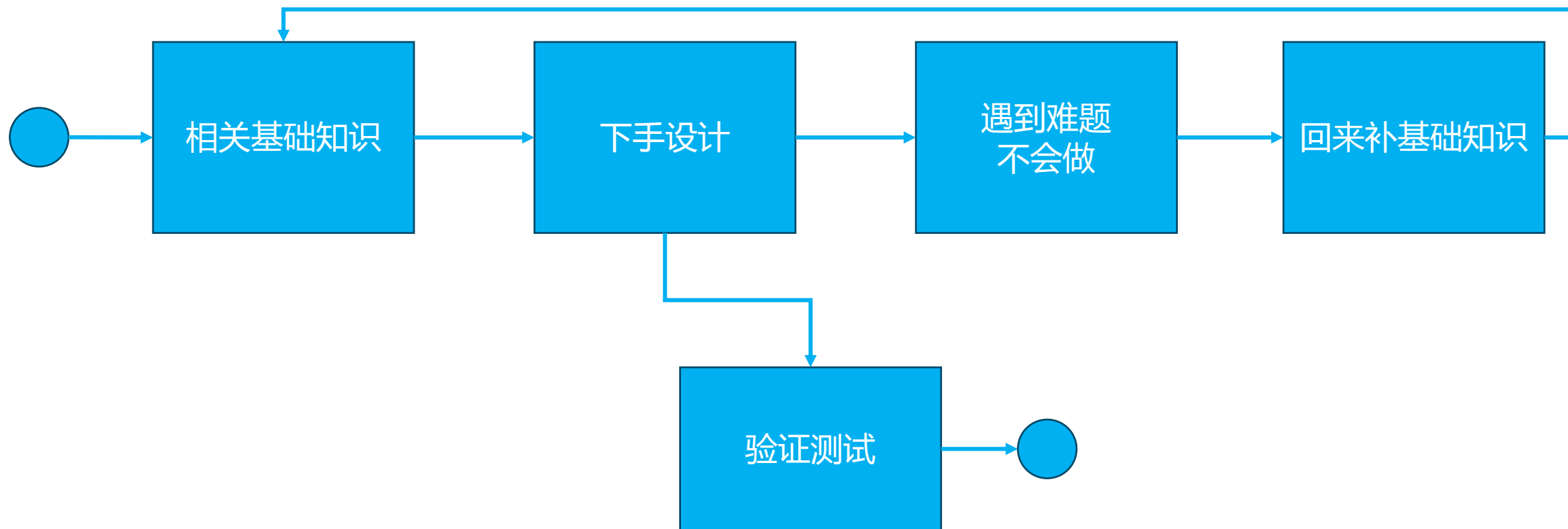
## 课程定位

- “研究项目” 课程
- 模式：



- 手把手带同学们设计一款人工智能芯片（Verilog）
- 课程网站/参考书：稍后提供

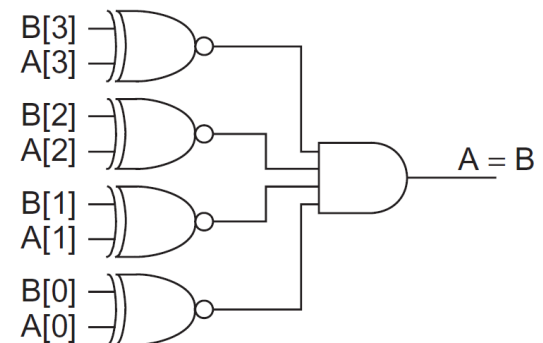
•课程节奏



•研究/工作 “预演”

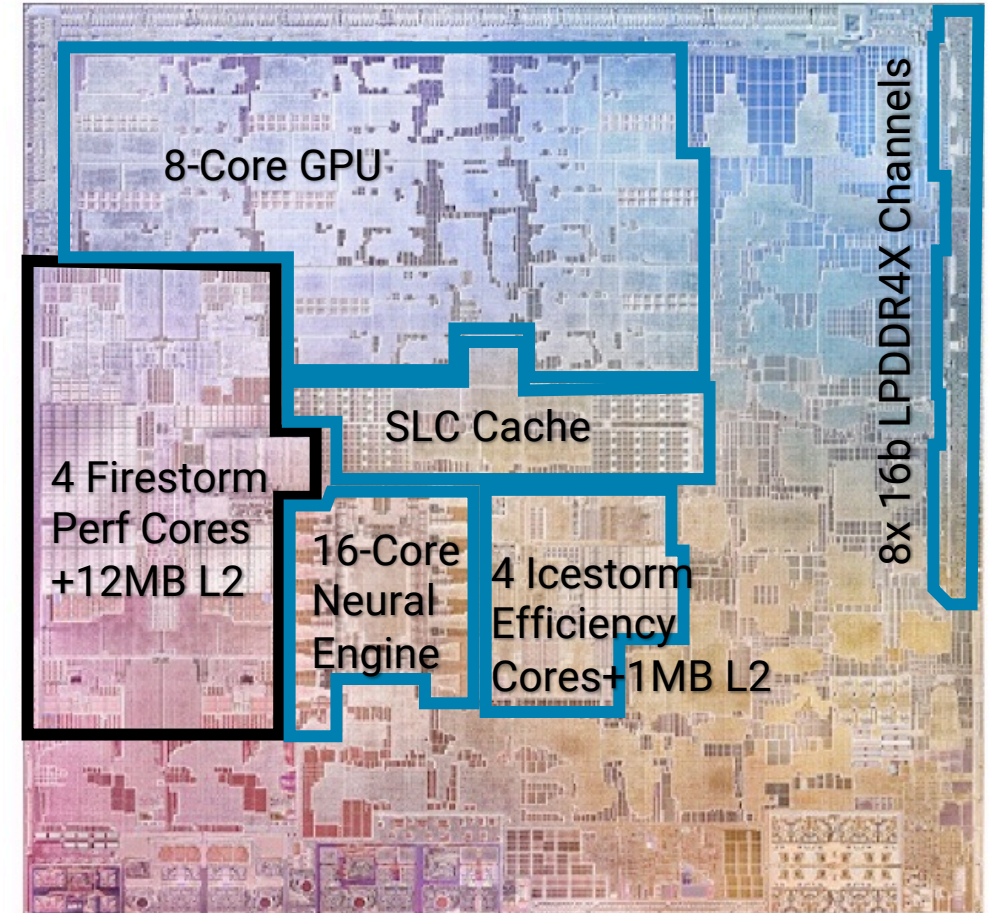
# Content

- Number 1 reason for students to enroll:
- “Gain more experience in AI ASIC design”
  
