北京大学
PEKING UNIVERSITY

22530007

人工智能与芯片设计

1-课程介绍

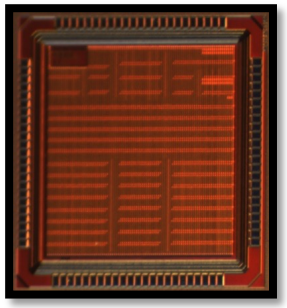
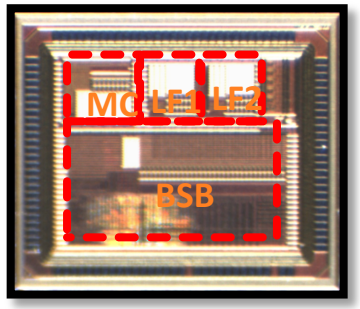
燕博南

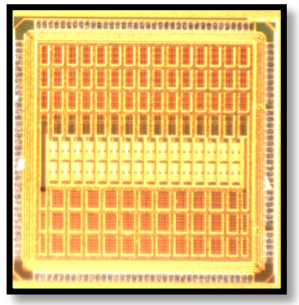
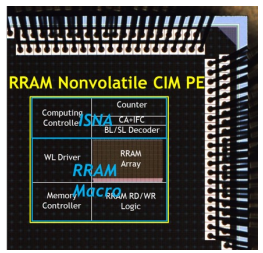
2023秋

课程教授：燕博南，Ph.D.

- 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

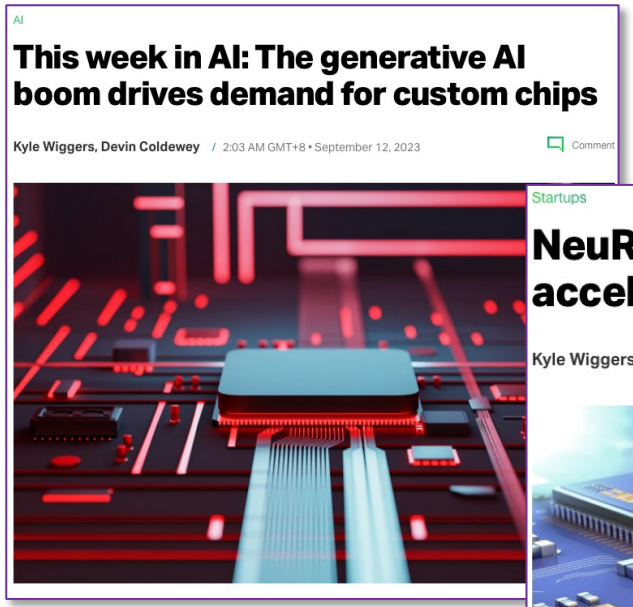
LABS hp

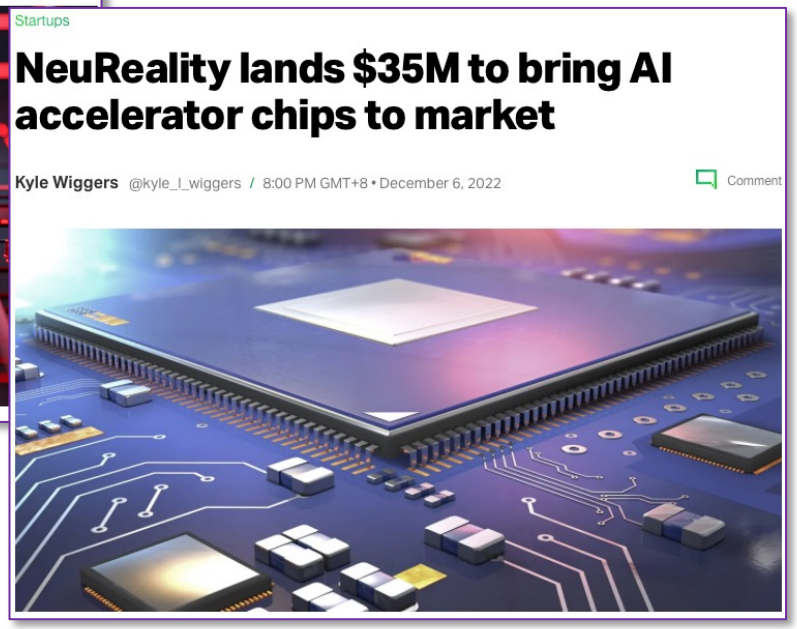
University of Pittsburgh

人工智能与芯片设计

Chip For AI

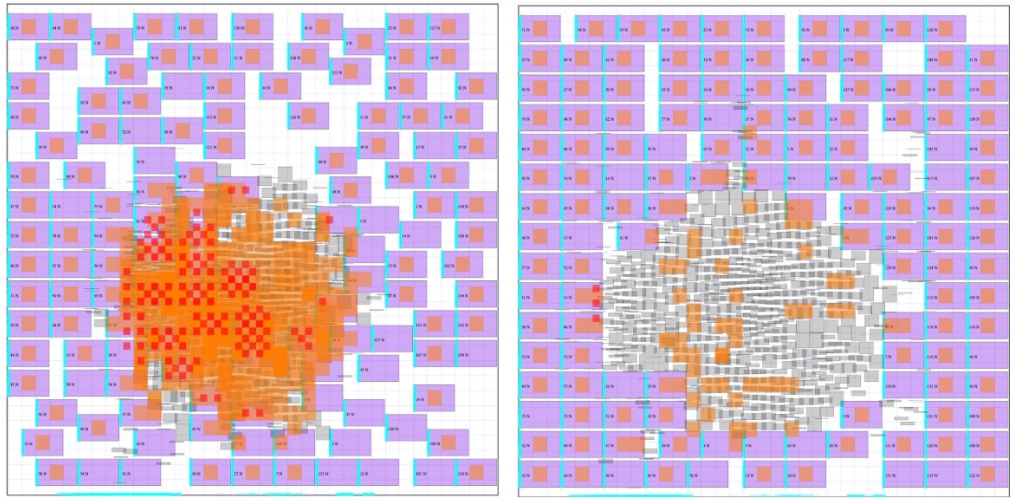
- 19 billion transistors
- CPU: 6-core CPU, 2 high-performance cores, and 4 high-efficiency cores
- GPU: 6-core with support for ray tracing
- Neural Engine: 16-core, 35 trillion operations per second
- Technology Node: 3nm





