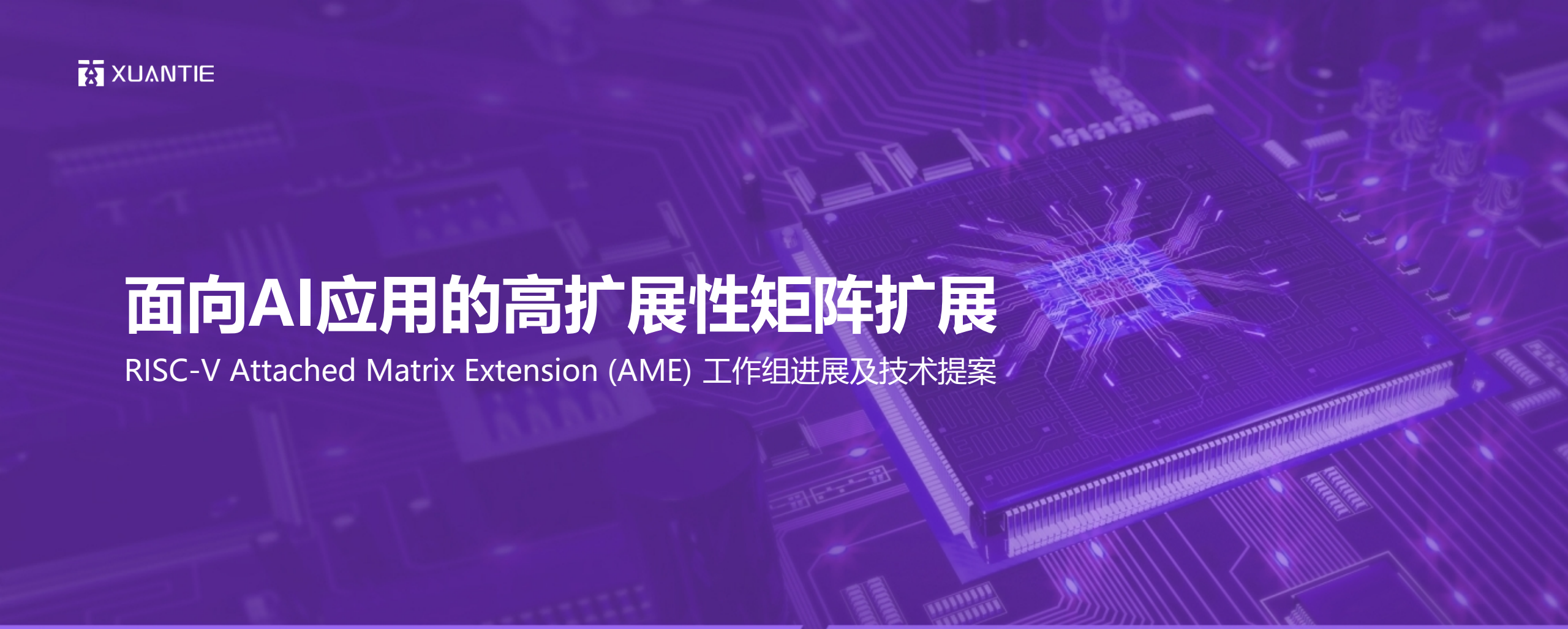
XUANTIE

面向AI应用的高扩展性矩阵扩展

RISC-V Attached Matrix Extension (AME) 工作组进展及技术提案



赵思齐

RISC-V 基金会 AME TG Chair

达摩院RISC-V及生态 玄铁架构组

XUANTIE

CONTENTS

目录

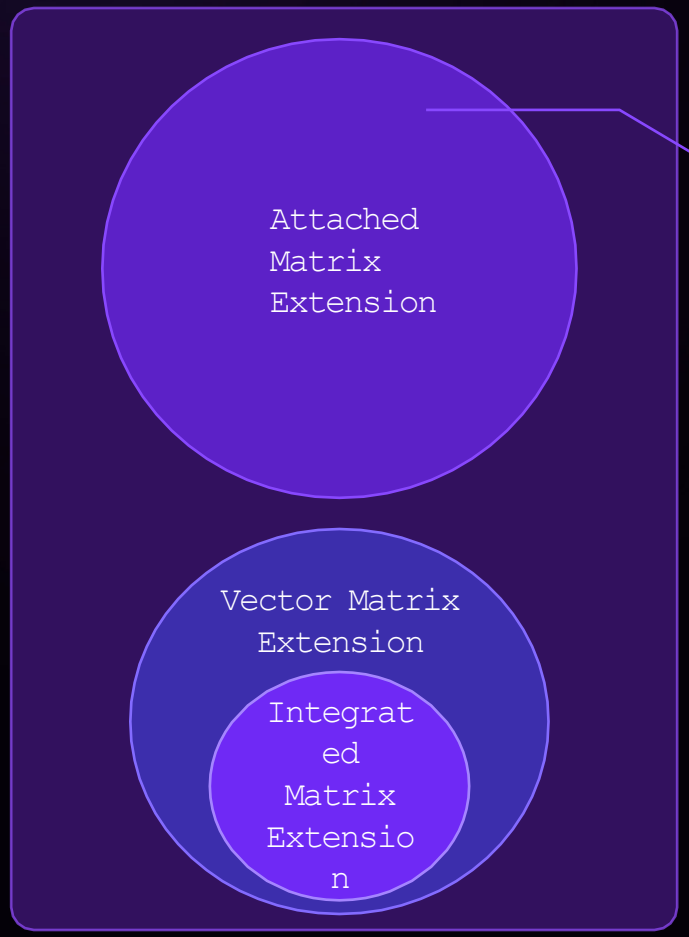
AME扩展的目标

- AME扩展的指令是CPU指令流的一部分
- AME扩展为矩阵运算提供额外的架构状态。即软件可见的矩阵寄存器
- AME扩展提供新增的矩阵和向量运算指令
- AME扩展以AI应用为主要目标，兼顾HPC和嵌入式应用场景

AME与其他扩展的关系

XUANTIE

RISC-V ISA **架构状态**



AME

全新设计的矩阵架构状态

- AME扩展为RISC-V架构设计一种没有受到已有架构设计约束的，以AI应用为主要目标，兼顾HPC和嵌入式的矩阵乘法扩展。
- 由于没有受到已有架构设计的约束，AME具有充分的想象空间：