- Components of AI ASIC design:
- 1. Logic & Transistor Circuits and low-level blocks:
  - how to achieve desired function of low-level chip building blocks
  - state machines and clocking
  - performance/cost/power tradeoffs
  - physical realization concerns (floorplanning, clock distribution, pwr distribution)
  - Provides “**Bottom-up**” knowledge



# Content

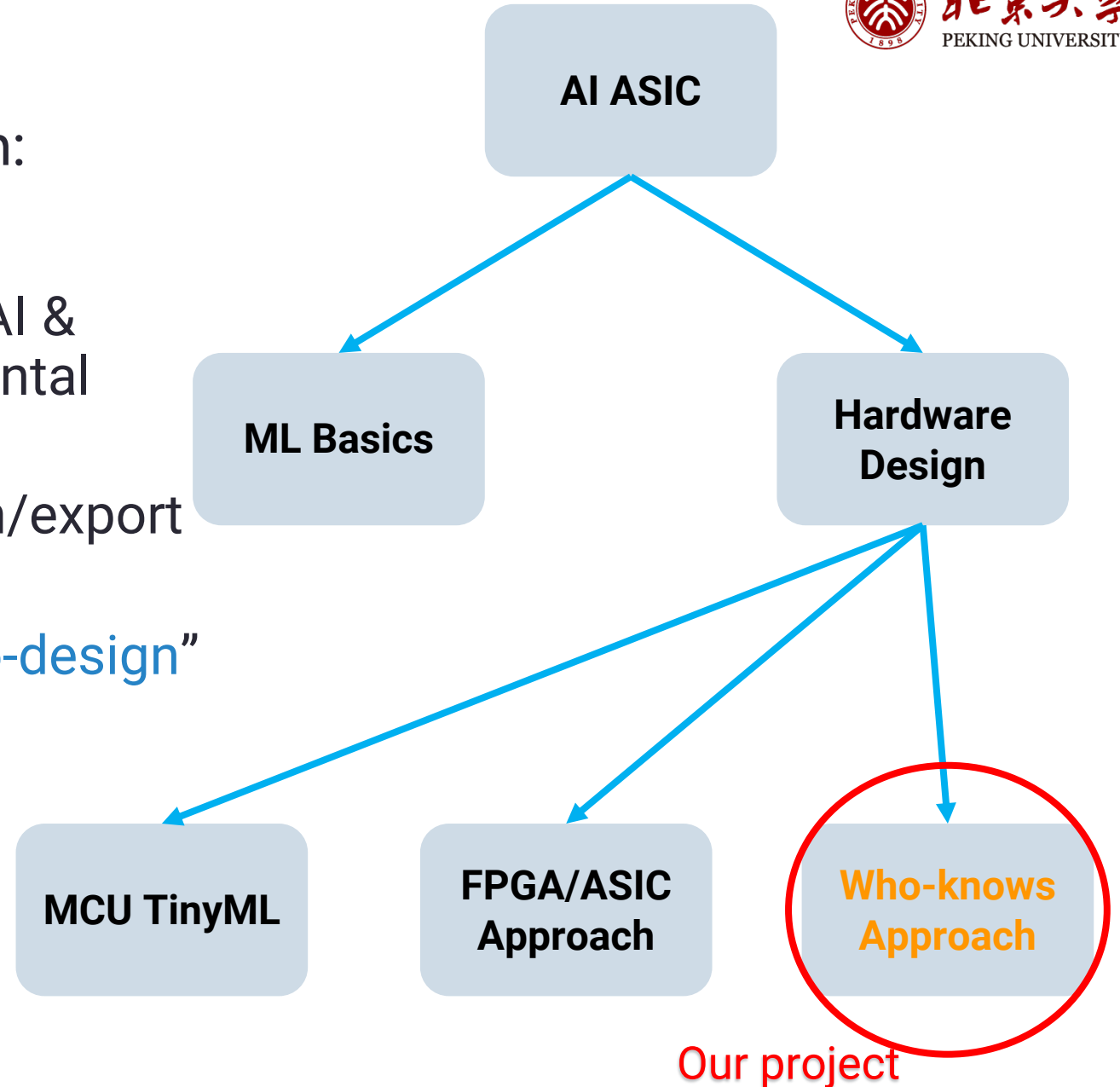
- Components of AI ASIC design:
- 2. Chip Architecture and high-level blocks:
- How building blocks are assembled to achieve high-level functionality
  - The programmable architectures start from a standard “execution model” – ISA
  - Accelerators start from an algorithm or set of algorithms.
- Provides “**Top-down**” knowledge



Apple M1 processor (2020)  
8-core ARM, 16 billion transistors

# Content

- Components of AI ASIC design:
- 3. Going Intelligent
- Instrumentalism perspective of AI & Machine Learning (ML) fundamental knowledge
- Use Pytorch framework to obtain/export models
- Provides “Software-Hardware Co-design” knowledge



# Project: ML加速器设计

## Tensor Processing Unit (2017)

Google发布

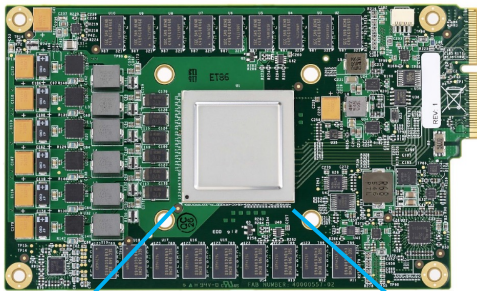


Figure 3. TPU Printed Circuit Board. It can be inserted in the slot for an SATA disk in a server, but the card uses PCIe Gen3 x16.

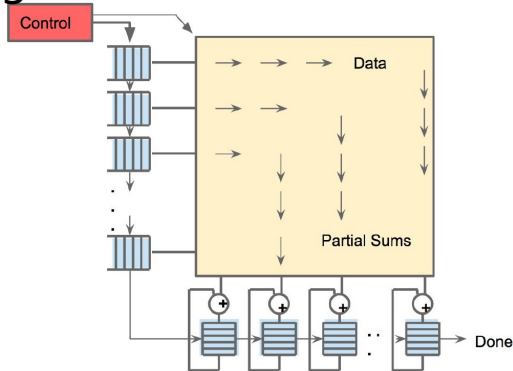
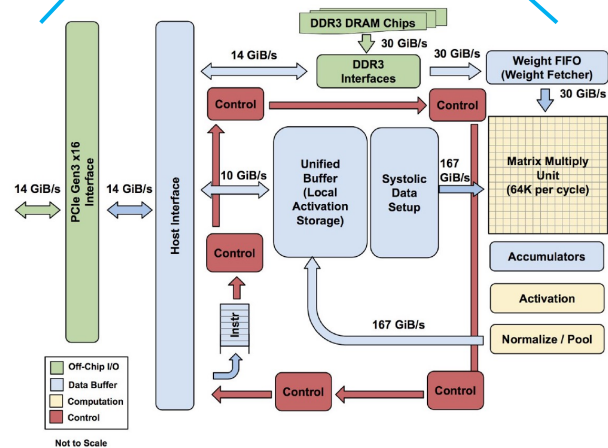


Figure 4. Systolic data flow of the Matrix Multiply Unit. Software has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.



