北京大学
PEKING UNIVERSITY

4803019

人工智能芯片：设计流程与实践

AI ASIC: Design and Practice (ADaP)

Fall 2023

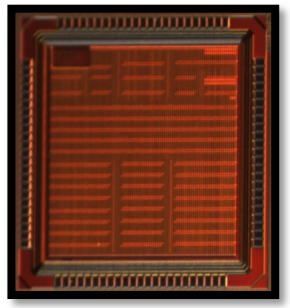
燕博南

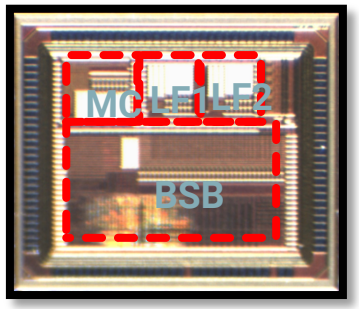
课程教授：燕博南，Ph.D.

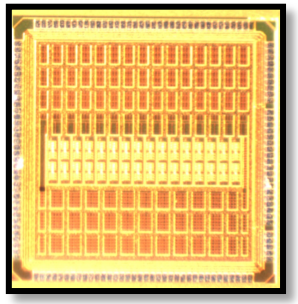
北京大学 PEKING UNIVERSITY

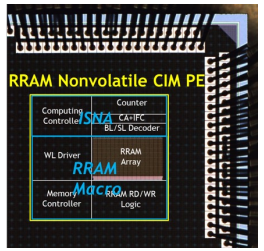
- 2020-Now：Work as Peking University
- Education:
  - PhD Duke University, 2020
- Research:
  - In-Memory Computing Circuits & Systems
  - Domain-Specific Accelerator Chips
  - Emergin Artificial Intelligence Processor
- Bilibili: Dr燕同学

What I am working on









Where I worked

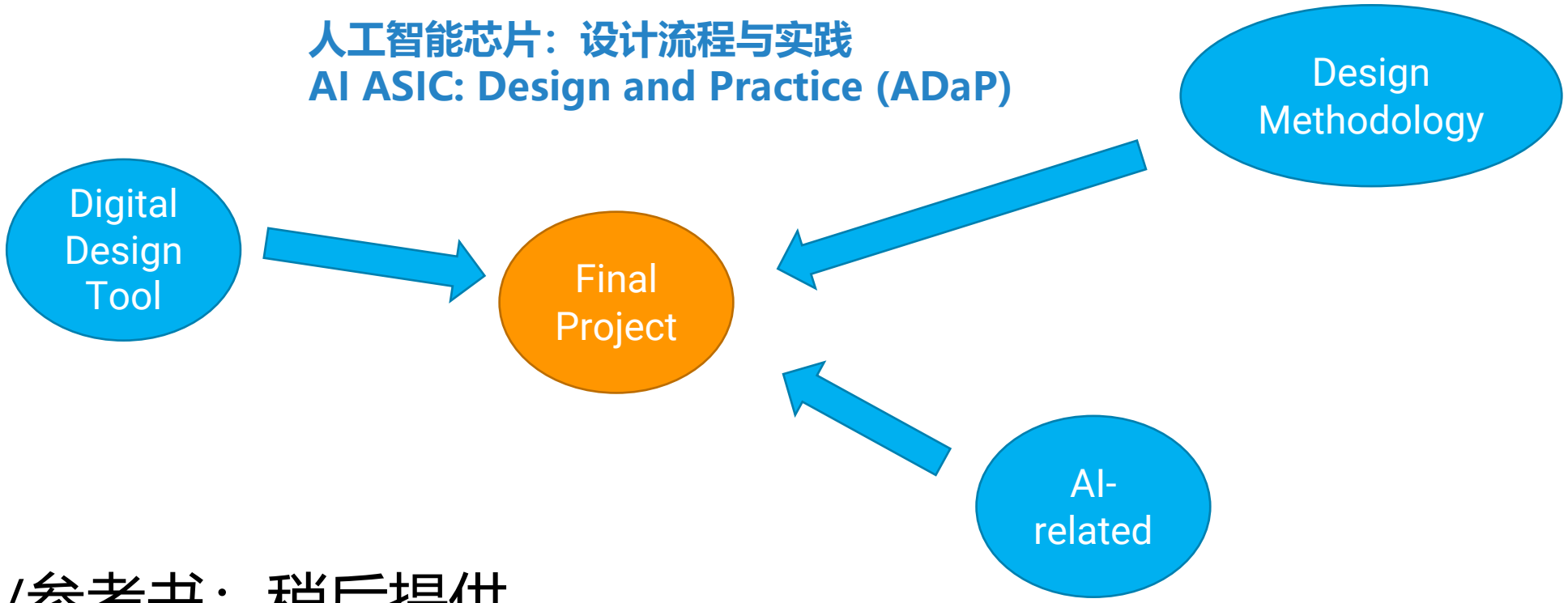
Duke

LABS hp

University of Pittsburgh

课程定位

- “研究项目”课程
- 模式：

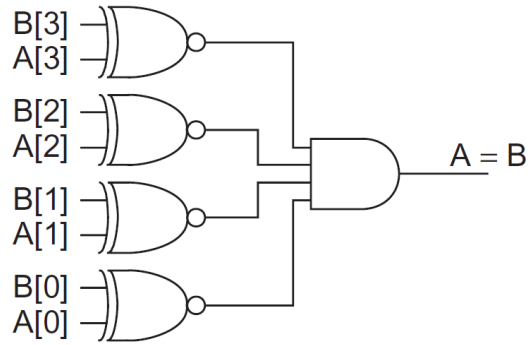


- 课程网站/参考书：稍后提供

Context of ADaP

- Number 1 reason for students to enroll in ADaP:
- “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