极致吞吐 | **极高能效** | **极强的扩展性和应用范围** | **完善的应用支持**

IME — **复用已有Vector架构定义**

VME — **复用已有Vector架构定义 + 额外的累加寄存器架构状态**

AME扩展的特点

XUANTIE

AI为重点的全新设计

为矩阵运算量身定制全新的架构定义
针对AI应用重点设计

灵活的集成方式

SoC平台可以自由选择CPU核与AME单元的比例
自由定制算力

更灵活的实现

更少的对微架构设计的假设
更多可能的实现方式

更高的极致性能

独立的运算单元
更多的架构创新，比如更加放松的内存模型

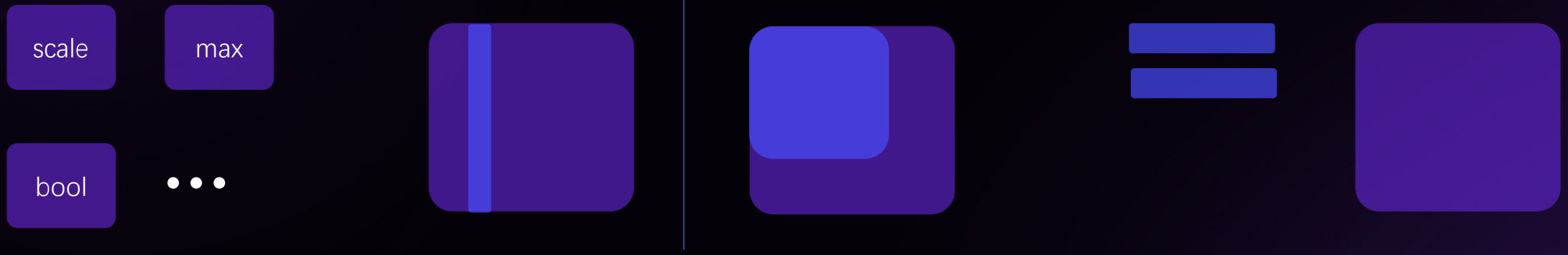
现有AME TG的工作讨论点

- **矩阵**tile**状态设计讨论**
- Point-wise / Element-wise **操作相关的定义讨论**
- Relaxed**内存模型**