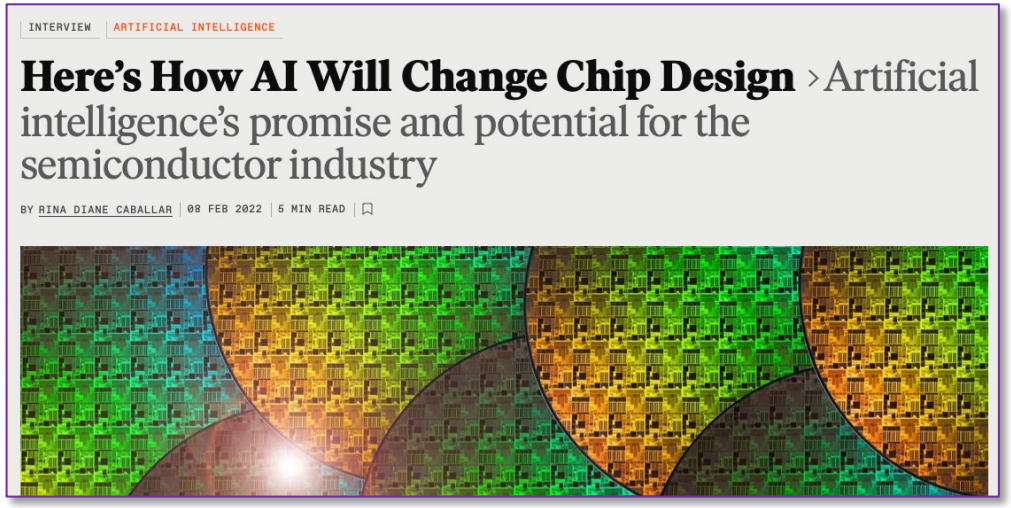
人工智能与芯片设计

AI for Chip



Extended Data Fig. 4 | Visualization of Ariane placements. Left, zero-shot placements from the pre-trained policy; right, placements from the fine-tuned policy. The zero-shot placements are generated at inference time on a previously unseen chip. The pre-trained policy network (with no fine-tuning) reserves a convex hull in the centre of the canvas in which standard cells can be placed, a behaviour that reduces wirelength and that emerges only after many hours of fine-tuning in the policy trained from scratch.

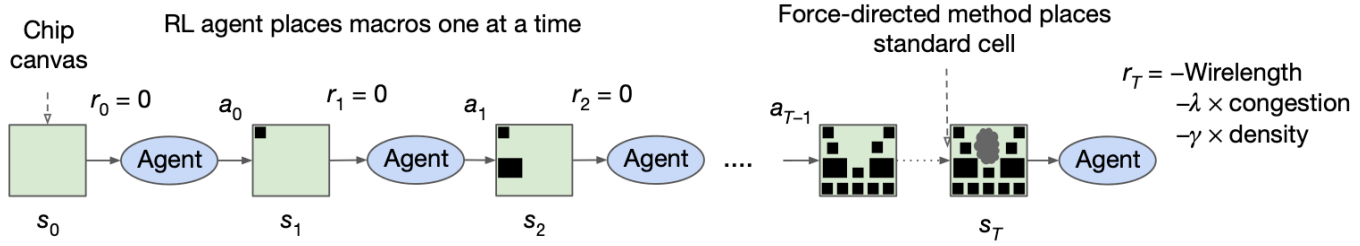
Too many transistors!



INTERVIEW | ARTIFICIAL INTELLIGENCE

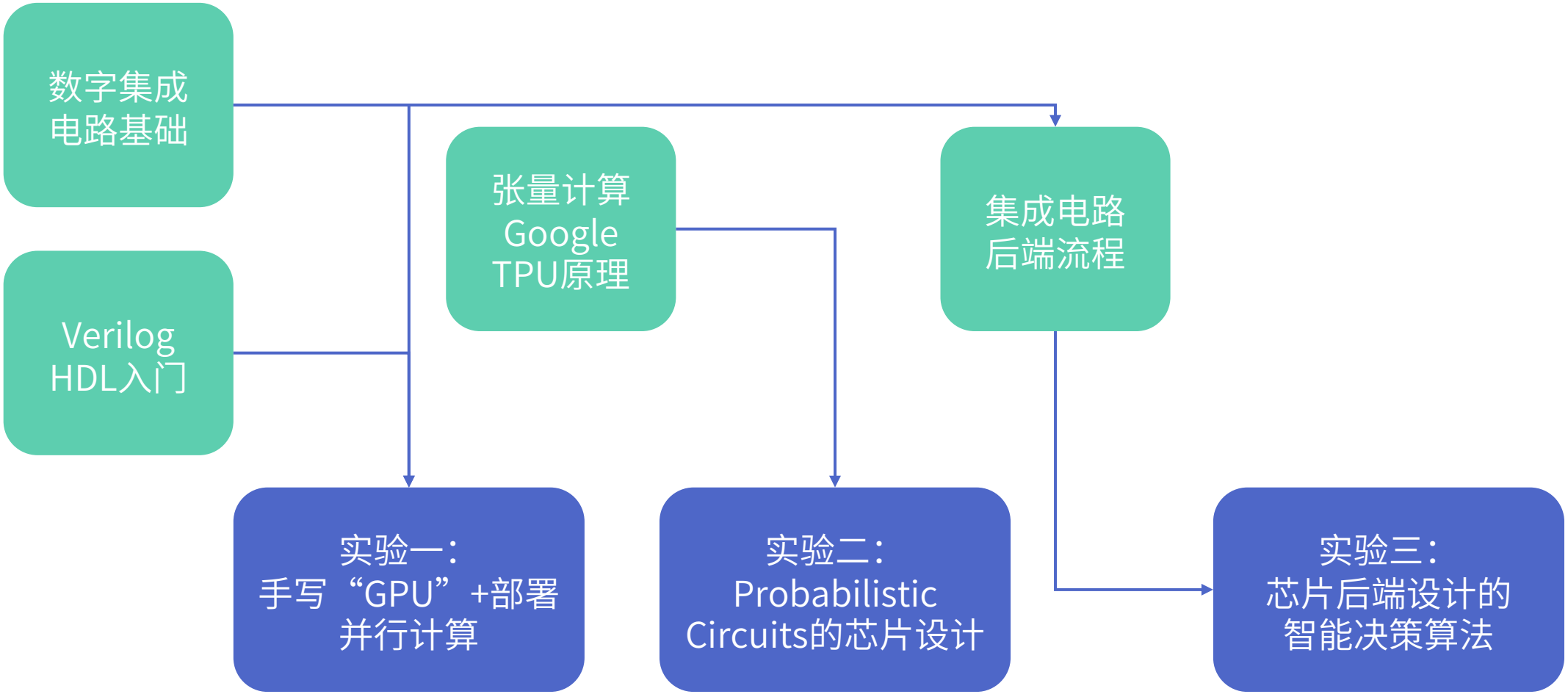
Here's How AI Will Change Chip Design › Artificial intelligence's promise and potential for the semiconductor industry

