RISC-V 中国峰会

RISC-V与虚拟指令技术结合打造创新的计算架构

杨宜 博士
奕行智能 COO

EVAS INTELLIGENCE | 奕行智能

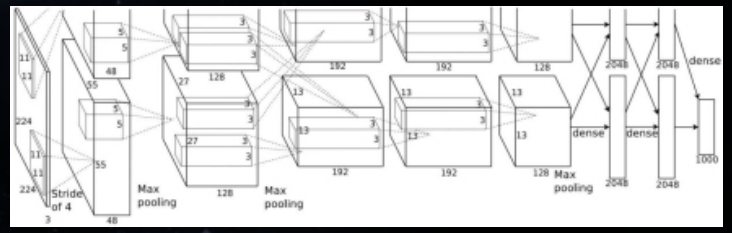
目录

1 AI的发展改变了软件编程的范式
2 AI处理器在计算效率与通用性方面的挑战
3 RISC-V + RVV是AI计算范式的最佳选择
4 AI处理器指令的路径选择
5 RISC-V + VISA的AI计算架构

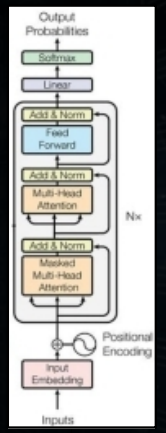
AI的发展改变了软件编程的范式，Hardware 3.0时代到来

Software 1.0	Software 2.0	Software 3.0
computer code	weights	prompts
↓	↓	↓
computer	neural net	LLM

became programmable in **~1940s**



fixed function neural net
e.g. AlexNet: for image recognition **~2012**



LLM = programmable neural net! **~2019**

CPU dominated	GPGPU dominated	DSA will dominate
↑	↑	↑
Hardware 1.0	**Hardware 2.0**	**Hardware 3.0**