矩阵 tile 寄存器的可选项

矩阵tile类型 **矩阵tile形状** **矩阵tile数目**

矩阵 tile 寄存器的可选项

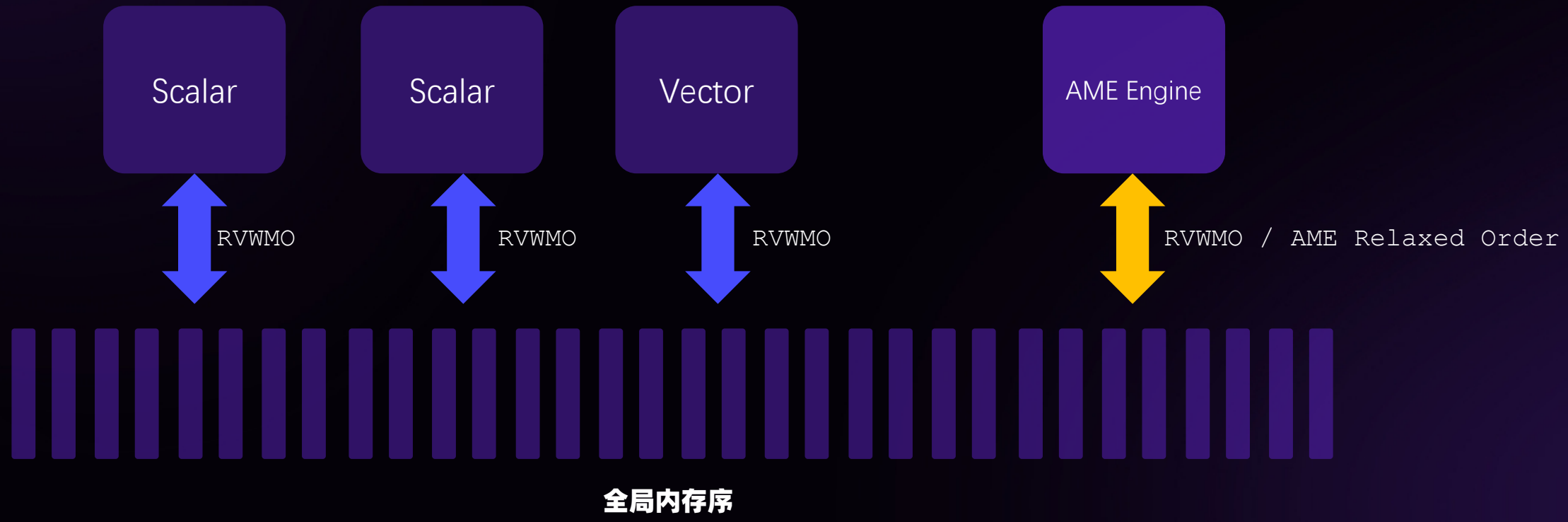


Point-wise / element-wise **操作类型**

如何利用架构状态表示向量

与Vector**扩展如何配合**

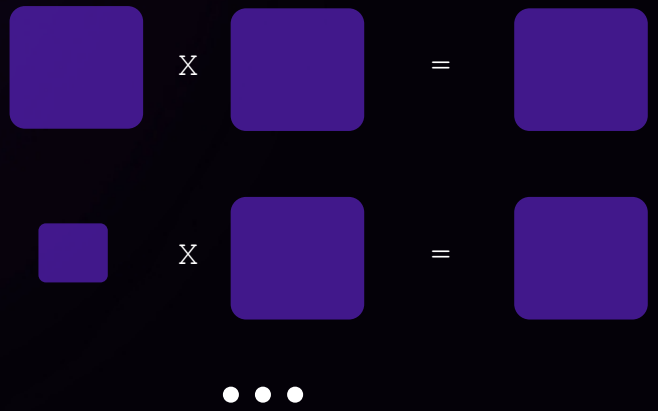
Relaxed内存序



现有提案

XUANTIE

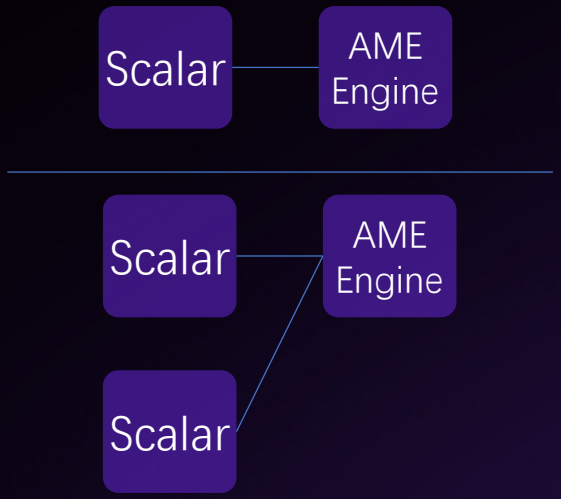
统一的Tile



多种计算指令支持

Bf16
fp8
fp4
Microscaling
format

数据类型支持



