AI芯片的思考

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem

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
   - 通用图形处理器 GPU
   - AI专用处理器 NPU/TPU
3. **GPU详解**
   - 英伟达GPU架构发展
   - Tensor Core和NVLink
4. **国外 AI 芯片**
   - 特斯拉 DOJO 系列
   - 谷歌 TPU 系列
5. **国内 AI 芯片**
   - 壁仞科技芯片架构
   - 寒武纪科技芯片架构
6. **AI芯片的思考**
   - SIMD&SIMT与编程体系
   - AI芯片的架构思路与思考

Talk Overview

I. AI 芯片的思考

- SIMD & SIMT 区别与联系
- SIMT 与 CUDA 编程
- GPU 在 SIMT 编程本质
- SIMD、SIMT 与 DSA 架构
- DSA（AI芯片）架构主要形态
- AI 芯片架构的黄金十年

基于硬件编程，让生态繁荣

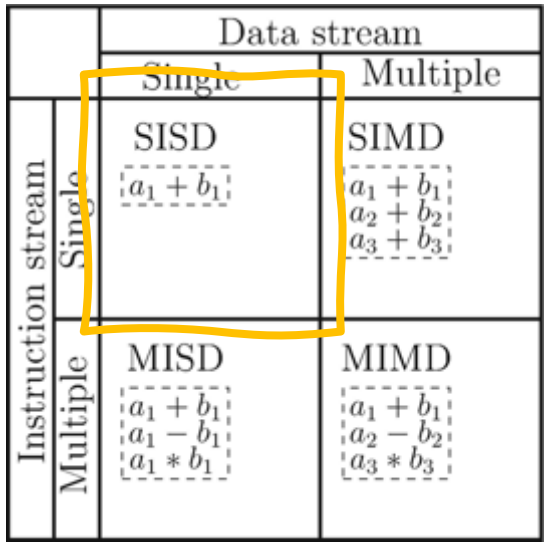
并行处理硬件架构

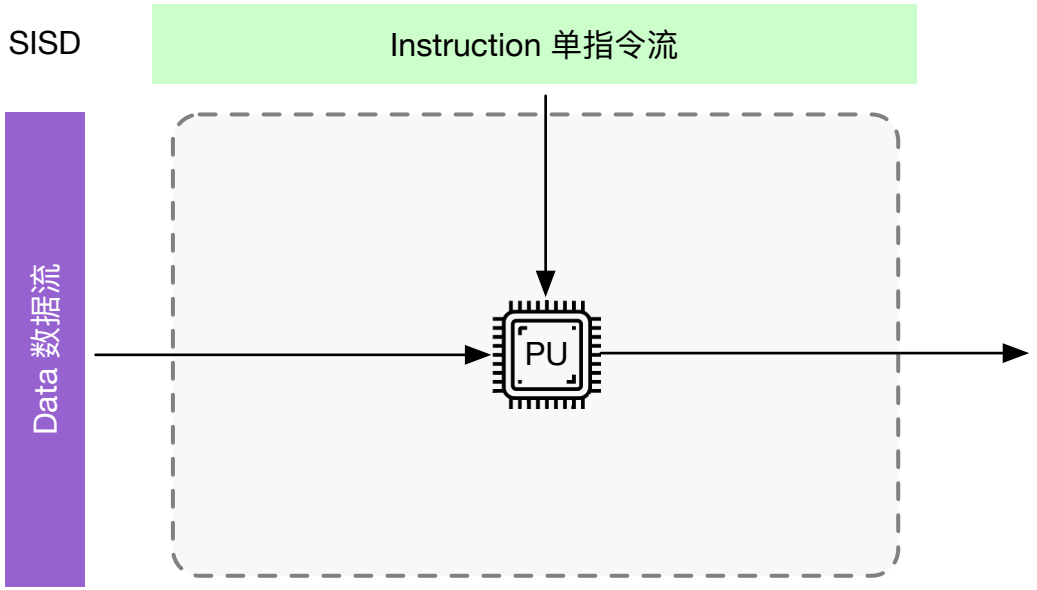
- 单指令流单数流（SISD）系统。
- 单指令流多数据流（SIMD）系统。
- 多指令流单数据流（MISD）系统。
- 多指令流多数据流（MIMD）系统。