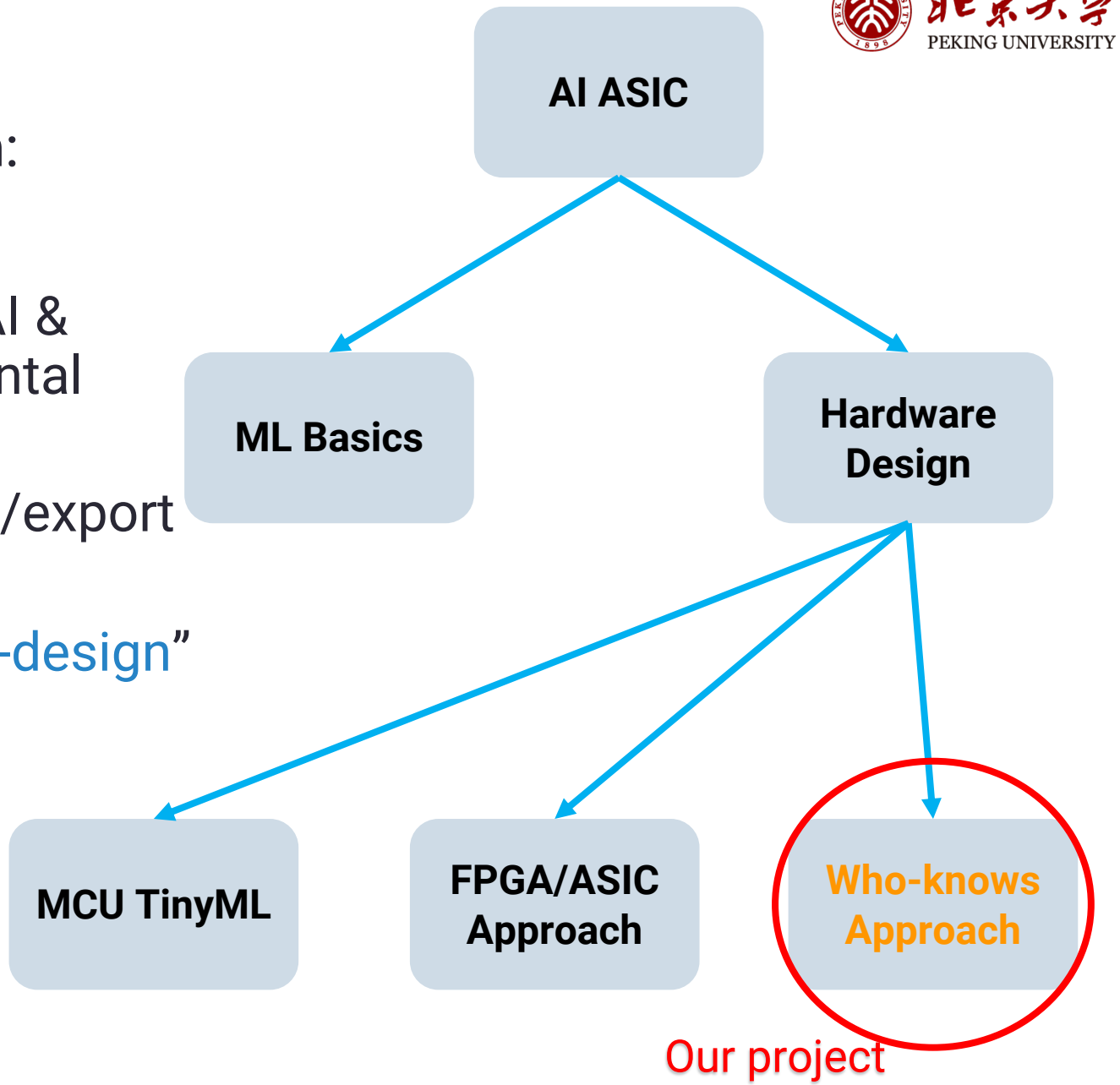
Context of ADaP

- 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

- **19 billion transistors**
- **CPU: 6-core CPU, 2 high-performance cores, and 4 high-efficiency cores**
- **GPU: 6-core with support for ray tracing**
- **Neural Engine: 16-core, 35 trillion operations per second**
- **Technology Node: 3nm**

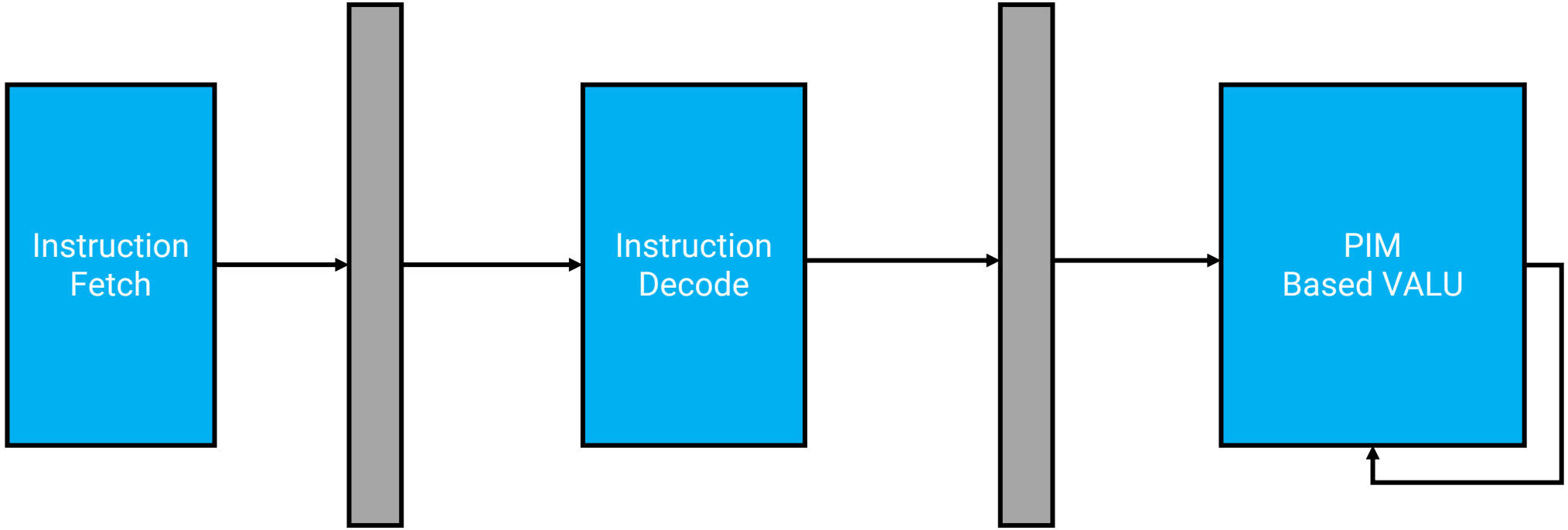
Context of ADaP

- 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: General Purpose VIP

- Group of 2
- Vector In-Memory Processor (VIP)