- 仿照TPU结构
- 定制TPU并部署简单算法
- 完成形式:
  - 全班合作

- 二维阵列乘法PE
- 专用数字电路  
安排“输入时序”



# Grading

Assignment	Assignment 1	10%	30%
	Assignment 2	10%	
	Assignment 3	10%	
Presentation	In-Class Update	10%	40%
	Project Update 1	10%	
	Project Update 2	10%	
	Final Presentation	10%	
Paper	Final Paper	30%	30%

Late homework incur penalties as follows:

- Submission is 0-24 hours late: total score is multiplied by 0.9
- Submission is 24-48 hours late: total score is multiplied by 0.8
- Submission is more than 48 hours late: total score is multiplied by the Planck constant (in J·s)

You are **encouraged** to complete the assignments/projects with peers

- ✓ discussion
- ✓ compare your answers

But DO NOT COPY

**抄袭零容忍！ 后果很严重！**

# ■ ■ This course prepares you as an digital designer

- What we assume you already knows:
  - Basic digital logic
  - Know the basics of CMOS technologies
  - Python & C, knows how to code



- You will become (at least):
  - A master of Verilog HDL
  - A rookie of In-Memory Computing
  - A rookie of CPU hardware designer

1. 你的专业？研究方向？
2. 描述你过去的数字电路设计相关的课程（凡是用到数字电路设计）
3. 除上述课程经验外，您是否有数字设计和使用ASIC或FPGA的额外经验工具？
4. 自己有没有训练过神经网络？



- **End of Intro I**

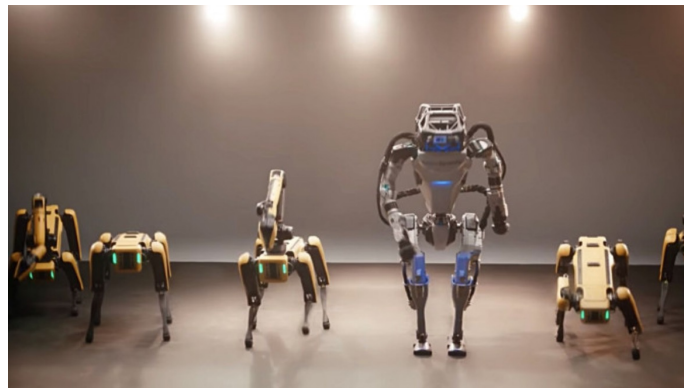
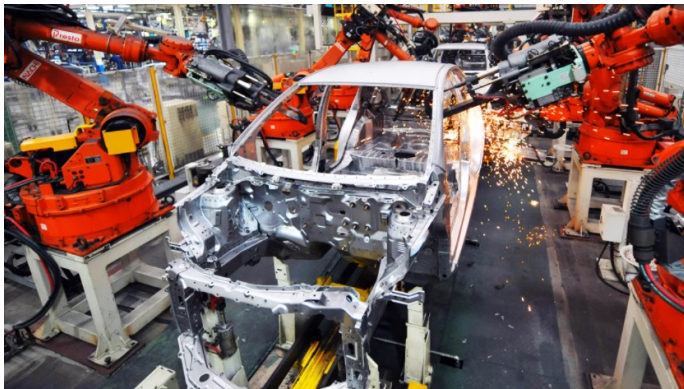
# 定义

**Intelligence** might be defined as the ability to learn and perform suitable techniques to solve problems and achieve goals, appropriate to the context in an uncertain, ever-varying world.

- A fully pre-programmed factory robot is flexible, accurate, and consistent but not intelligent.

**Artificial Intelligence (AI)**, a term coined by emeritus Stanford Professor John McCarthy in 1955, was defined by him as “the science and engineering of making intelligent machines”.

**Machine Learning (ML)** is the part of AI studying how computer agents can improve their perception, knowledge, thinking, or actions based on experience or data.



## 定义



In **supervised learning**, a computer learns to predict human-given labels, such as dog breed based on labeled dog pictures; **unsupervised learning** does not require labels, sometimes making its own prediction tasks such as trying to predict each successive word in a sentence; **reinforcement learning** lets an agent learn action sequences that optimize its total rewards, such as winning games, without explicit examples of good techniques, enabling autonomy

**Deep Learning** is the use of large multi-layer (artificial) neural networks that compute with continuous (real number) representations, a little like the hierarchically organized neurons in human brains.

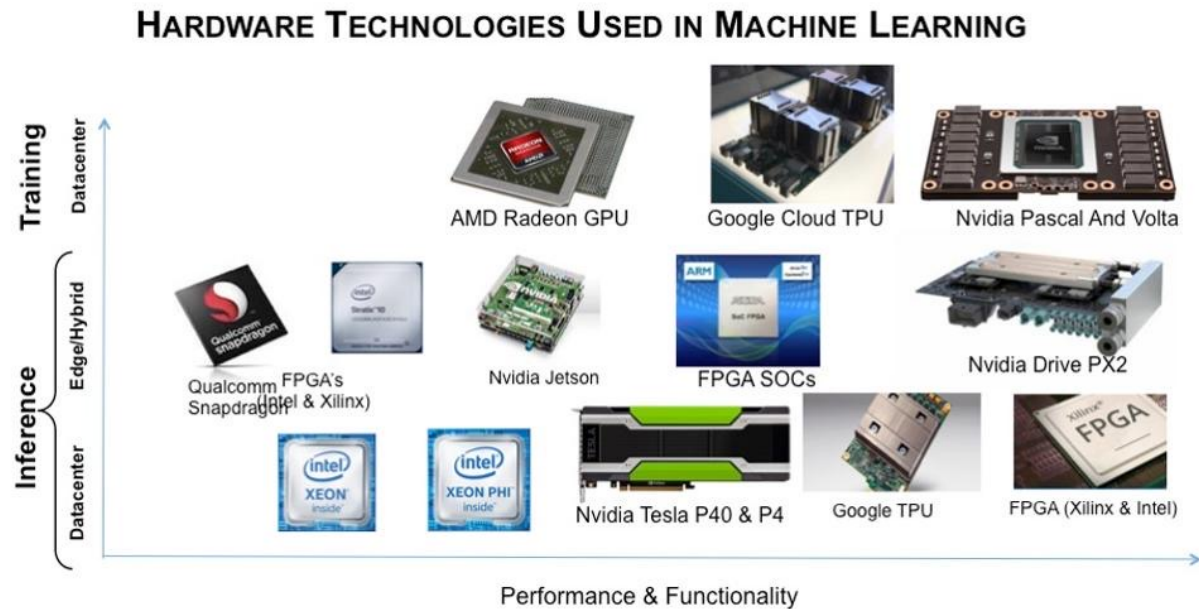
- It is currently the most successful ML approach, usable for all types of ML, with better generalization from small data and better scaling to big data and compute budgets.