		Data stream	
		Single	Multiple
Instruction stream	Single	SISD $a_1 + b_1$	SIMD $a_1 + b_1$ $a_2 + b_2$ $a_3 + b_3$
	Multiple	MISD $a_1 + b_1$ $a_1 - b_1$ $a_1 * b_1$	MIMD $a_1 + b_1$ $a_2 - b_2$ $a_3 * b_3$

并行计算处理硬件架构（I）：SISD 系统

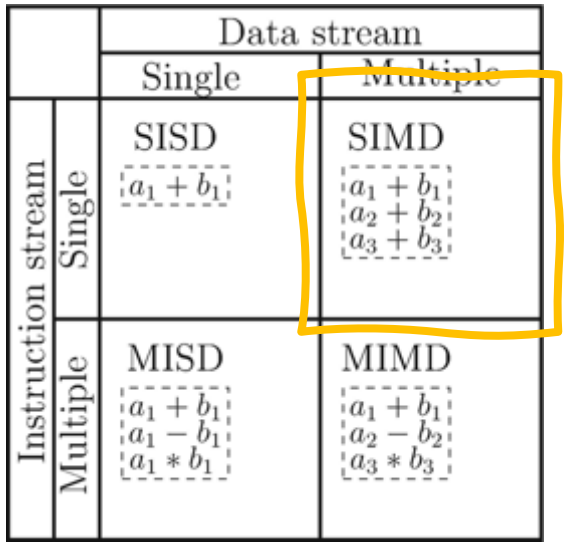
- 每个指令部件每次仅译码一条指令，而且在执行时仅为操作部件提供一份数据
- 串行计算，硬件不支持并行计算；在时钟周期内，CPU只能处理一个数据流。