BY RINA DIANE CABALLAR | 08 FEB 2022 | 5 MIN READ



Google's approach

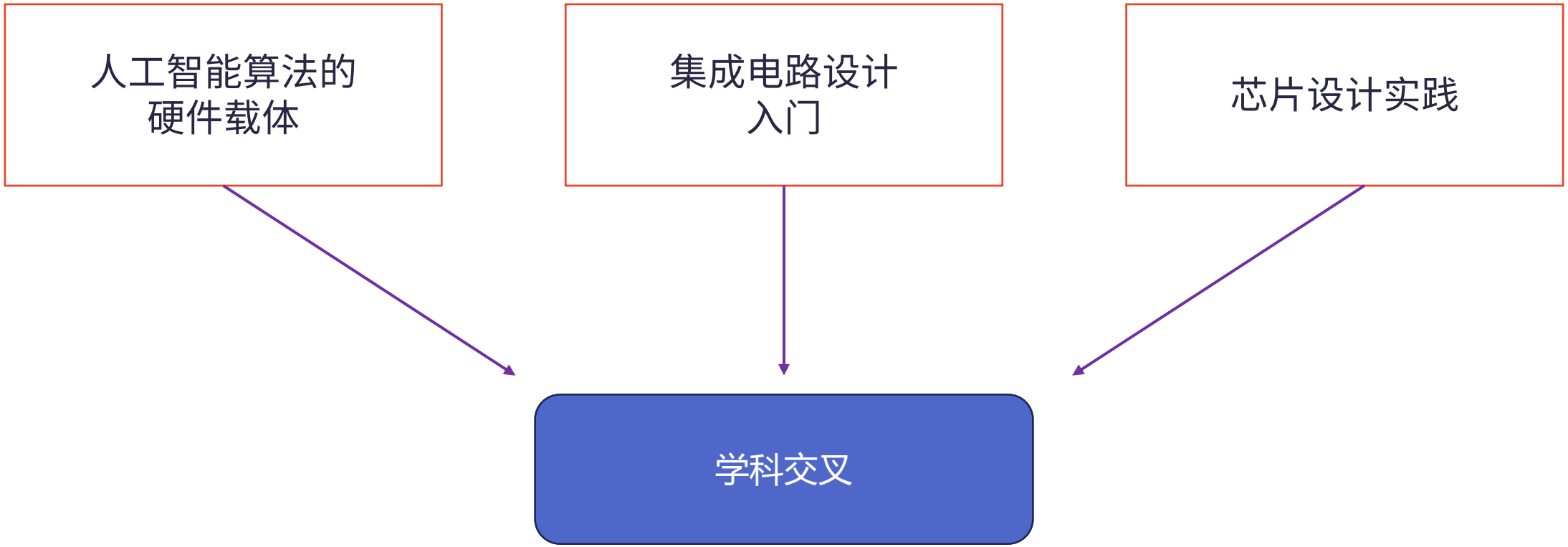
课程思维导图


数字集成
电路基础
Verilog
HDL入门
张量计算
Google
TPU原理
集成电路
后端流程
实验一：
手写“GPU”+部署
并行计算
实验二：
Probabilistic
Circuits的芯片设计
实验三：
芯片后端设计的
智能决策算法

课程考核

- 学期成绩评定：
  - 作业：20%（两次作业）
  - 讨论课（40%）
    - 实验随堂程序：10% × 3
    - 论文分享报告：10%
  - 实验书面报告，三选二（20% × 2）
  - 课程网站稍后推出