Vs conventional vector processor:
Combing register & ALU >> PIM-Based VALU

Grading

北京大学 PEKING UNIVERSITY

Assignment	Assignment 1	10%	20%
	Assignment 2	10%	
Presentation	In-Class Update	5%	40%
	Project Update 1	5%	
	Project Update 2	5%	
	Final Presentation	25%	
Paper	Final Paper	40%	40%

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)

评价方式：互评/相互打分

Integrity

北京大学 PEKING UNIVERSITY

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

北京大学 PEKING UNIVERSITY

- 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

背景调查+自我介绍

北京大学 PEKING UNIVERSITY

1. 简要自我介绍
2. 你的专业？研究方向？
3. 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）
4. 除上述课程经验外，您是否有数字设计和使用ASIC或FPGA相关工具的额外经验？
5. 自己有没有训练过神经网络？

北京大學
PEKING UNIVERSITY

- **End of Intro I**

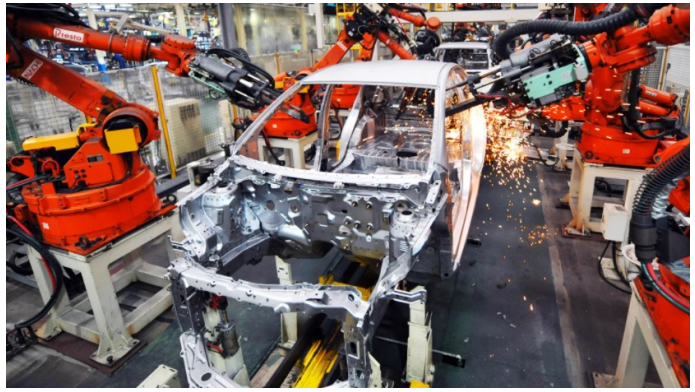
定义

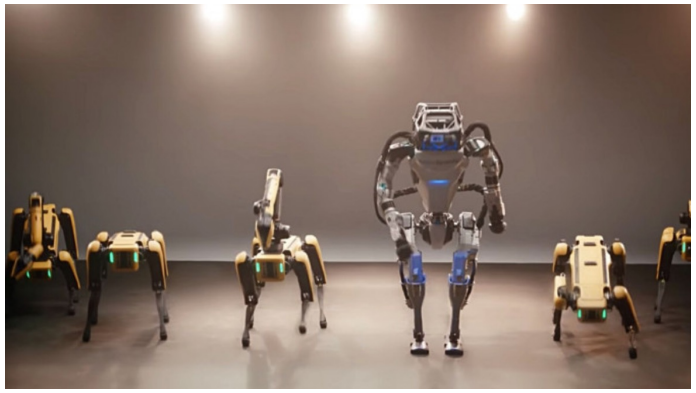
Intelligence might be defined as the ability to learn and perform suitable techniques to solve problems and achieve goals, appropriate to the context in an uncertain, ever-varying world.

- 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.





AI-Definitions-HAI.pdf (stanford.edu)

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.

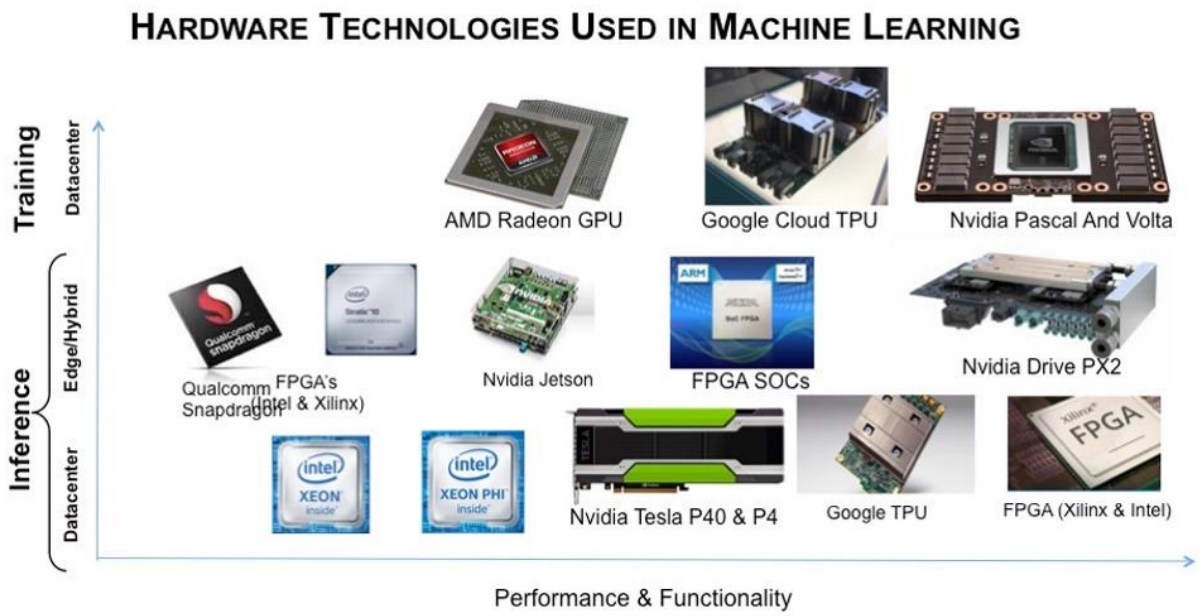
- It is currently the most successful ML approach, usable for all types of ML, with better generalization from small data and better scaling to big data and compute budgets.

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.