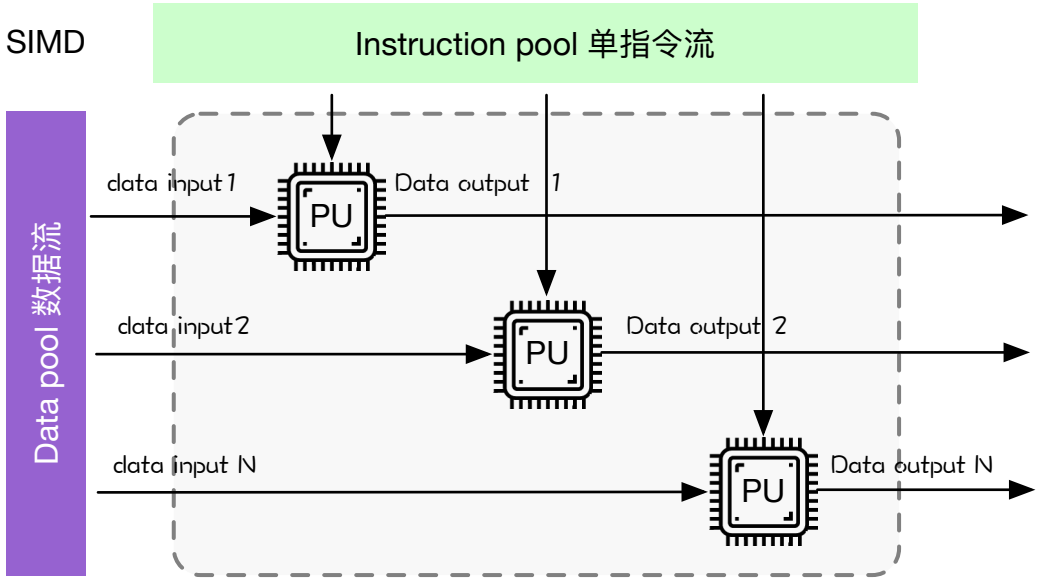




并行计算处理硬件架构（II）：SIMD 系统

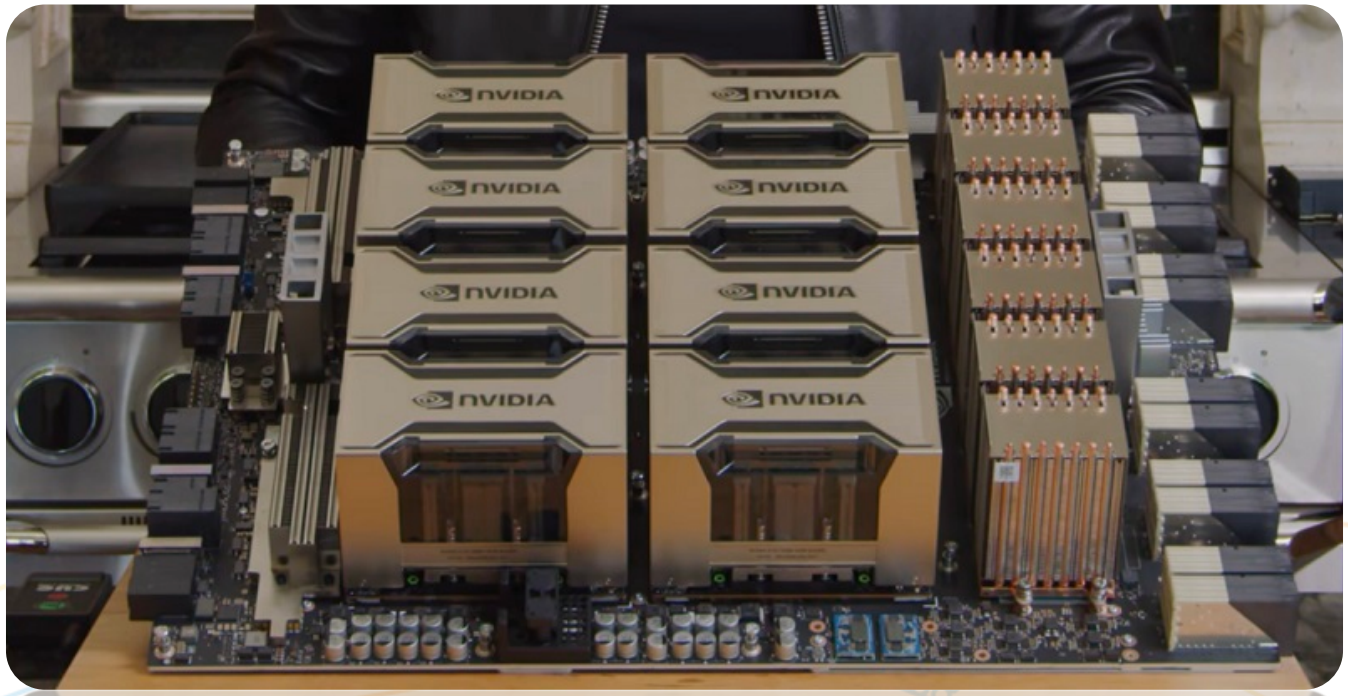
- 一个控制器控制多个处理器，同时对一组数据中每一个分别执行相同操作
- SIMD主要执行向量、矩阵等数组运算，处理单元数目固定，适用于科学计算
- 特点是处理单元数量很多， 但处理单元速度受计算机通讯带宽传递速率的限制