以终为始，你能学到什么



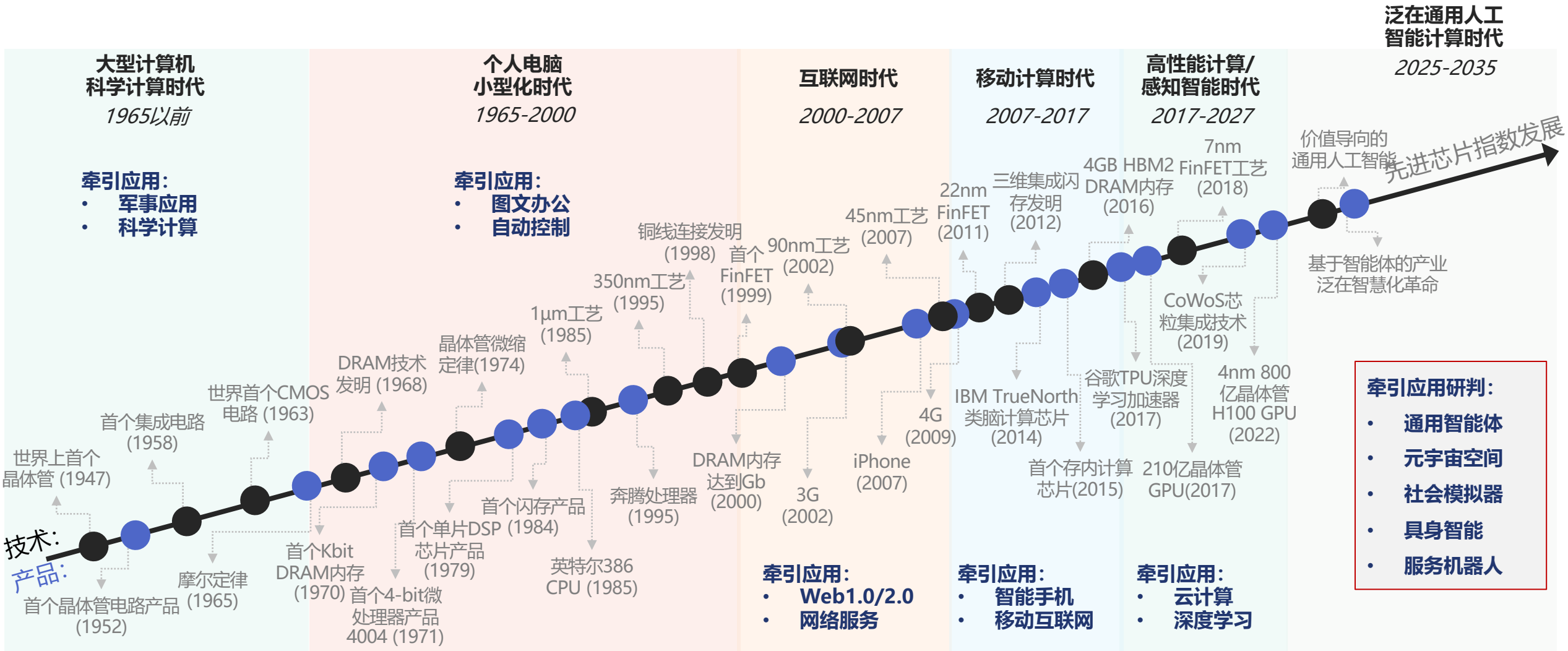
背景调查+自我介绍

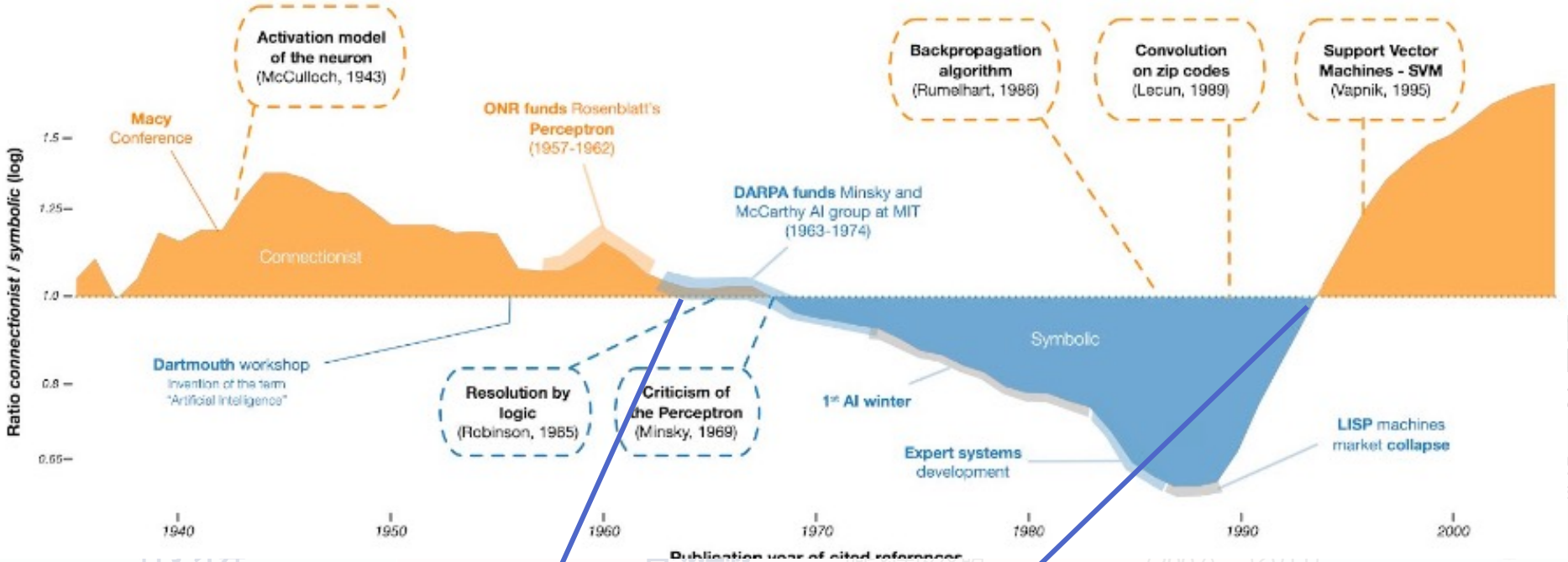
- 姓名
- 专业/学院（校）
- 感兴趣的学术/研究方向
- 之前是否学习过
  计算机组成原理/计算(微)架构/微机原理/操作系统/编译原理?
- 想从本课程学到什么?