AI-Definitions-HAI.pdf (stanford.edu)

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

北京大学 PEKING UNIVERSITY

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

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

What is the Jetson AGX Xavier's energy efficiency?

Specs & Definition

- Area Efficiency:
  - Unit: OPS/mm^2 [operations per second per mm^2]
  - Throughput/Area
  - Peak/Average/Sparse..

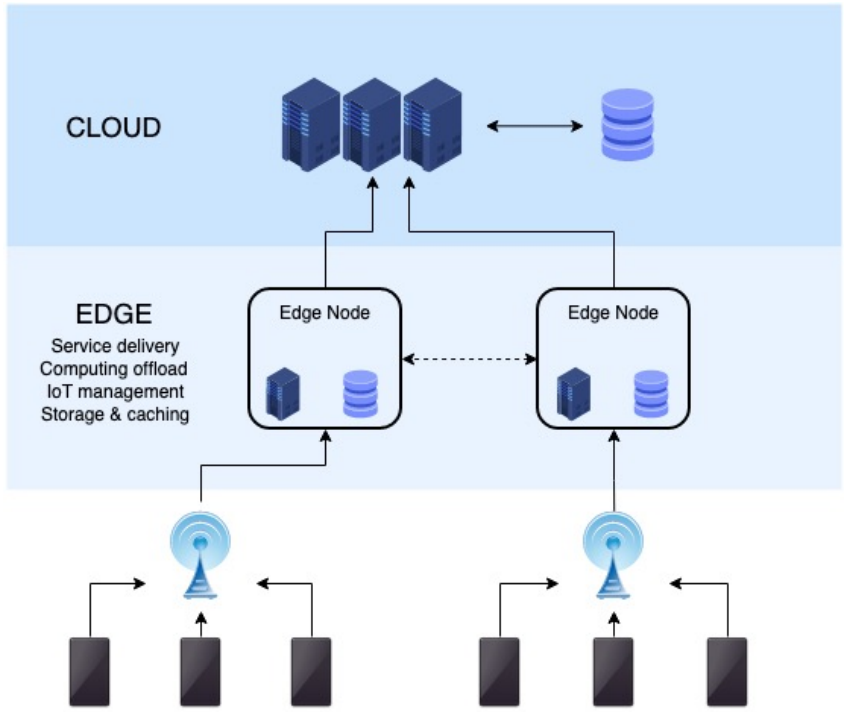
Processor A does INT8 Add, 1k times/second, area: 10mm^2, what is the area efficiency?

- Memory Density:
  - Unit: bit/mm^2 [bit per mm^2]
  - Storage Capacity/Area

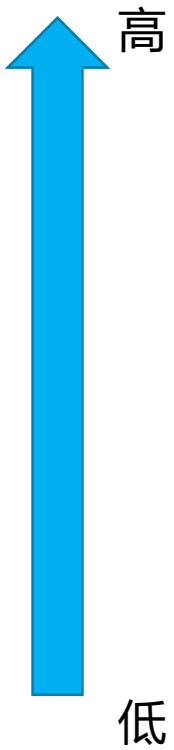
Memory A has 1Kb, area: 10mm^2, what is the density?

训练？推理？云？边缘？

- 算力层次



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

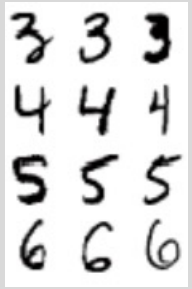


Ref. MLPerF

常见任务与数据集

北京大学 PEKING UNIVERSITY

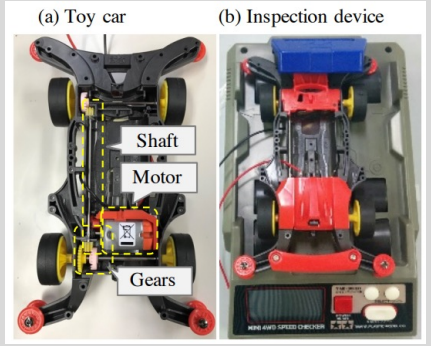
MNIST: Handwritten Datasets



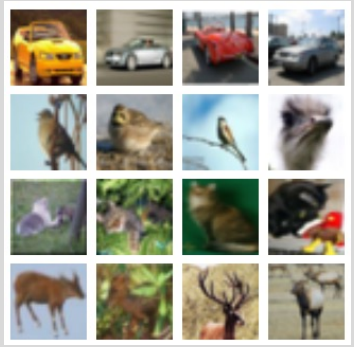
SVHN: Streat View House Number

COCO: Object detection dataset

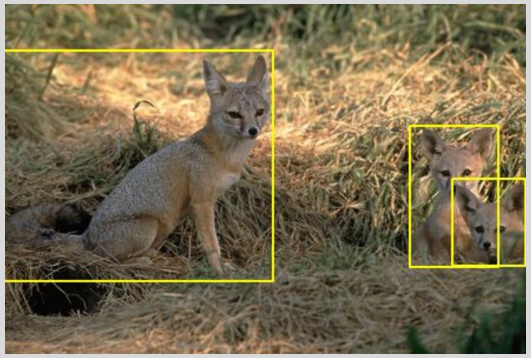
ToyADMOS: machine operating sounds dataset of normal machine operating



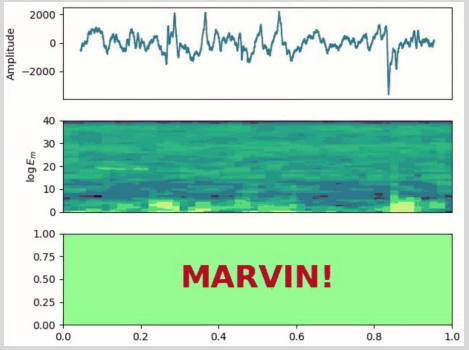
CIFAR: Image Classification with 100-class (CIFAR100) / 10-class (CIFAR10)