1. 为什么AI编程关注SIMT和SIMD？

GPU 真香

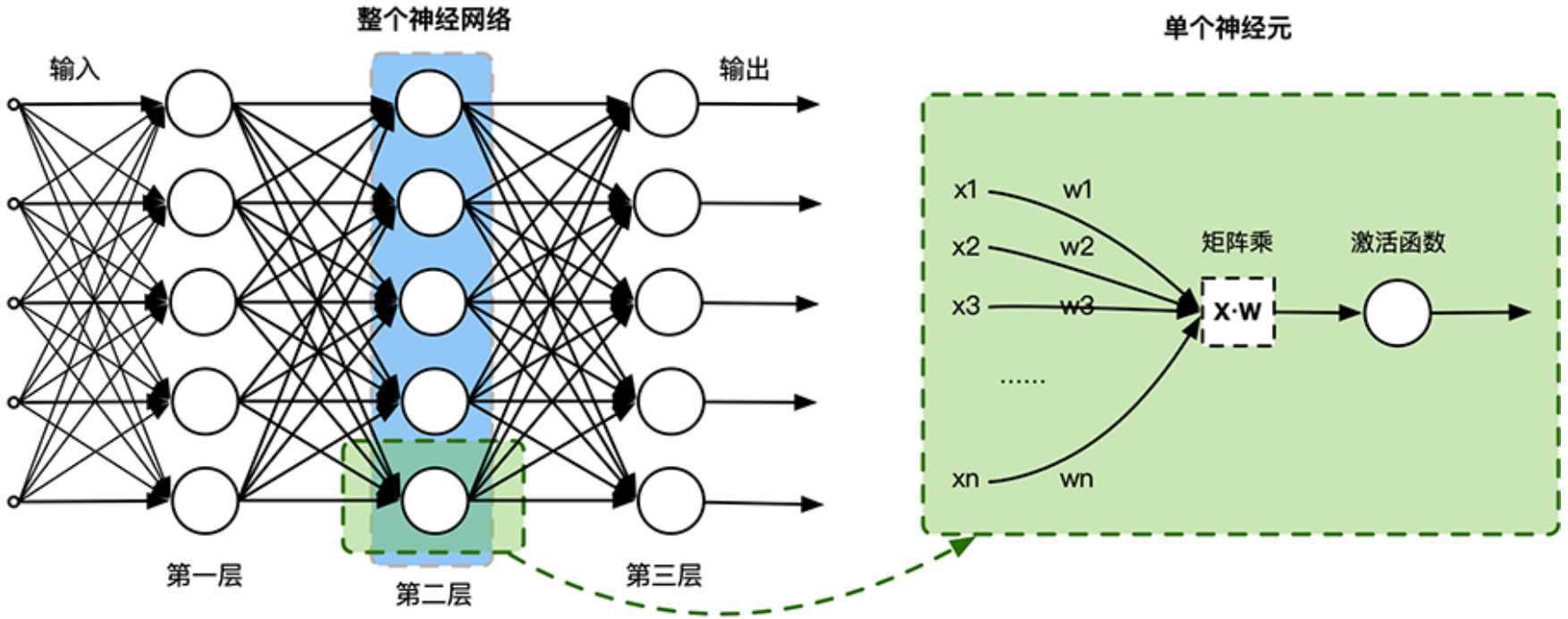


CUDA 真香

英伟达 真香

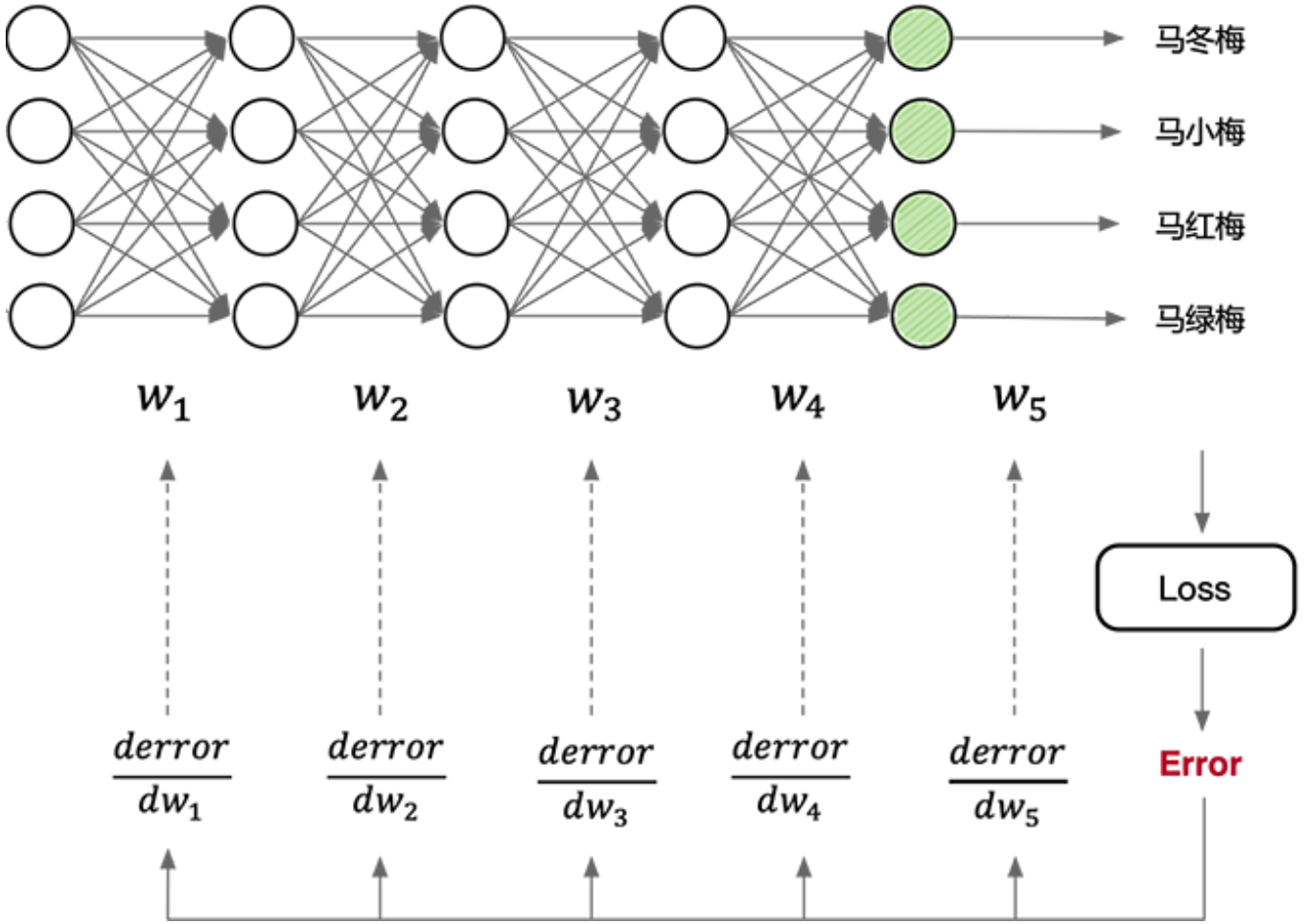
Example: Deep Neural Networks

- inspired by neuron of the brain
- Neurons arranged in layers
- Computes non-linear activation function of the weighted sum of input values