Introduction to AI Chips

Why ASIC

Historic Perspective: AI & Chips





泛在通用人工智能计算时代
2025-2035

性能计算/
智能时代
017-2027

先进芯片指数发展

技术:
产品:

牵引应用研判:
- 通用智能体
- 元宇宙空间
- 社会模拟器
- 具身智能
- 服务机器人

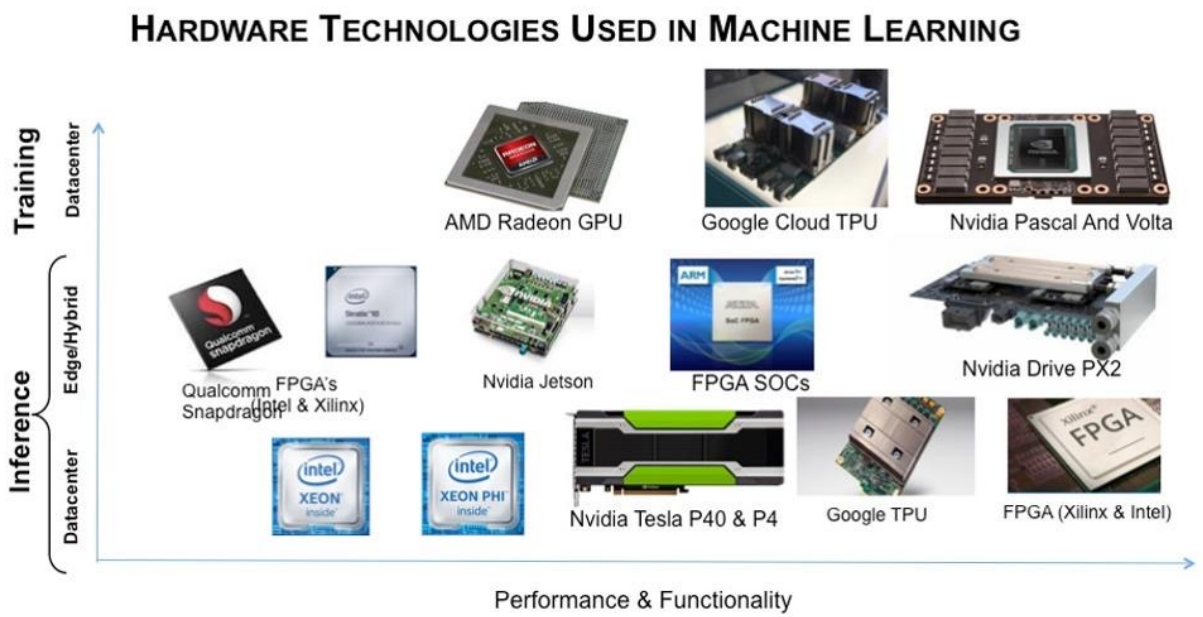
牵引应用:
- Web1.0/2.0
- 网络服务

牵引应用:
- 智能手机
- 移动互联网

牵引应用:
- 云计算
- 深度学习

为什么要有“人工智能芯片”？

- 硬件平台多种多样



Power?

Area? Volume?

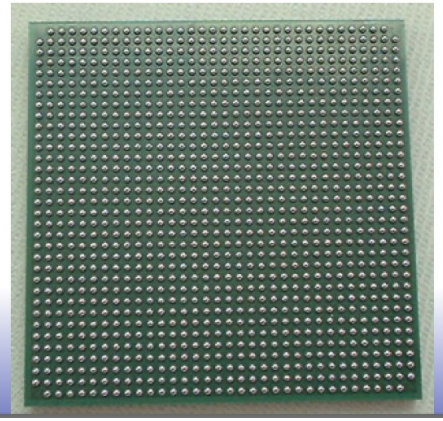
Performance?

Application-Specific Integrated Circuits (ASIC)

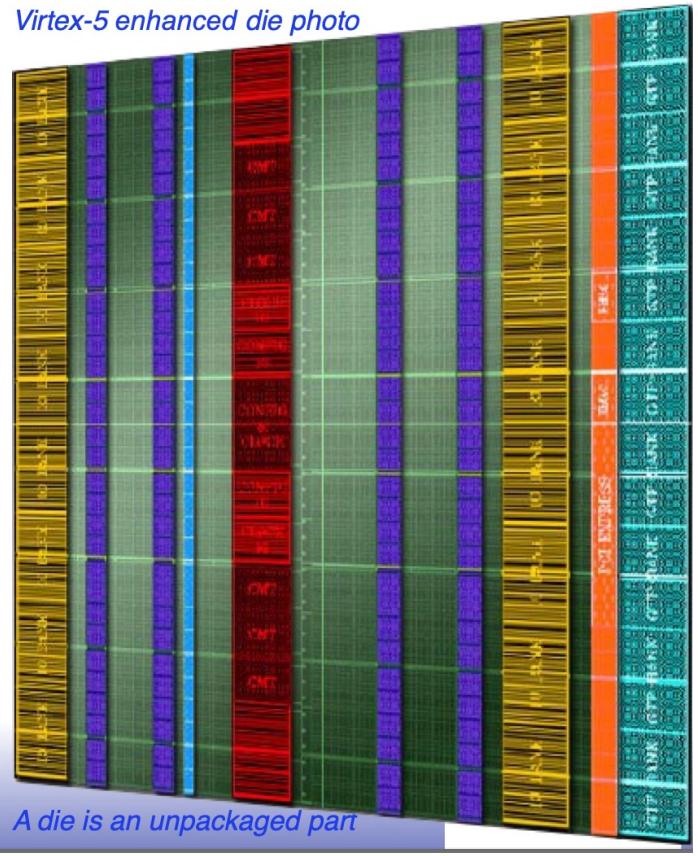
- 硬件永远不够用！

Field Programmable Gate Array (FPGA)