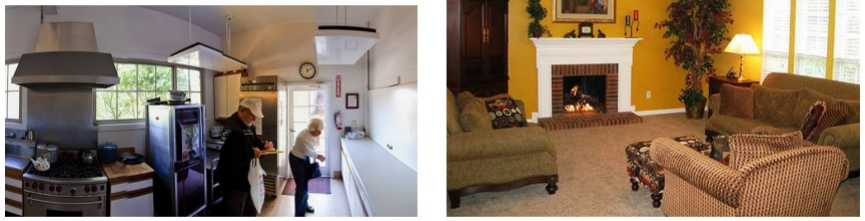
ImageNet: Image Classification



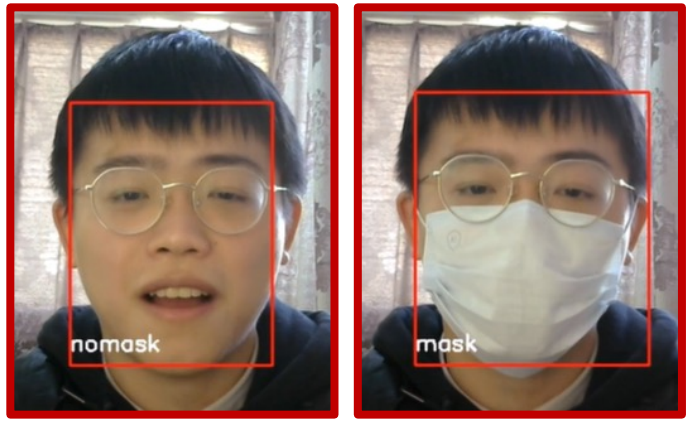
Key Words Spotting



Visual Wake Words



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

Source: Bilibili

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:

FPGA/ASIC路线的哲学问题

北京大学 PEKING UNIVERSITY

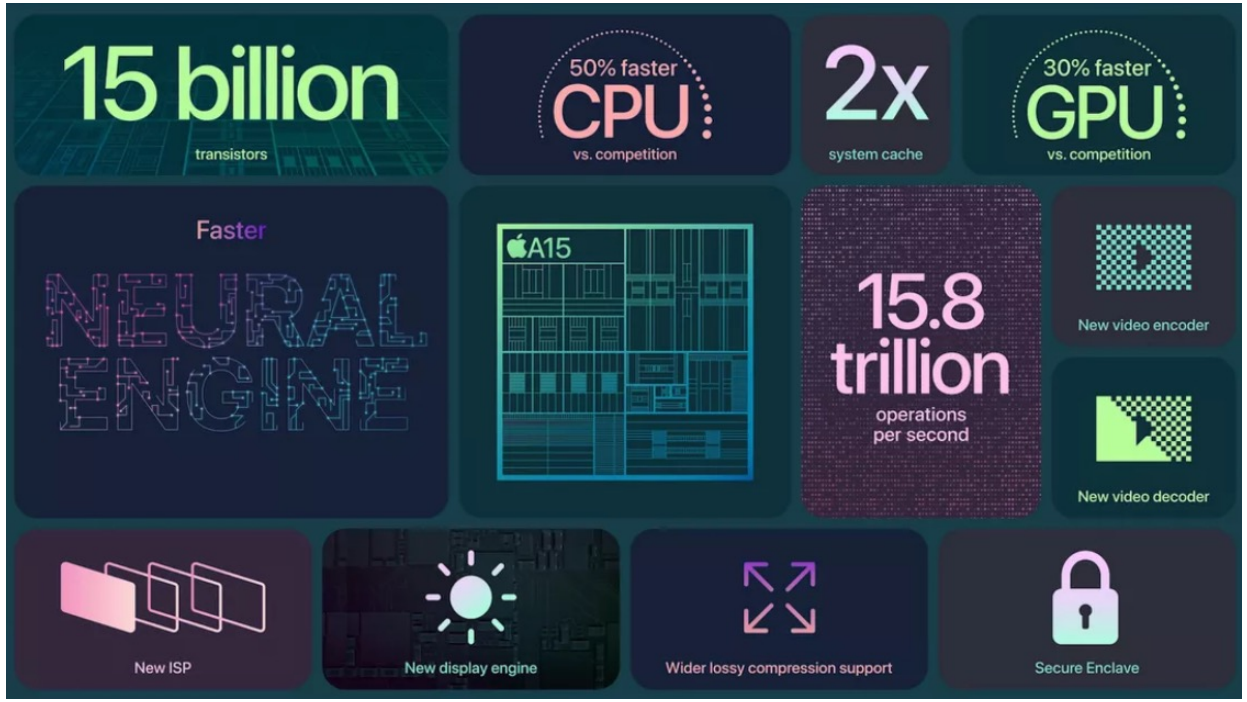
- 为什么用加速器?
  - Domain-Specific Accelerator
  - 提升算力
  - 提升效率
  - 相对降低成本
- 为什么用FPGA?
  - 可重构
  - 快速开发
  - 原型设计、硬件模拟
- 为什么用ASIC?
  - 从最底层(Gate、Transistor)开始优化
  - 灵活度高，完全符合应用需求