灵活的集成方式

玄铁AME张量算力引擎-Tensor Processing Engine

XUANTIE

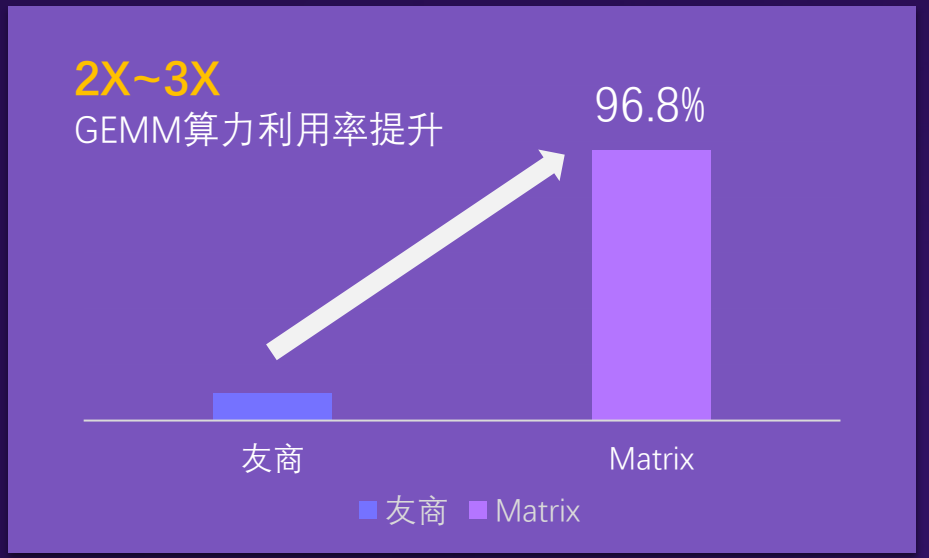
XT-TPE

玄铁全新设计的张量算力引擎

- **TPE支持玄铁自定义AME矩阵扩展指令集**
  - **提供矩阵算力支持**
  - **提供轻量向量算力**(element-wise)**支持**

- **TPE提供大模型计算范式原生支持**
  - **支持**bf16、fp8、fp4、microscaling format**等数据类型**
  - **支持**A16W4、A8W4**等大模型混合精度量化算法加速**
  - TPE**支持可配置峰值算力，单**PE**提供：**