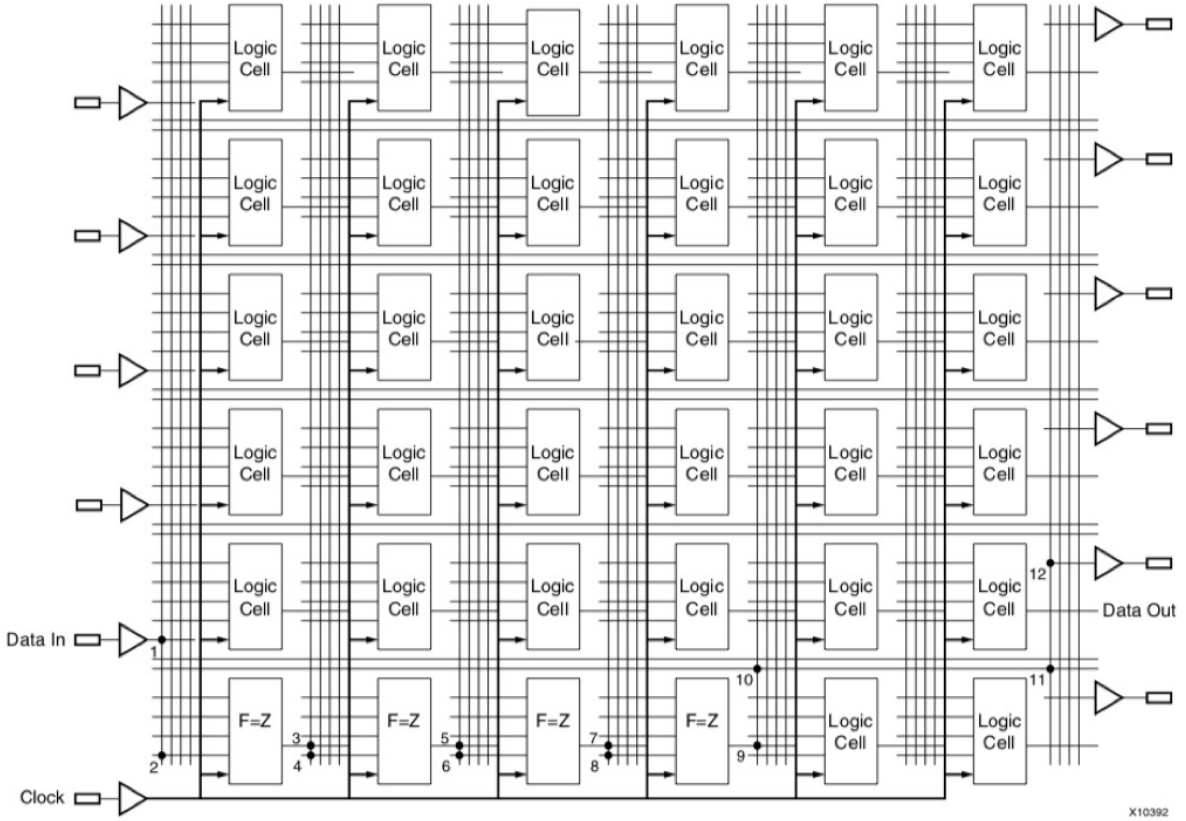
FPGA: Xilinx Virtex-5 XC5VLX110T



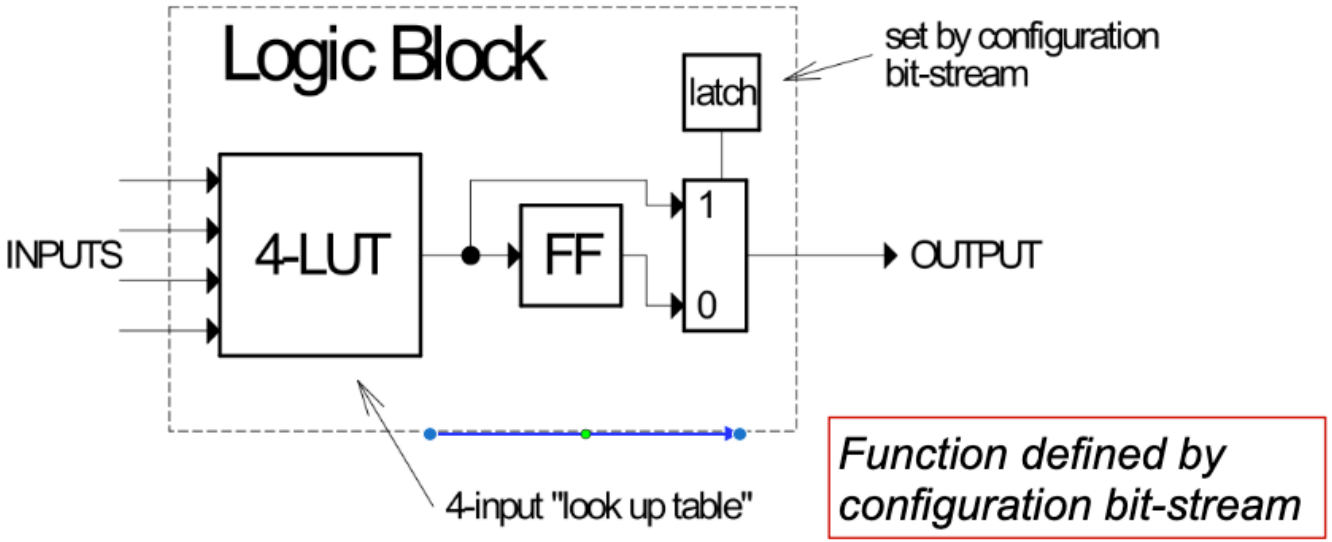
Virtex-5 enhanced die photo



A die is an unpackaged part



FPGA building block



Look up table (LUT)

- implements combinational logic function

Register (Flip-flop)

- optionally stores output of LUT

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
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..

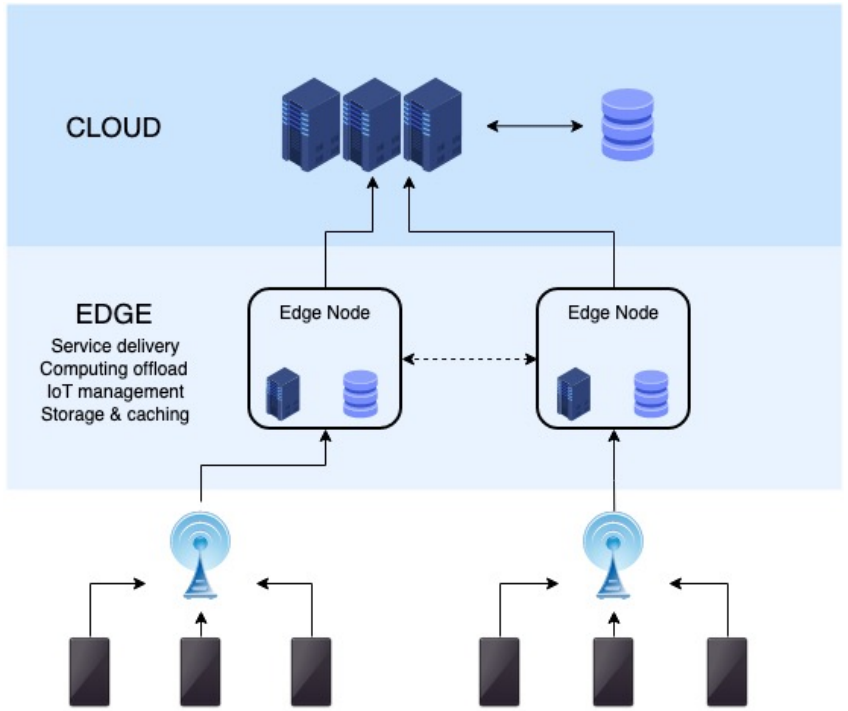
Processor A does INT8 Add, 1k times/second, area: $10mm^2$, what is the area efficiency?