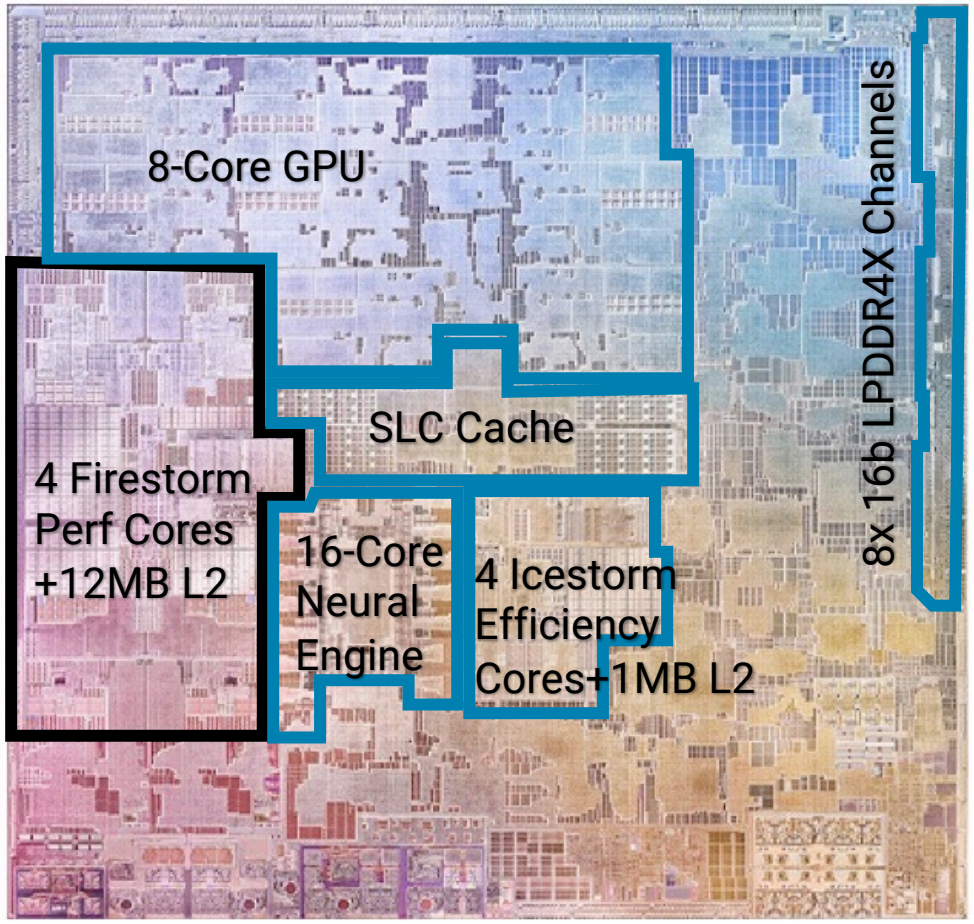
FPGA/ASIC路线的哲学问题

北京大学 PEKING UNIVERSITY

Heterogenous Computing SoC

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



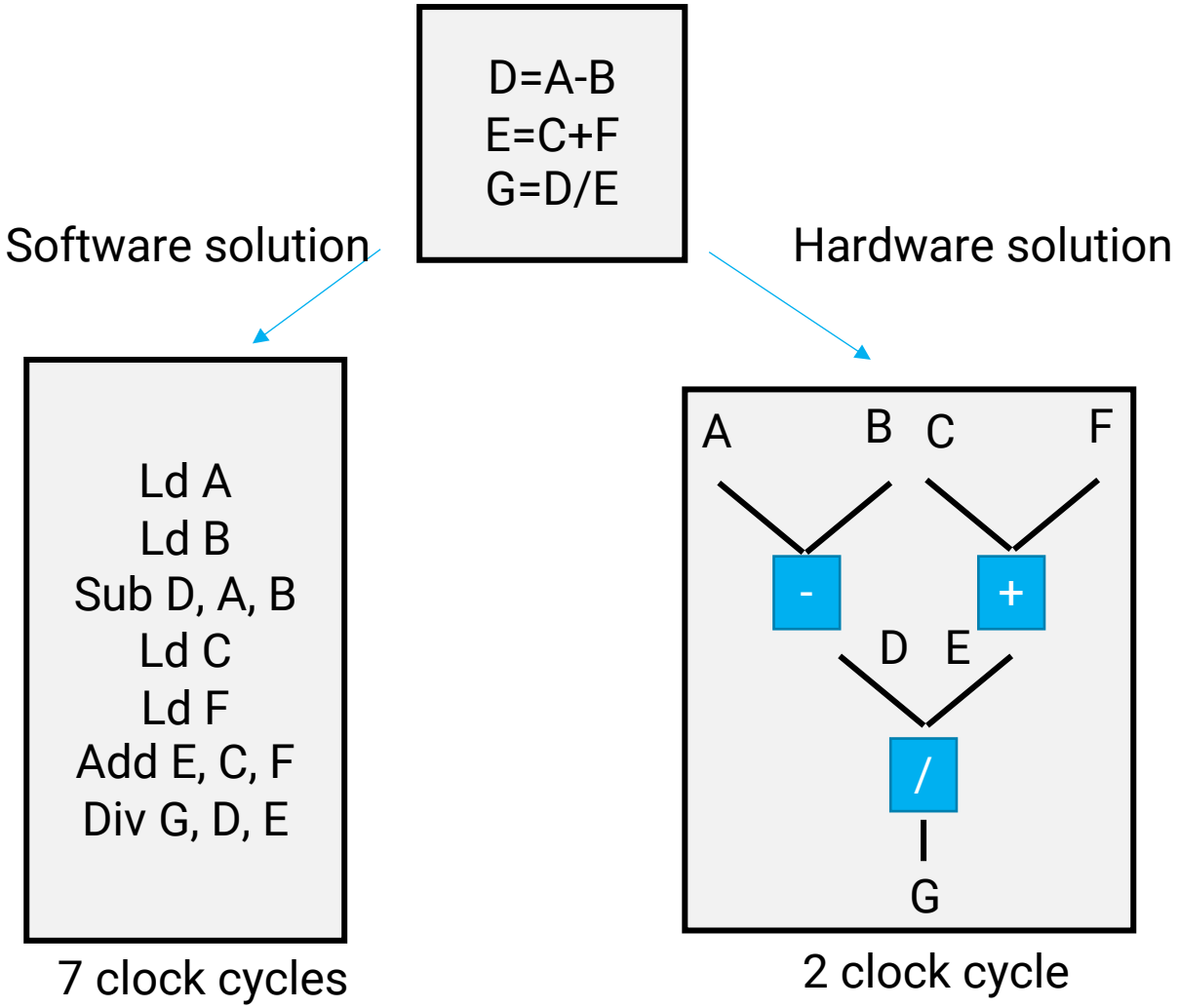


Apple M1 processor (2020)
8-core ARM, 16 billion transistors

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



Source: Mark McDermott

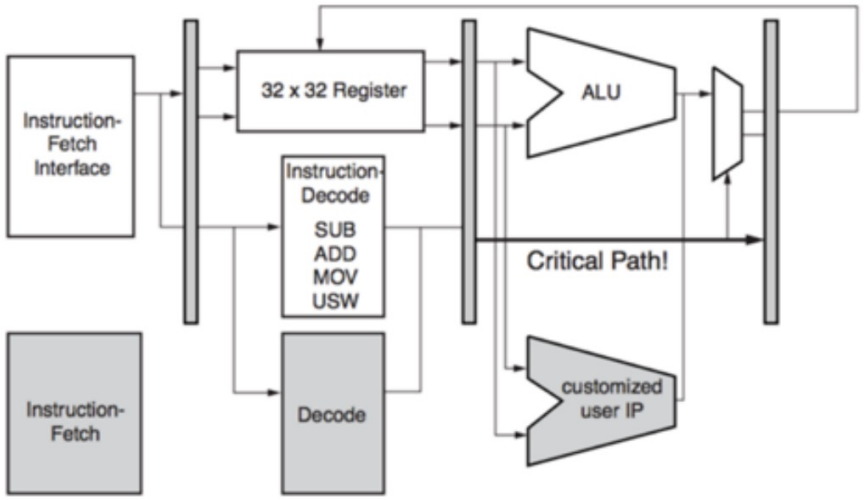
紧密连接 Tightly Coupled

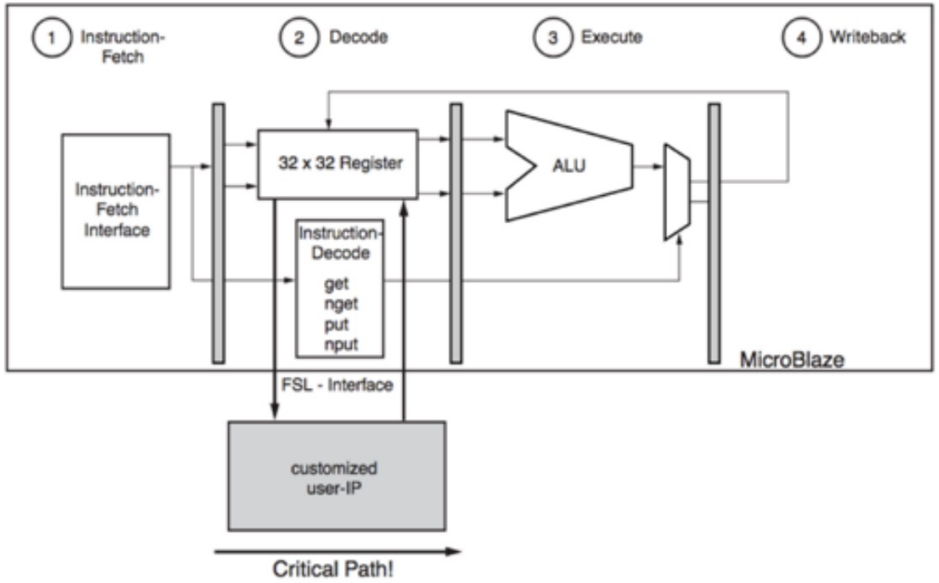
Integrated with processor control logic

– 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



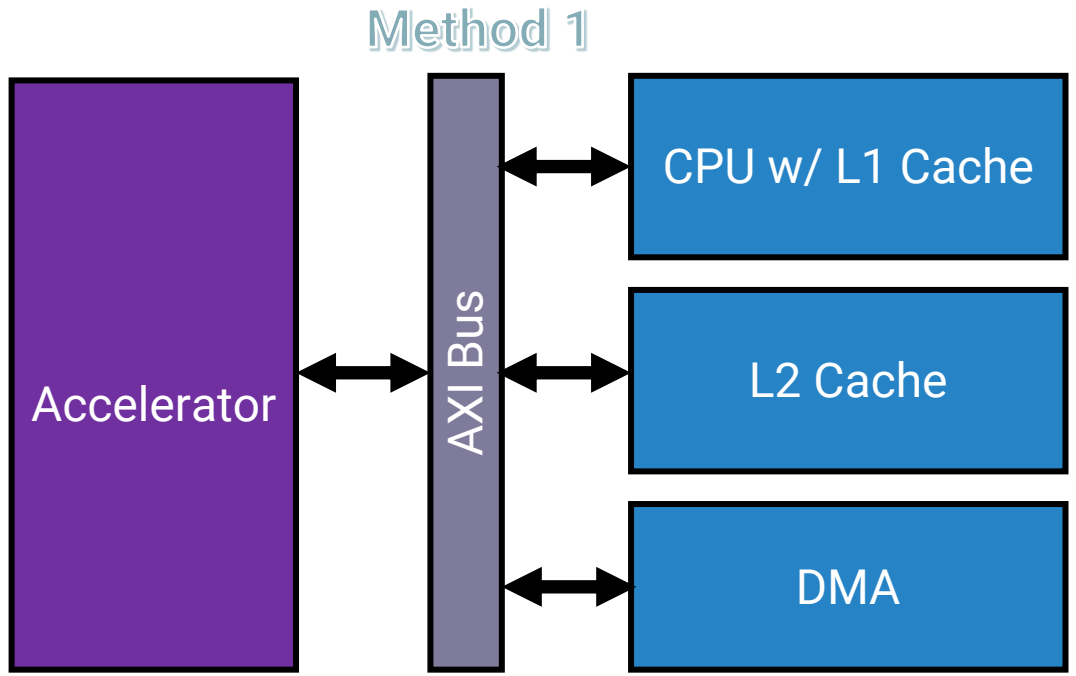


Source: Mark McDermott

松弛连接

Loosely-Coupled Co-processors