    fp16/bf16: 4TFlops fp8/int8：8TFlops fp4: 16TFlops



玄铁AME张量算力引擎-Tensor Processing Engine

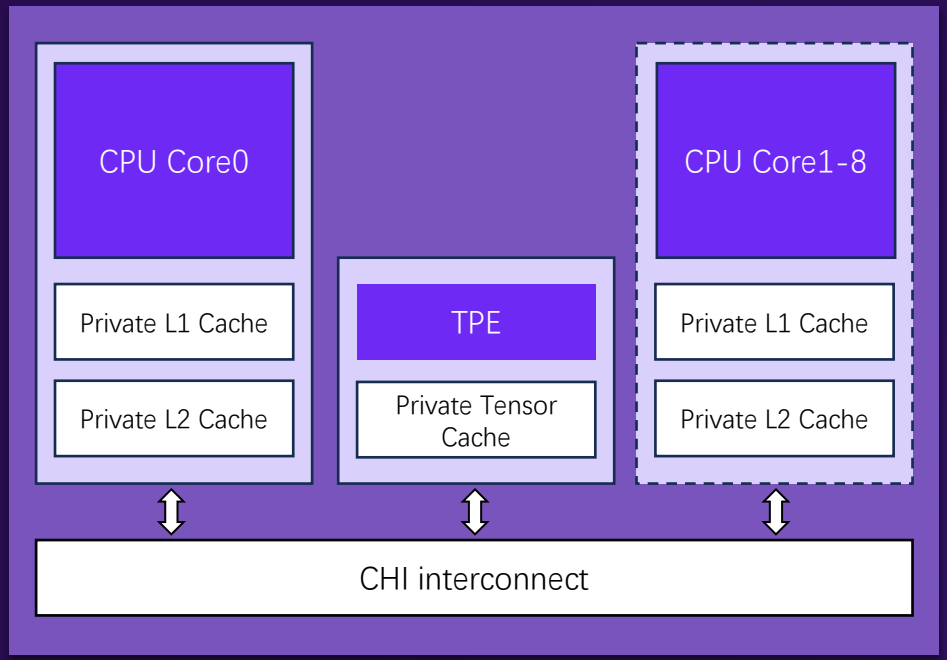
XUANTIE

XT-TPE

玄铁全新设计的张量算力引擎

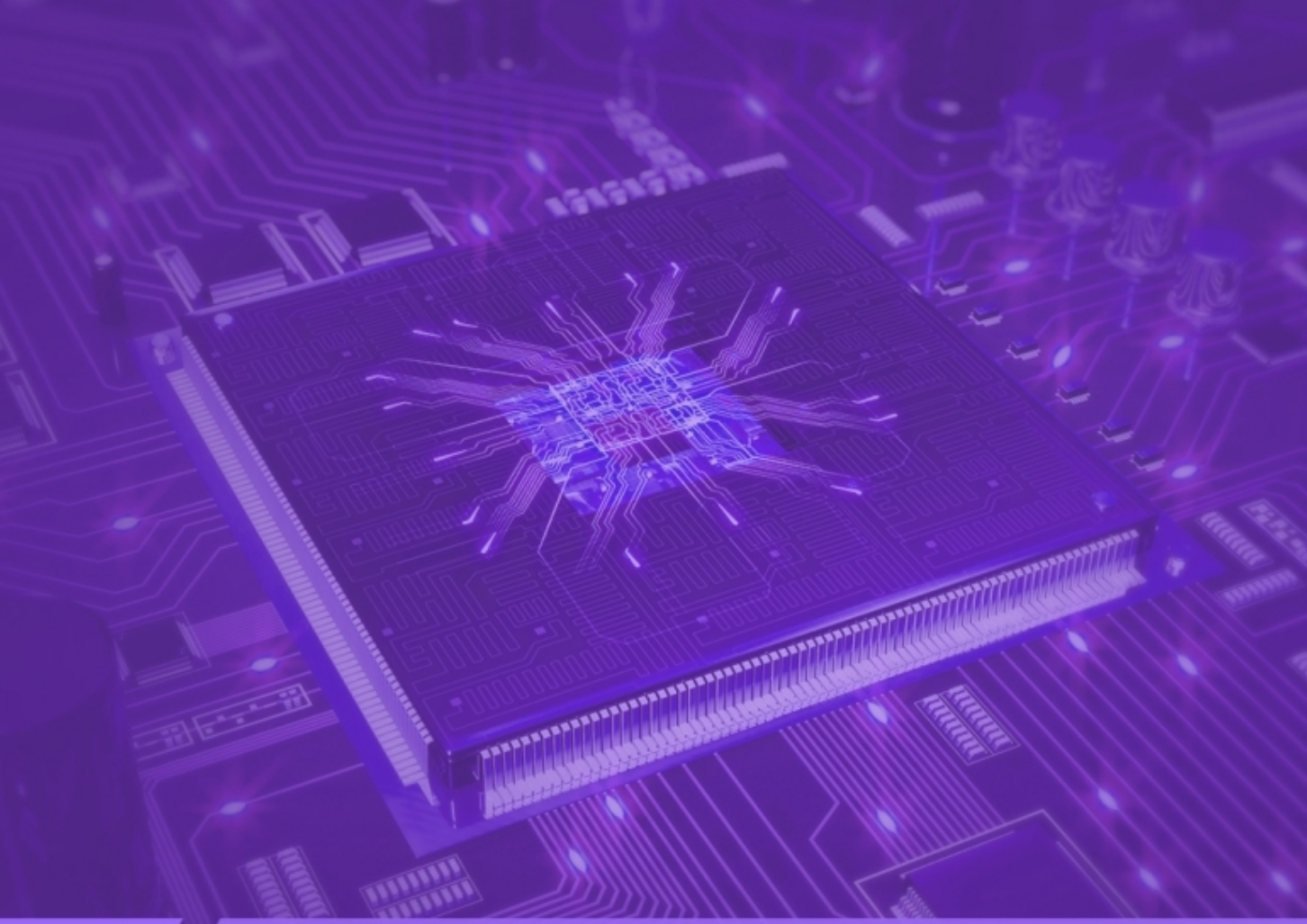
- **TPE与标量流水线解耦**
  - **支持单核私有/多核共享，灵活适配不同应用场景**
  - **支持独立电源域和时钟域**
    高面效（area-efficiency)、高能效（power-efficiency)
  - Matrix-Vector**轻量多线程支持**

- **TPE创新存储架构优化，强力提升AI算力兑付率**
  - **支持与标量访存解耦的矩阵访存架构**
  - **支持私有**Tensor Cache，**提供跨**Cacheline**高带宽访存能力**
  - **硬件维护**Cache**一致性，遵循**RVWMO**访存序模型**
  - **高效硬件预取策略，支撑跨**Loop**及时、准确的数据预取**
  - **提供**CHI**一致性内存访问接口**



XUANTIE

Thank You



玄铁公众号

玄铁官网

公司地址：浙江省杭州市余杭区阿里巴巴西溪园区C区
如需了解更多，可以咨询您的【玄铁专家】
您也可前往玄铁官网：xrvm.cn 或通过 xuantie@service.alibaba.com ，
riscv_techsales@service.alibaba.com 与我们联系