Source: Andrej Karpathy on Software 3.0: Software in the Age of AI

AI处理器在计算效率与通用性方面的核心矛盾

AI计算是领域特定范式，却因模型多样性和海量编程用户需求，需兼顾领域内通用性和专用性

算力利用率双重要求

硬件计算单元利用率

最大化硬件的计算资源使用率，避免算力浪费
优化计算任务的并行能力、访存效率，适配不同硬件架构

用户编程易用性

兼容主流框架，提供高层抽象，降低开发者使用硬件的门槛
隐藏底层硬件细节，同时允许高级用户进行深度优化

AI计算领域特殊性

模型算法多样性，海量算子

模型计算模式多样（密集/稀疏/动态），需高效适配
海量算子支持需兼顾性能与利用率，避免硬件低效

海量开发者生态需求

既要满足研究人员的快速实验需求（易用性）
又要满足工业界的高效部署（利用率）

核心矛盾：领域专用效率 vs 编程通用性

构造AI处理器过程中面临诸多挑战

1 AI计算架构设计耗时长

从零开始构造AI计算架构所需要的时间较长，涉及复杂的技术决策与优化过程。

2 指令系统打磨时间久

从零开始构造一套经过产品与生态认可的指令系统需要大量时间进行验证与完善。

3 AI编译软件落地周期久

从零开始构造后端编译软件并达到成熟可商用水平，需经历长时间的调试和优化。

4 生态兼容难题

自主构建的指令系统需获得广泛生态支持，这一过程存在较高的门槛与不确定性。

RISC-V + RVV 是AI计算架构的最佳选择

国家八部门联合起草指导政策，鼓励全国使用开源RISC-V芯片

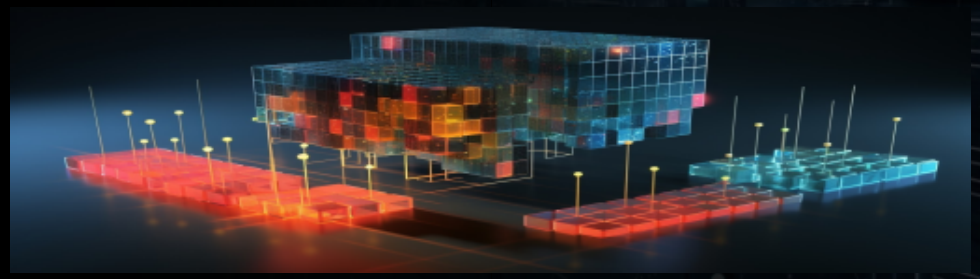
发布于2025-03-05 11:40:53 综合报道 分享 海报分享

架构优势

- 开放的图灵完备指令，定制灵活
- RVV可变长向量，适应多样化数据
- 天然SIMD并行，内存访问高效

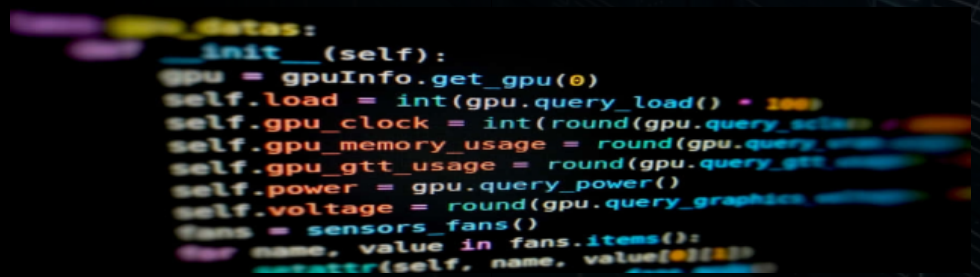
技术适配性

- 更容易实现软硬协同设计
- RVV向量操作直接对应AI的张量计算
- 向量掩码操作天然支持稀疏矩阵运算
- 可根据性能需求配置不同规模的向量单元



生态优势

- GCC、LLVM等主流编译器已支持RISC-V
- 主流AI框架正积极适配RISC-V平台
- 开源调试和性能分析工具日趋完善



AI处理器指令的路径选择

CPU与GPU作为成熟的处理器架构其原子指令都是微指令，软硬件之间有着清晰的分工合约。

而作为AI处理器的指令选择有以下不同策略：

固化的ASIC

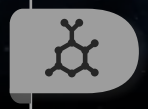
优势：针对单一算法专用性强 / 能效高

劣势：一旦有新的模型算法产生，ASIC则无法适配或性能较低

高层次粗颗粒度指令

优势：单条指令可完成复杂操作，简化了软件设计的复杂度

劣势：算子数量多，若全部指令化的面积开销大



低层次细粒度微指令

优势：可图灵完备，灵活性高，可实现复杂的算法逻辑

劣势：指令数量多，调度开销大，软件开发周期长

有没有一种创新的计算架构，

既能够保留算子级粗颗粒指令的语义，

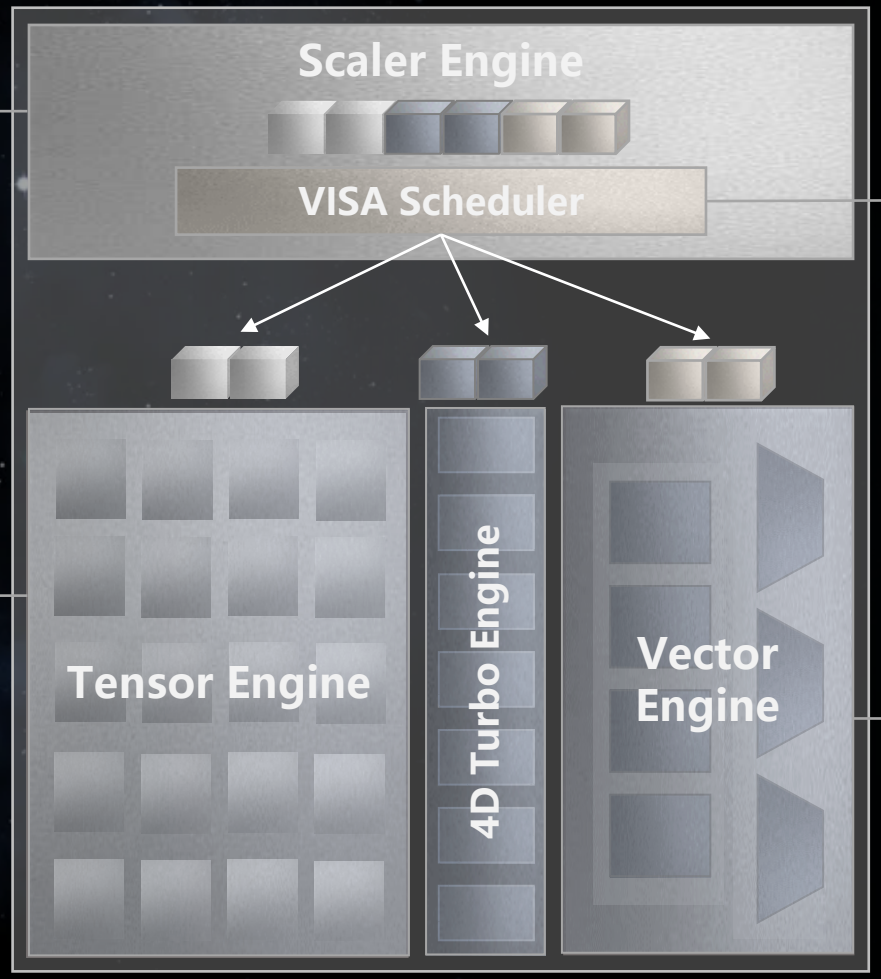
给编程者更好更高效的编程界面，

同时又能保证细颗粒度指令的图灵完备性？

EVAS计算架构设计：RISC-V + VISA 融合

VISA = Virtual Instruction Set Architecture 虚拟指令集架构

标量引擎不直接参与AI计算，负责整个系统的协调和控制



VISA调度器保证粗粒度宏指令的编排和乱序发射

张量引擎负责专门处理矩阵运算和张量计算

RISC-V RVV向量引擎提供AI专用的硬件扩展，保证细粒度微指令的高效执行

EVAMIND™ AI内核

Epoch – 首款RISC-V + VISA的AI计算架构产品

Epoch – 新一代AI计算架构产品，即将揭晓

RISC-V搭配独有的虚拟指令集技术，兼顾通用性和专用性

集成多组EVAMIND AI内核，具备图灵完备向量引擎 + 大尺寸张量单元

兼顾AI专用与通用计算任务，硬件级指令化提升效率

FP8/INT4原生支持，2-4倍计算吞吐提升

支持INT4,INT8,FP8,FP16,BF16,TF32,FP32 等多种浮点和定点数据类型

支持大模型特需的混合精度计算（权重W与激活A采用不同的数据类型）

多种并行及流水掩盖计算方式，实现计算资源的极致利用率

矩阵，向量，数据引擎充分流水并行，最大化矩阵计算单元利用率

支持模型并行，数据并行，张量并行等多种分布式训练方式

EVAS INTELLIGENCE | 奕行智能

奕行智能科技有限公司（EVAS Intelligence）成立于2022年，
是一家专注于提供前沿的AI计算架构和高效能并行计算解决方案的通用计算芯片设计公司，
在上海、广州、南京、杭州、深圳、北京等地设有研发及运营中心。

奕行智能致力于以先进的计算架构、编译器软件工具为核心技术，
通过RISC-V开放指令集生态提供新一代通用和专用计算加速解决方案，
推动人类社会的进步和可持续的发展。


EVAS
INTELLIGENCE
奕行智能