**Narrow AI** is intelligent systems for one particular thing, e.g., speech or facial recognition.

**Human-level AI**, or **Artificial General Intelligence (AGI)**, seeks broadly intelligent, context-aware machines. It is needed for effective social chatbots or human-robot interaction.

# 为什么要有“人工智能芯片”？

- 硬件平台多种多样



Power?



Area? Volume?



Performance?



Application-Specific Integrated Circuits (ASIC)

- 硬件永远不够用!



## Specs & Definition

- Energy Efficiency/Power Efficiency:
  - Unit: Op/J [operations per Joule] ~ TOPS/W
  - Unit: OPS/W [operations per second per watt] ~ TOPS/W
  - Throughput/Power
  - Peak/Average/Sparse
- Examples:
  - Processor A does INT8 Add, 1k times/second, power: 1mW, what is the energy efficiency?
  - Processor B does FP64 Multiply, 100 times/second, power: 1mW, what is the energy efficiency?

**FLOPS/W**

# Example

## Technical Specifications

	Jetson AGX Xavier Series	
	AGX Xavier	AGX Xavier Industrial
<b>AI Performance</b>	32 TOPS	30 TOPS
<b>GPU</b>	NVIDIA Volta architecture with 512 NVIDIA CUDA cores and 64 Tensor cores	
<b>CPU</b>	8-core NVIDIA Carmel Armv8.2 64-bit CPU 8MB L2 + 4MB L3	
<b>DL Accelerator</b>	2x NVDLA	
<b>Vision Accelerator</b>	2x 7-Way VLIW Vision Processor	
<b>Safety Cluster Engine</b>	-	2x Arm Cortex-R5 in lockstep
<b>Memory</b>	32GB 256-bit LPDDR4x 136.5GB/s	32GB 256-bit LPDDR4x (ECC support) 136.5GB/s
<b>Storage</b>	32GB eMMC 5.1	64GB eMMC 5.1

What is the Jetson AGX Xavier's energy efficiency?



<b>UPHY</b>	8x PCIe Gen4   8x SLVS-EC 3x USB 3.1 Single Lane UFS	8x PCIe Gen4 3x USB 3.1 Single Lane UFS
<b>Power</b>	10W   15W   30W	20W   40W
<b>Networking</b>	10/100/1000 BASE-T Ethernet	
<b>Display</b>	Three multi-mode DP 1.2a/e DP 1.4/HDMI 2.0 a/b	
<b>Other I/O</b>	USB 2.0 UART, SPI, CAN, I2C, I2S, DMIC & DSPK, GPIOs	
<b>Mechanical</b>	100mm x 87mm 699-pin connector Integrated Thermal Transfer Plate	

- Area Efficiency:

- Unit: OPS/mm<sup>2</sup> [operations per second per mm<sup>2</sup>]
- Throughput/Area
- Peak/Average/Sparse..

Processor A does INT8 Add, 1k times/second, area: 10mm<sup>2</sup>, what is the area efficiency?

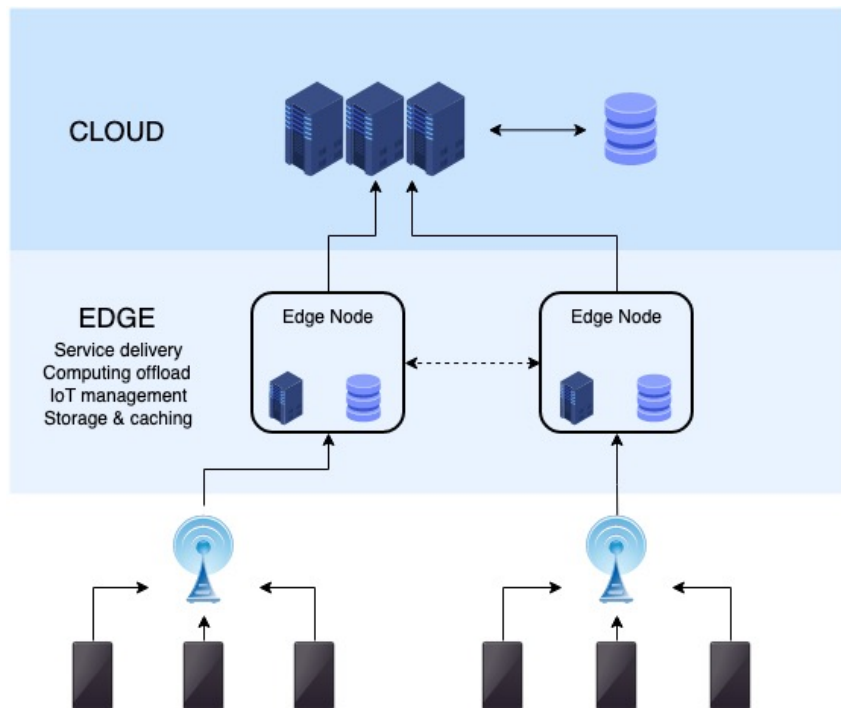
- Memory Density:

- Unit: bit/mm<sup>2</sup> [bit per mm<sup>2</sup>]
- Storage Capacity/Area

Memory A has 1Kb, area: 10mm<sup>2</sup>, what is the density?