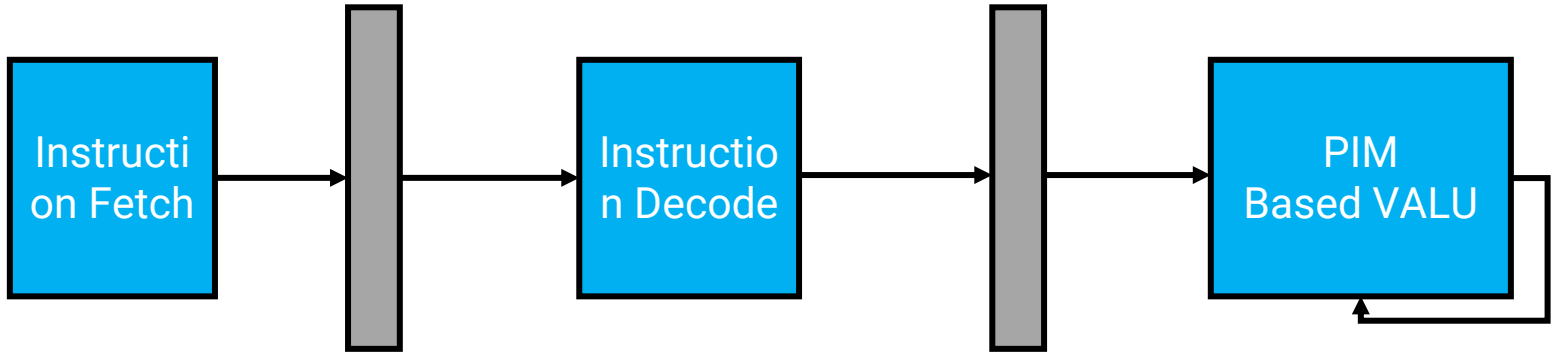
– Used for larger tasks than is the case for tightly-coupled coprocessors
– Task runs in parallel with main processor
– May take many cycles per task
– Large amounts of data that coprocessor may access independent of main processor – May or may not use the standard coprocessor interface



Source: Mark McDermott

Project Revisit: General Purpose VIP

- Vector In-Memory Processor (VIP)
- Qs:
  - Is this AI ASIC? Yes!
  - Is Programmable? Yes!
  - Expected efficiency vs. CNN accelerators? Lower!

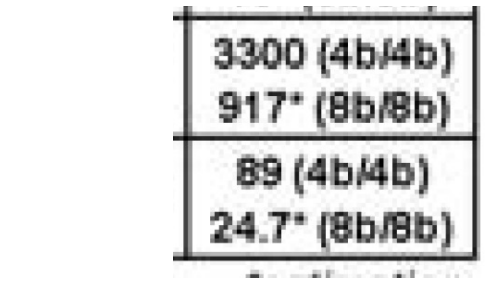


Vs conventional vector processor:
Combing register & ALU >> PIM-Based VALU