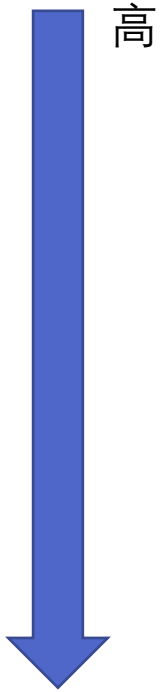
- Memory Density:
  - Unit: bit/mm^2 [bit per mm^2]
  - Storage Capacity/Area

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

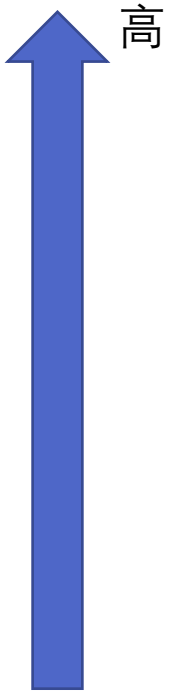
训练？推理？云？边缘？

- 算力层次





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



算力

规模/功耗

Ref. MLPerF

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:

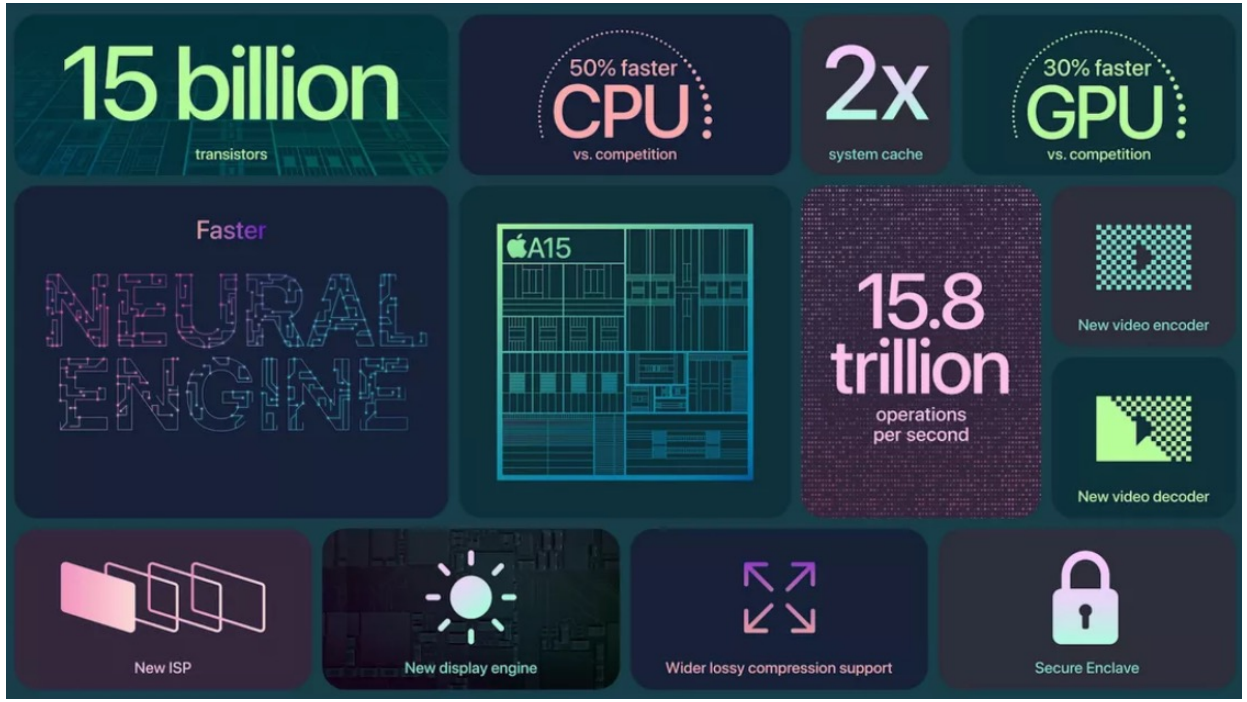
FPGA/ASIC路线的哲学问题

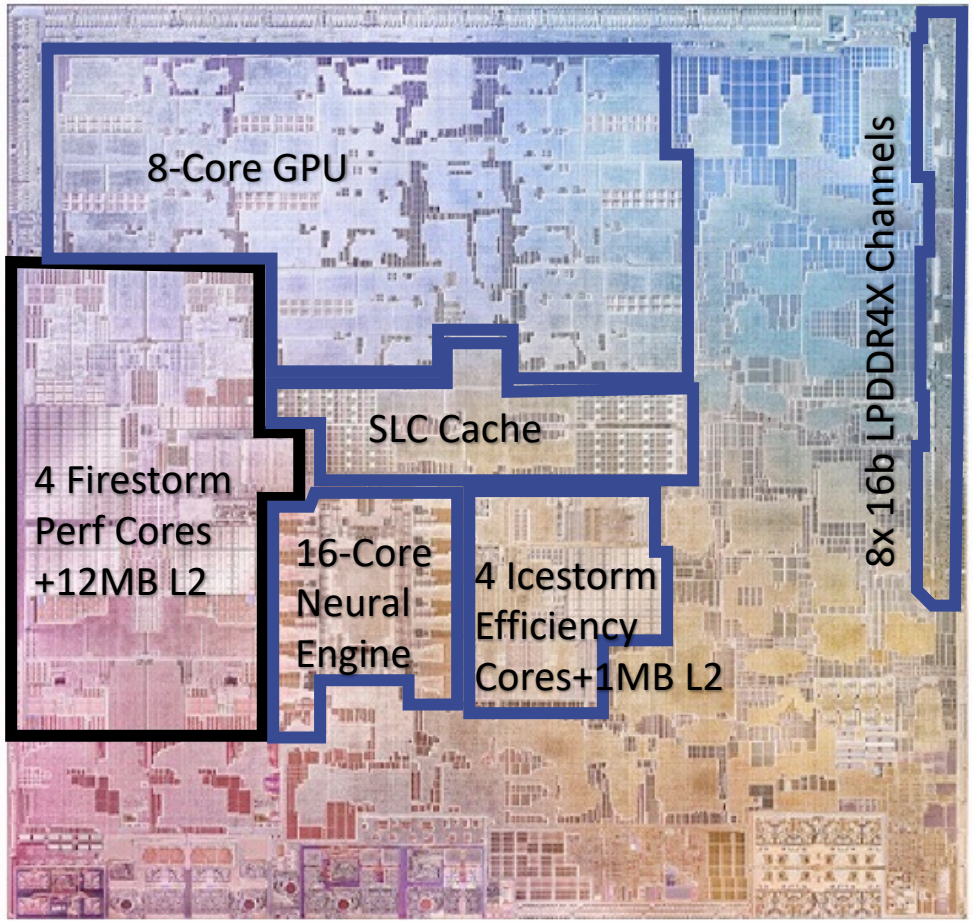
- 为什么用加速器?