# ■ ■ 训练? 推理? 云? 边缘?

- 算力层次



算力

规模/功耗

高

高

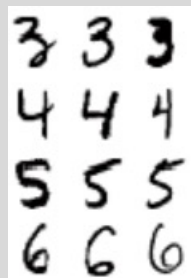
**Training: HPC**  
**Training**  
**Inference: Datacenter**  
**Inference: Edge**  
**Inference: Mobile**  
**Inference: Tiny (TinyML)**

低

低

# 常见任务与数据集

**MNIST:**  
Handwritten  
Datasets



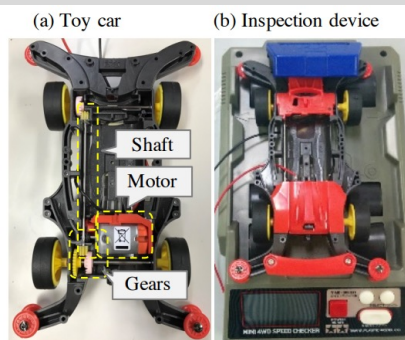
**SVHN:** Street View House Number



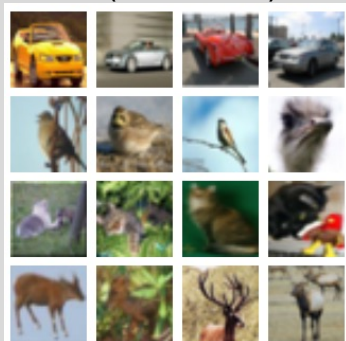
**COCO:** Object detection dataset



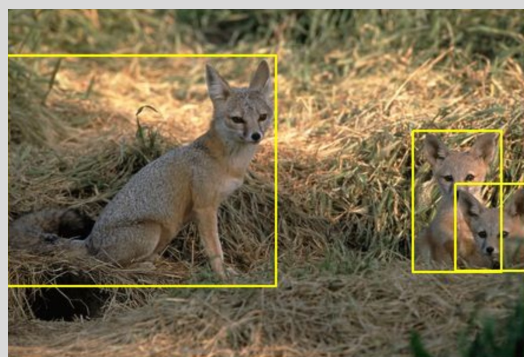
**ToyADMOS:** machine  
operating sounds dataset of  
normal machine operating



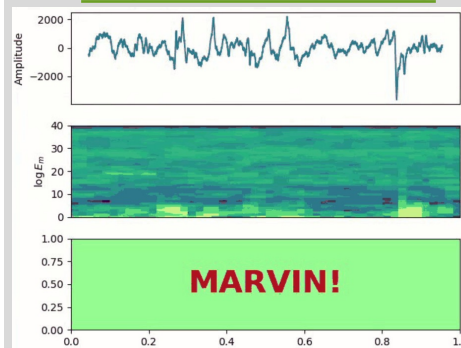
**CIFAR:** Image Classification  
with 100-class (CIFAR100) /  
10-class (CIFAR10)



**ImageNet:** Image Classification



**Key Words Spotting**



**Visual Wake Words**

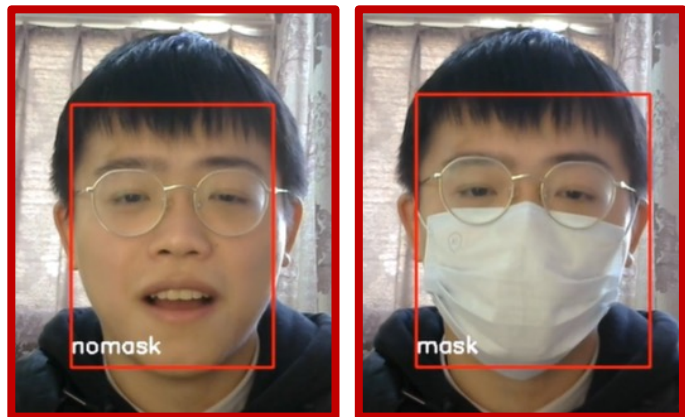


(a) 'Person'



(b) 'Not-person'

# 智能照进生活



检测戴不戴口罩



“自”行车



机载烟花



检测是不是在打瞌睡



跟着人的伞



跟拍的无人机



自动驾驶玩具车

# 几个定义

- Hardcore IP: 硬核IP
  - 固定的设计, 下游开发人员不能改变的功能块
- Softcore IP: 软核
  - 用Verilog等硬件描述语言描述的功能块
- System-on-a-chip (SoC): 片上系统
  - 单个芯片上集成一个完整的系统, 一般包含
    - CPU、GPU、NPU...,
    - 总线
    - 片上存储
    - GPIO、对外的接口
- ASIC: Application-Specific Integrated Circuits 专用集成电路

SoC Example:



## • 为什么用加速器?

- Domain-Specific Accelerator
- 提升算力
- 提升效率
- 相对降低成本

## • 为什么用FPGA?

- 可重构
- 快速开发
- 原型设计、硬件模拟

## • 为什么用ASIC?

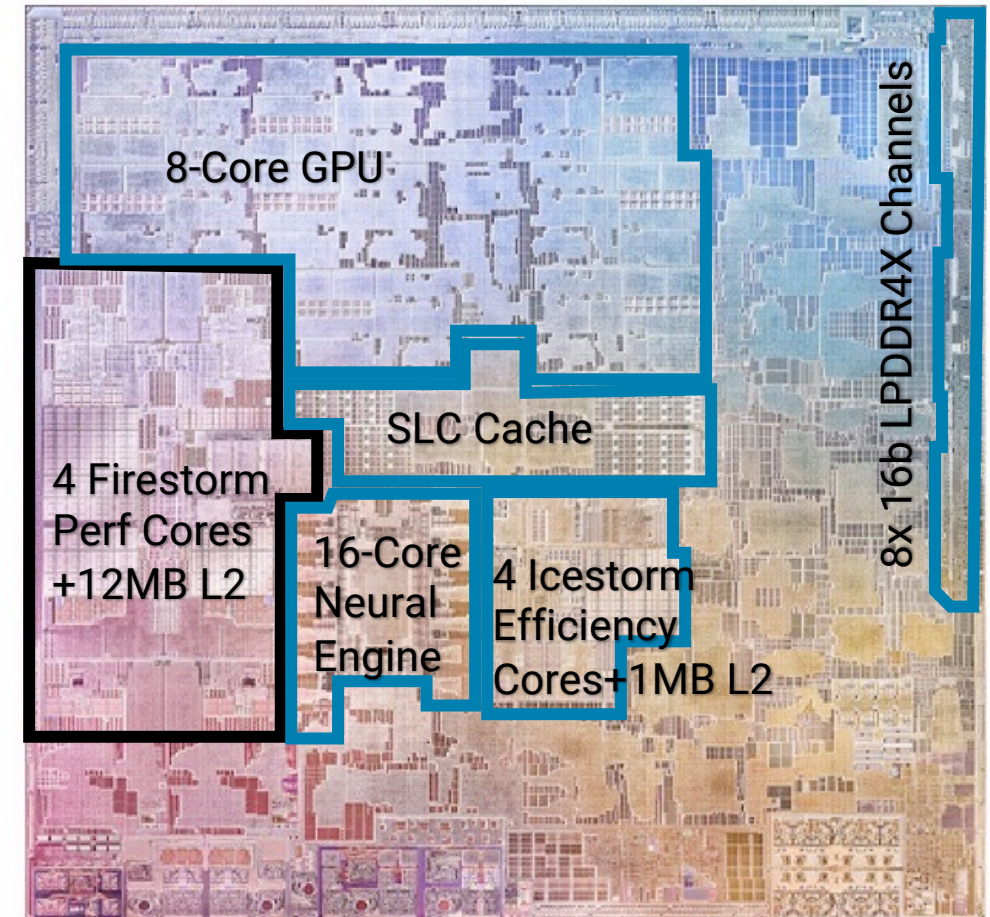
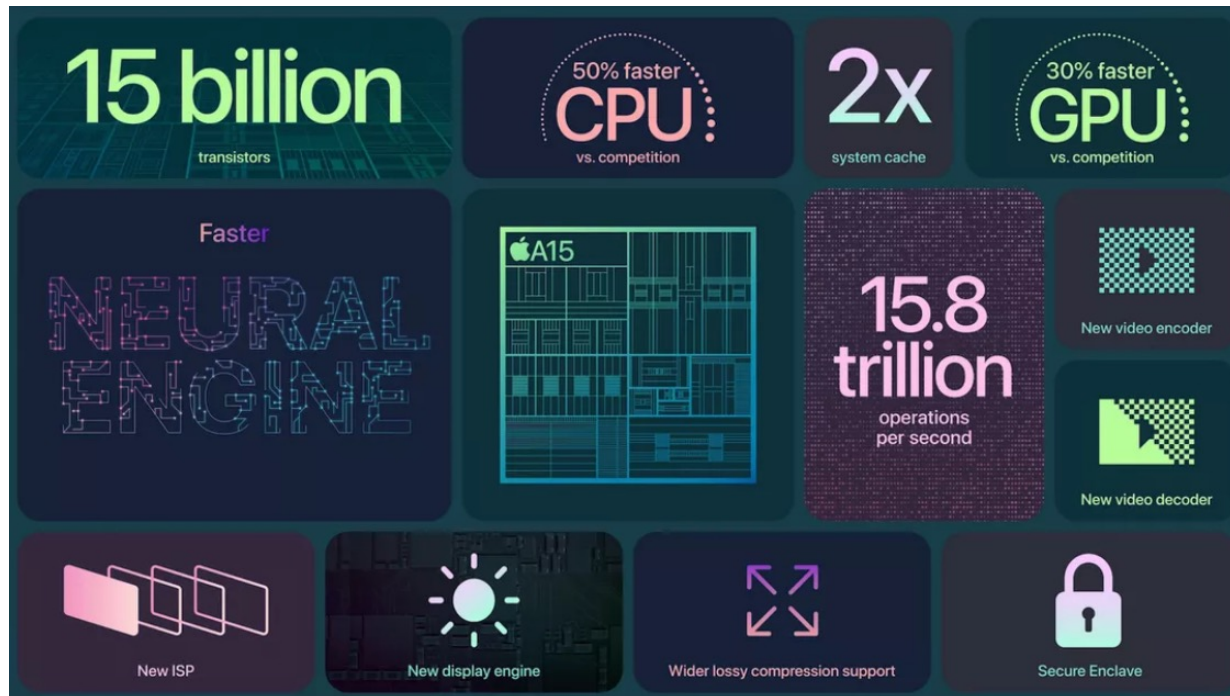
- 从最底层(Gate、Transistor)开始优化
- 灵活度高, 完全符合应用需求



# FPGA/ASIC路线的哲学问题

## Heterogenous Computing SoC

- Hardware accelerators
- Co-processors
- Tons of on-chip memories



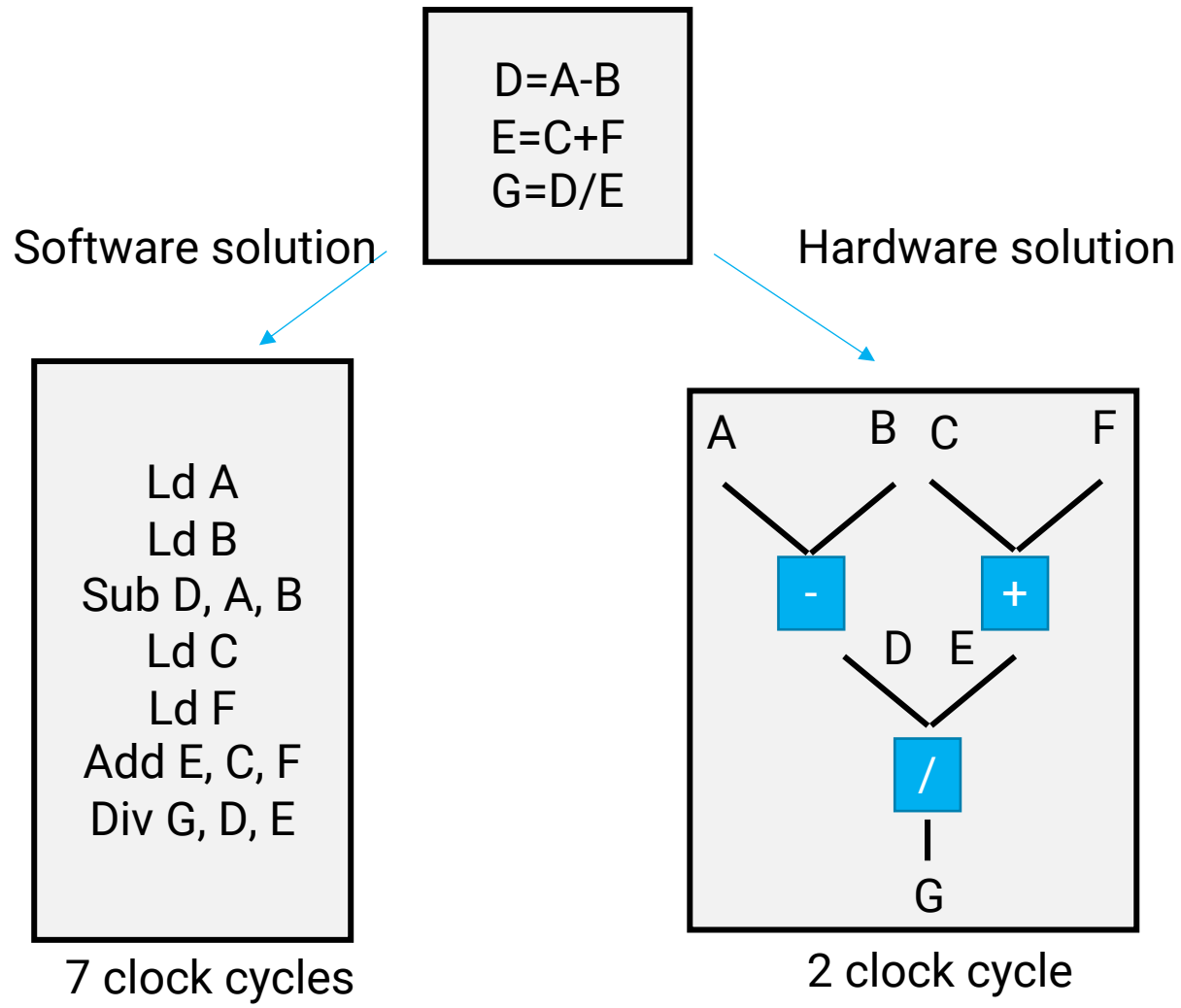
Apple M1 processor (2020)  
8-core ARM, 16 billion transistors

# 如果你是一家SoC的架构设计师，你需要考虑...



## OPs/\$ or OPs/Joule

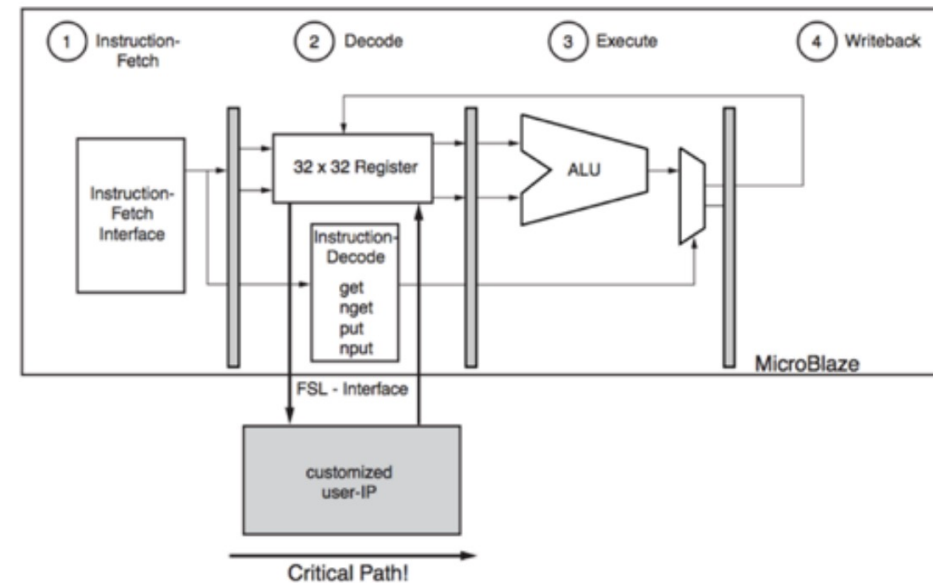
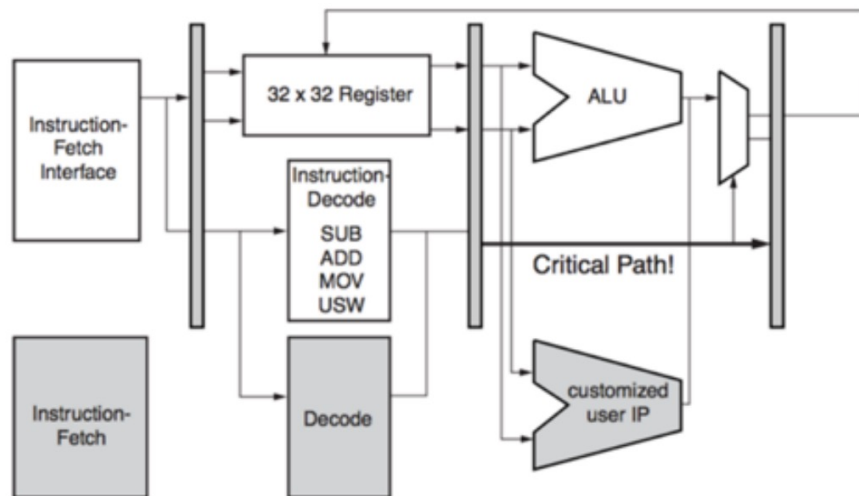
- Exploit problem specific parallelism, at thread and instructions level
- Custom operational units or “instructions” match the set of operations needed for the algorithm (replace multiple instructions with one), custom word width arithmetic, etc.
- Remove overhead of instruction storage and fetch, ALU multiplexing



# 紧密连接 Tightly Coupled

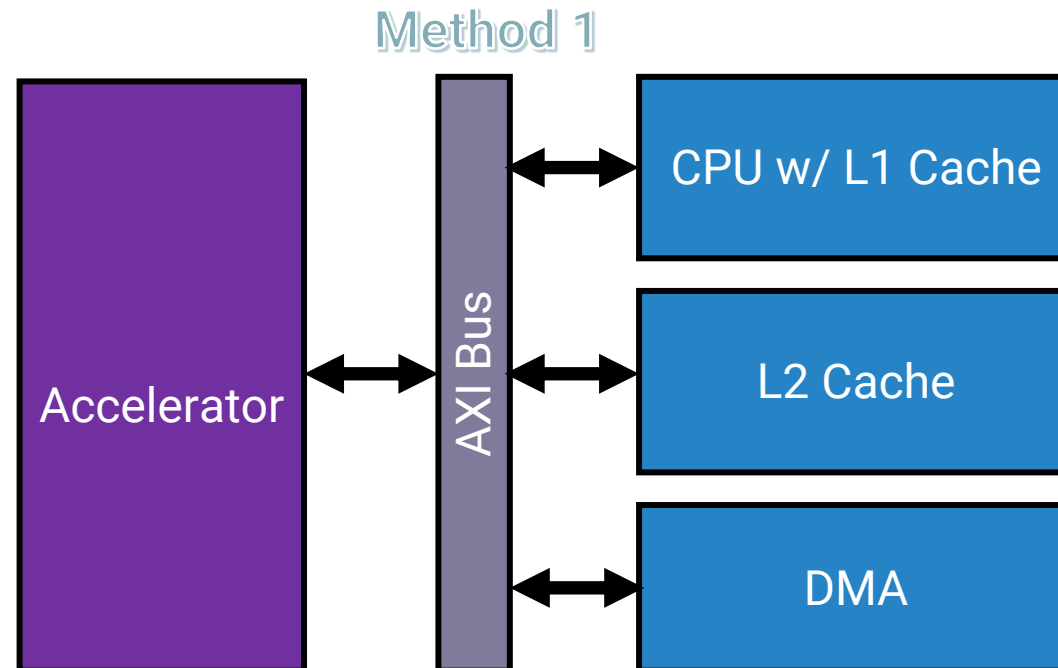
## Integrated with processor control logic

- Task typically completes in a few cycles – Small amounts of data
- Processor stalls waiting for the coprocessor
- Communication with coprocessor typically via registers and dedicated control signal



## Loosely-Coupled Co-processors

- Used for larger tasks than is the case for tightly-coupled coprocessors
- Task runs in parallel with main processor
- May take many cycles per task
- Large amounts of data that coprocessor may access independent of main processor – May or may not use the standard coprocessor interface



# Efficiency Calculation Example 1

	ISSCC'18 [1]	ISSCC'18 [2]	ISSCC'19 [3]	ESSCIRC'19 [4]	ISSCC'20 [5]	ISSCC'20 [6]	This work
Technology	65nm	65nm	55nm	65nm	7nm	28nm	22nm
MAC operation	Analog	Analog	Analog	Digital	Analog	Analog	Digital
Array Size	4Kb	16Kb	3.8Kb	16Kb	4Kb	64Kb	64Kb
Cell Type	S6T	10T	T8T	6T	8T	6T	6T
Push rule	Yes	No	Yes	NA	Yes	NA	No
Macro size (mm <sup>2</sup> )	NA	0.067	NA	0.2272	0.0032	NA	0.202
Bitcell Area (um <sup>2</sup> )	0.525	NA	0.865	NA	0.053	0.25	0.379
Power Supply(V)	1&0.8	1.2&0.9	1	0.6~0.8	0.8	0.7~0.9	0.72
Inputs Bits	1	7	4	1~16	4	4~8	1~8
Weight bits	1	1	5	4/8/12/16	4	4/8	4/8/12/16
Output Bits	1	7	7	8~23	4	12 (4b/4b) 20 (8b/8b)	16 (4b/4b) 24 (8b/8b)
Cycle time (ns)	2.3	150	10.2	NA	5.5	4.1 (4b/4b) 8.4 (8b/8b)	10 (4b/4b) 18* (8b/8b)
Throughput (GOPs)	1780	10.67	17.6	567 (1b/1b)	372.4 (4b/4b)	124.88 (4b/4b) 30.48 (8b/8b)	3300 (4b/4b) 917* (8b/8b)
Energy Efficiency (TOPS/W)	55.6	28.1	18.4	117.3 (1b/1b)	262.3~610.5 (4b/4b)	68.44 (4b/4b) 16.63 (8b/8b)	89 (4b/4b) 24.7* (8b/8b)

3300 (4b/4b)
917* (8b/8b)
89 (4b/4b)
24.7* (8b/8b)

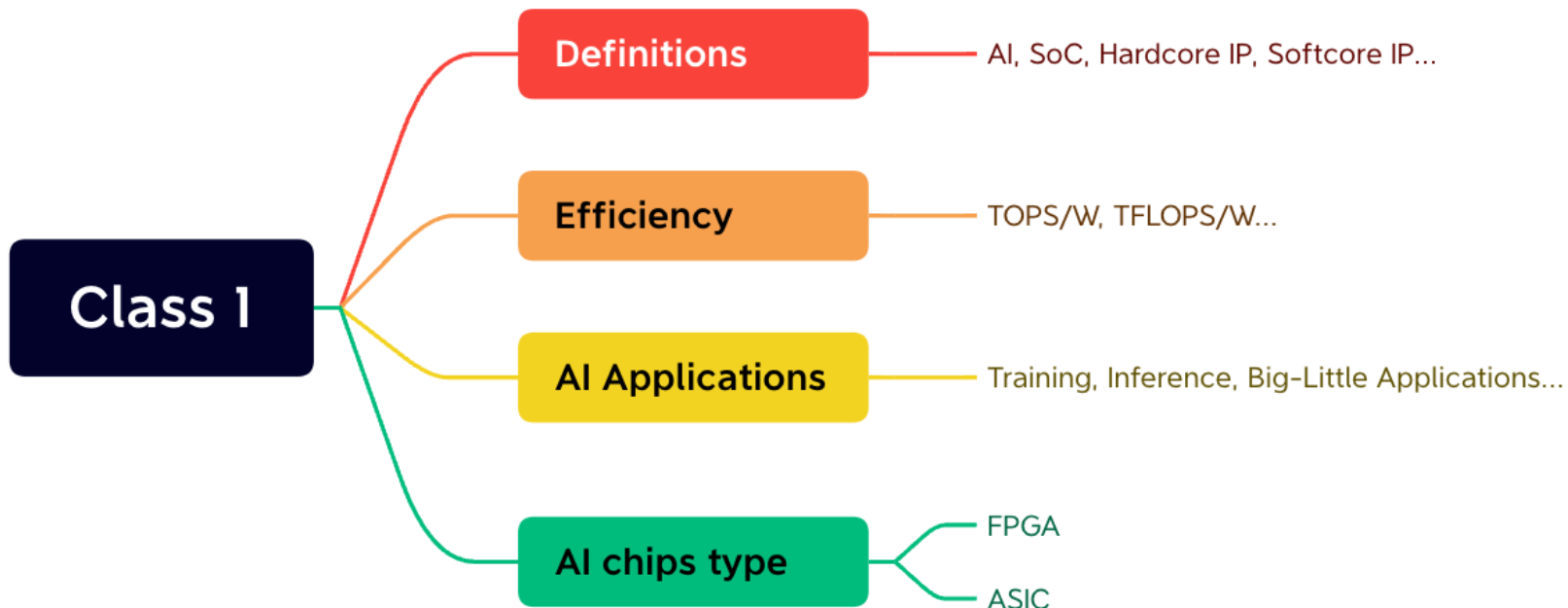


\*estimation

## Efficiency Calculation Example 2

Technology	7nm
Array Size	4kb
Macro Area (mm <sup>2</sup> ) *	0.0032
Input/Weight/Output Precision	4 / 4 / 4
Voltage Range (V)	0.65 ~ 1
Cycle Time @ 0.8V (ns)	5.5
Max Power @ 0.8V (mW)	1.42
Max Energy @ 0.8V (pJ)	7.8
Throughput (GOPS)	372.4
Energy Efficiency (TOPS/W)	262.3 ~ 610.5 351 in average

\* Including testing & reconfigurable blocks





- The End