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^2)	NA	0.067	NA	0.2272	0.0032	NA	0.202
Bitcell Area (um^2)	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)

*estimation



Y-. D. Chih, et al., ISSCC'2021

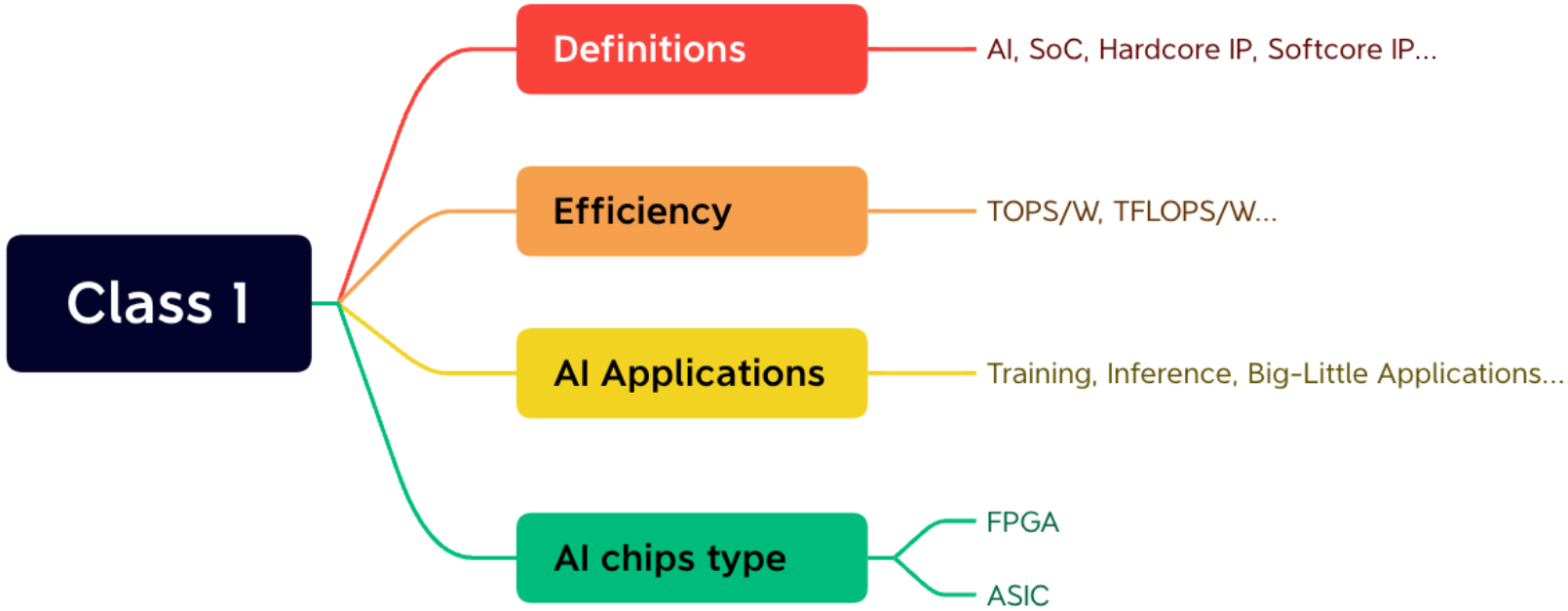
Efficiency Calculation Example 2

Technology	7nm
Array Size	4kb
Macro Area (mm^2) *	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

M. E. Sinangil, et al., JSSCC'2021

总结与思路梳理


Class 1
Definitions
AI, SoC, Hardcore IP, Softcore IP...
Efficiency
TOPS/W, TFLOPS/W...
AI Applications
Training, Inference, Big-Little Applications...
AI chips type
FPGA
ASIC

北京大學
PEKING UNIVERSITY

- The End