  - Domain-Specific Accelerator
  - 提升算力
  - 提升效率
  - 相对降低成本
- 为什么用FPGA?
  - 可重构
  - 快速开发
  - 原型设计、硬件模拟
- 为什么用ASIC?
  - 从最底层(Gate、Transistor)开始优化
  - 灵活度高，完全符合应用需求

FPGA/ASIC路线的哲学问题

Heterogenous Computing SoC

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



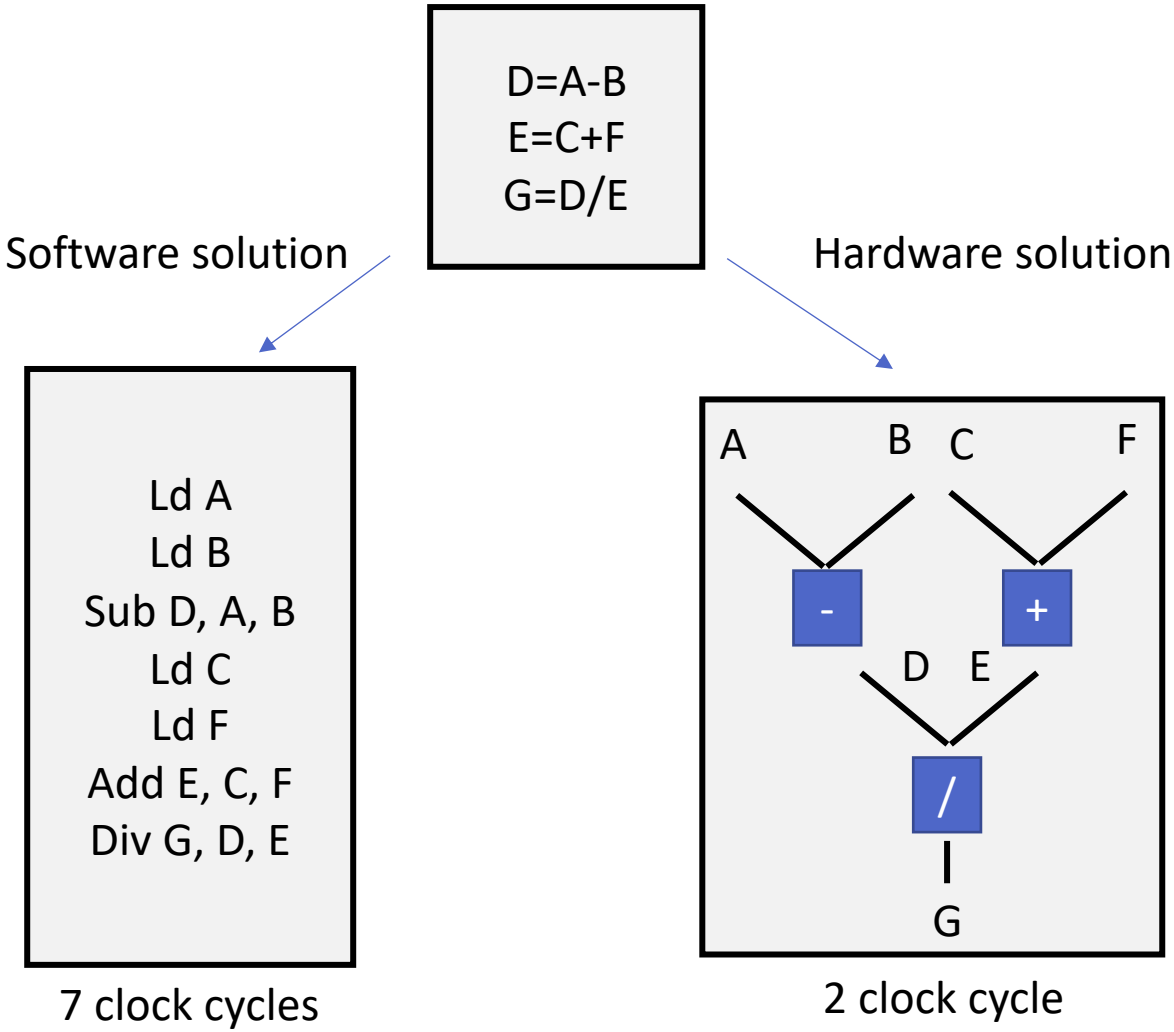


Apple M1 processor (2020)
8-core ARM, 16 billion transistors

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



Take-Aways

- AI chips are the foundation of AI
- Chip for AI & AI for Chip
- Why AI ASIC