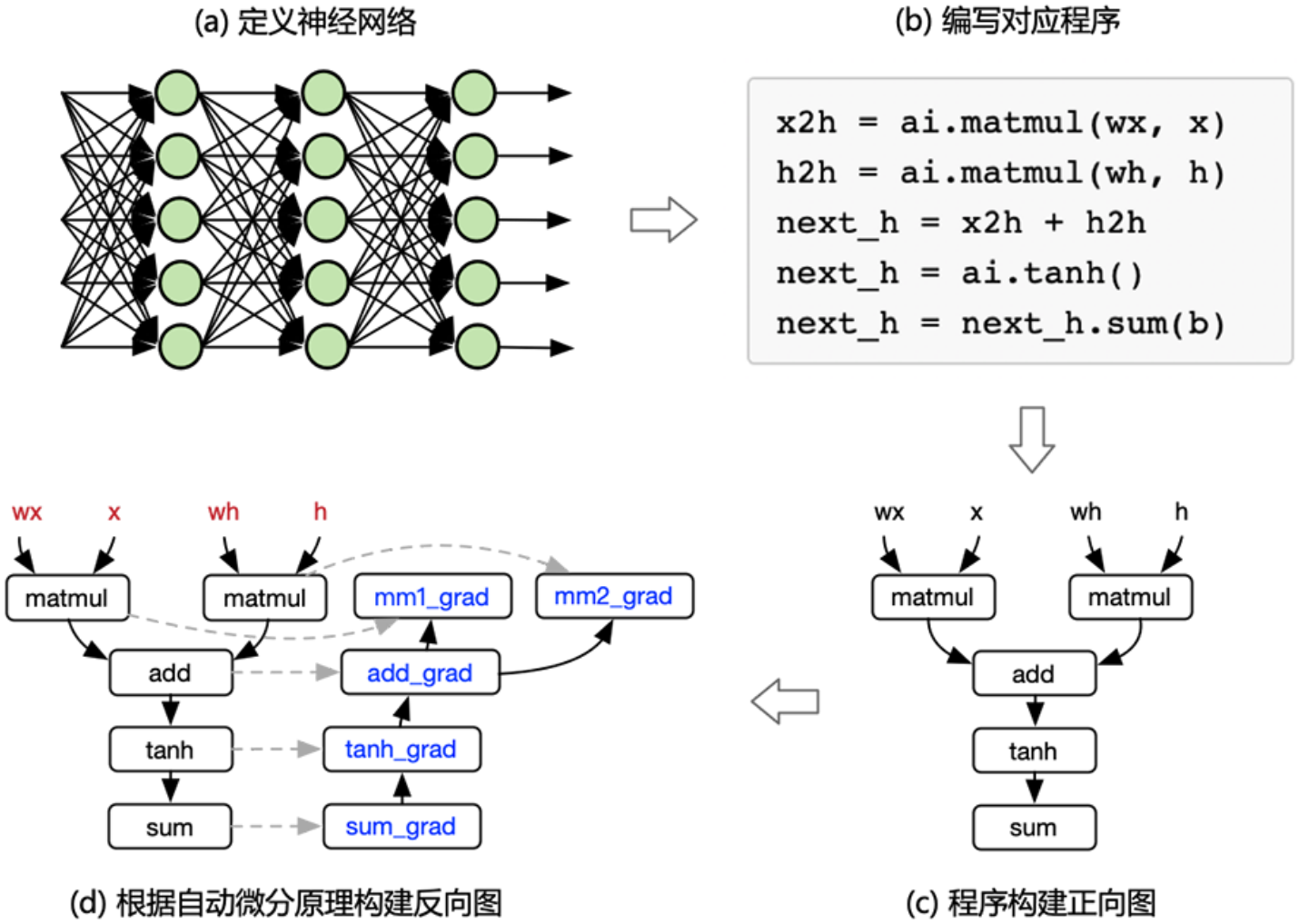
Example: Deep Neural Networks

- **Training 训练**
  - 将数据集以 mini-batch 反复进行前向计算并计算损失，反向计算梯度利用优化函数来更新模型，使得损失函数最小
- **Inference 推理**
  - 对神经网络模型执行一次前向计算的过程，到预测结果



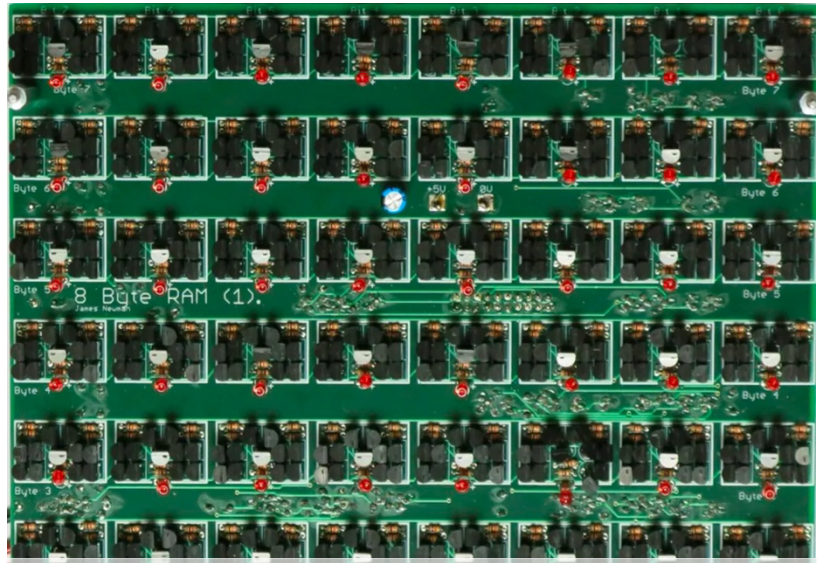
AI框架的开发流程

- SIMD、SIMT 在哪里？
- 这不应该是AI系统应该解决吗？
- SIMD 应该只对底层硬件设计有约束？



硬件执行模型和编程模型

- AI 框架中的算子、硬件提供的 Kernel 实现，需要根据硬件执行模型 Execution Model 来确定编程模型 Programming Model。
- 根据计算机系统设计里面的定义，AI 系统采用并行计算处理硬件架构以 SIMD 为主，但是编程模型却以生态强大的英伟达 CUDA 定义的 SIMT 方式为主。



思考

- SIMD 和 SIMT 之间怎么区别？
- SIMD、SIMT、DSA之间有什么样的区别联系？
- 在AI体系结构里面，SIMD 和 SIMT 的设计，对编程模型带来哪些挑战？

思考

- **编程的时候，请告诉我，哪个程序员谁的去做控制指令？SIMD了？**
- **SIMD 是不会暴漏给开发者，作为硬件并行处理架构，那么 SIMT 为啥会暴漏给开发者？**
- 英伟达的 SIMT 到底是编程模式还是硬件模式？
- 开发者到底是在控制线程还是只是简单编程？
- 线程的硬件执行是什么方式？SIMD？

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem