XUANTIE

2024 第4届 RISC-V 中国峰会

面向 AI 应用的玄铁扩展指令

仇径

阿里巴巴达摩院

XUANTIE

目录

Contents

01 向量 AI 扩展

SFU 扩展指令集

02 矩阵 AI 扩展

Attached Matrix 扩展指令集

03 总结

展望和鸣谢

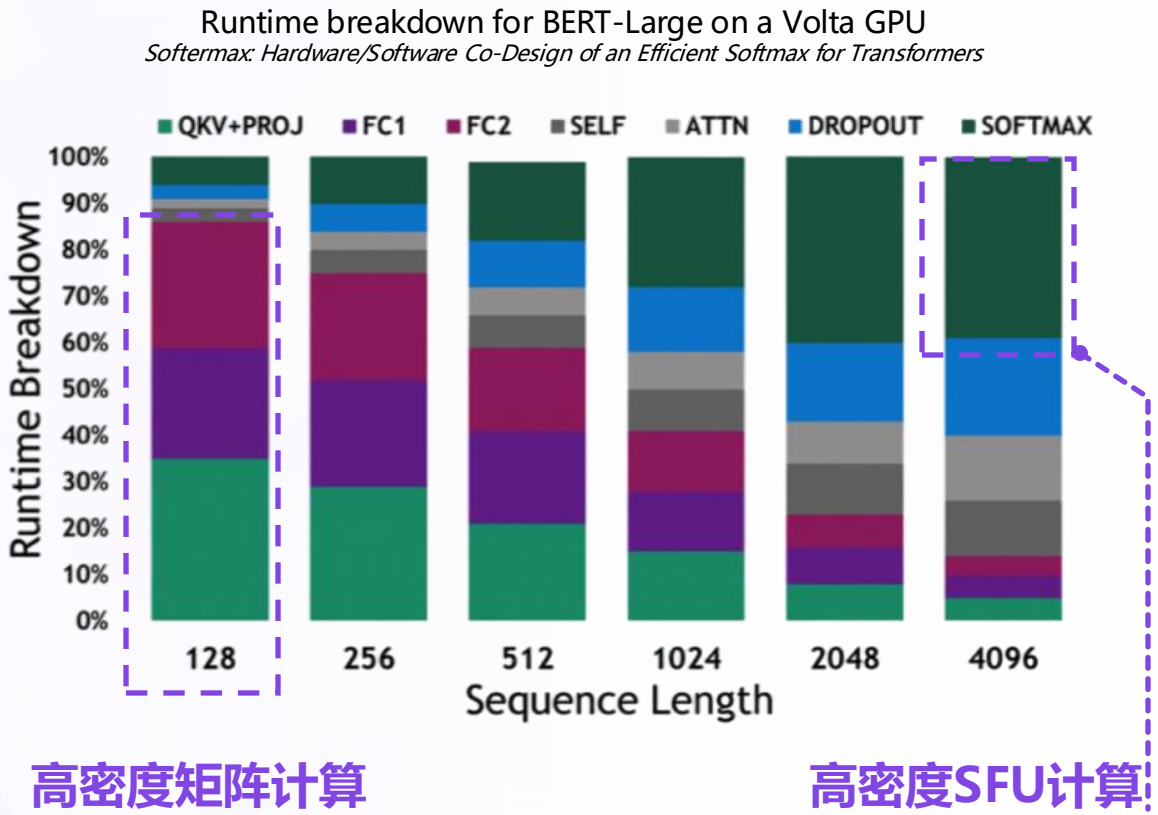
XUANTIE

01 向量 AI 扩展

SFU 扩展指令集 (Vector Based)

向量 SFU 扩展

XUANTIE



面向 AI 应用的 SFU 扩展

- Exponential function
- Reciprocal function
- Tangent function
- Sigmoid function
- SoftMax function
- Gelu function

现有实现方式

XUANTIE

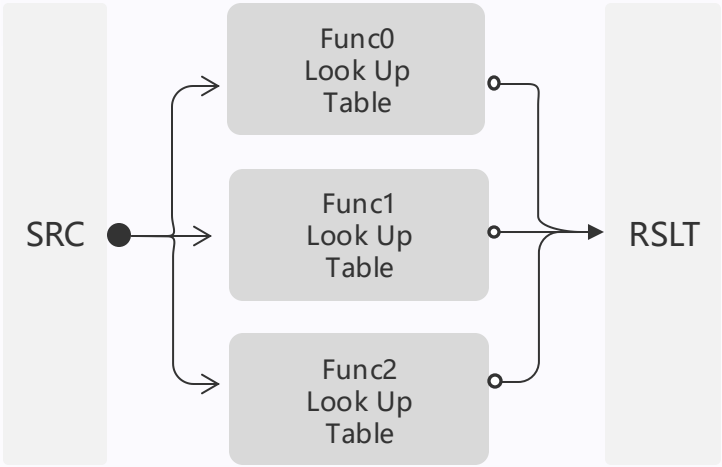
快速算法实现

```
static inline float fast_exp32(float y)
{
    union {
        float d;
        unsigned int x;
    } data = {y};

    data.x = (12102203 * y + 1064866816);

    return data.d;
}
```

Schraudolph N N. A Fast, Compact Approximation of the Exponential Function[J].

计算速度快

查找表实现



Lookup Table Example

RISC-V Vector Instruction

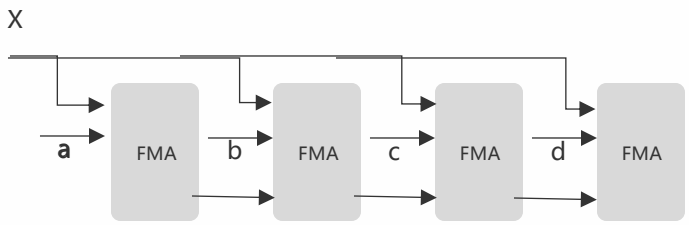
vfrec7.v

vfrsqrt7.v

实现简单

多阶泰勒展开

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots \quad \forall x$$



- ✓ 不同的展开阶数对应不同的精度需求
- ✓ 通过 FMA 的并行计算，可以得到较低的计算延迟
- ✓ FMA 阵列可以复用
- ✓ 预处理和后处理可以复用

精度实现灵活

分段多阶泰勒展开

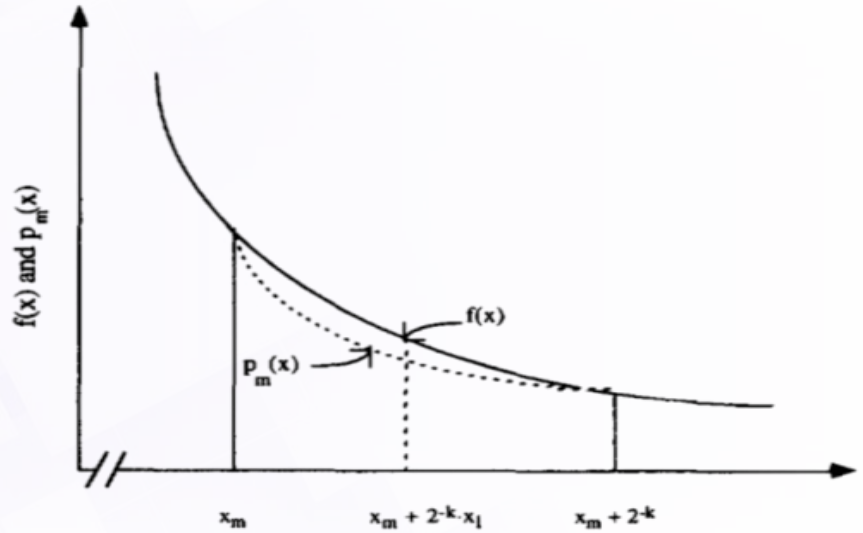
XUANTIE

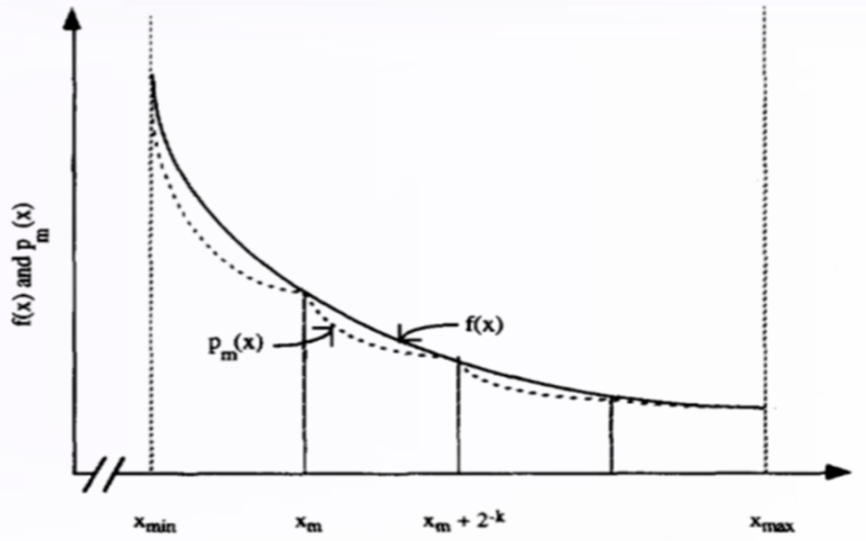
分段使用不同的展开阶数和参数 → 平衡精度和计算复杂度

使用查找表存储参数 → 加速计算

特殊值段特殊映射 → 高效计算和资源复用

输入范围	Tanh 的具体实现
特殊值	特殊值
[-inf, 0]	-tanh(-x)，x位于[0, +inf]
[0，2^(-10)]	x
[2^(-10), 2^(-1)]	分段泰勒展开实现，分为9段
[0.5, 4]	tanh(x) = 2 * sigmoid(2 * x) - 1
[4, +inf]	1





向量SFU扩展指令集

XUANTIE

遵循 RVV 扩展规则 玄铁自定义扩展指令规则

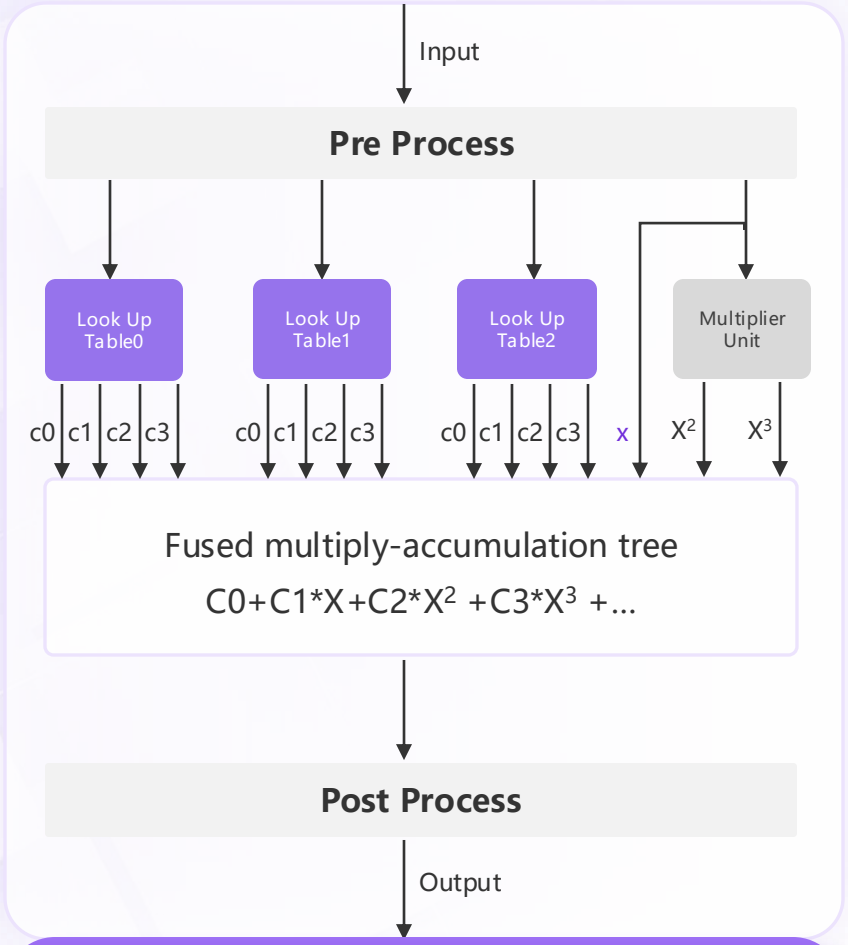
支持
- FP32 的计算
- mask 操作
- vta/vma 配置
- LMUL 操作

简化设计 不支持rounding模式设置，无非精确异常及下溢异常

指令	定义	助记符
vfexp2	Output vector 2^x of the input vector x	vfexp2.v vd, vs2, vm
vftanh	Output vector tanh(x) of the input vector x	vftanh.v vd, vs2, vm
vfsig	Output vector sigmoid(x) of the input vector x	vfsig.v vd, vs2, vm
vfrec	Output vector 1/x of the input vector x	vfrec.v vd, vs2, vm

硬件实现

XUANTIE



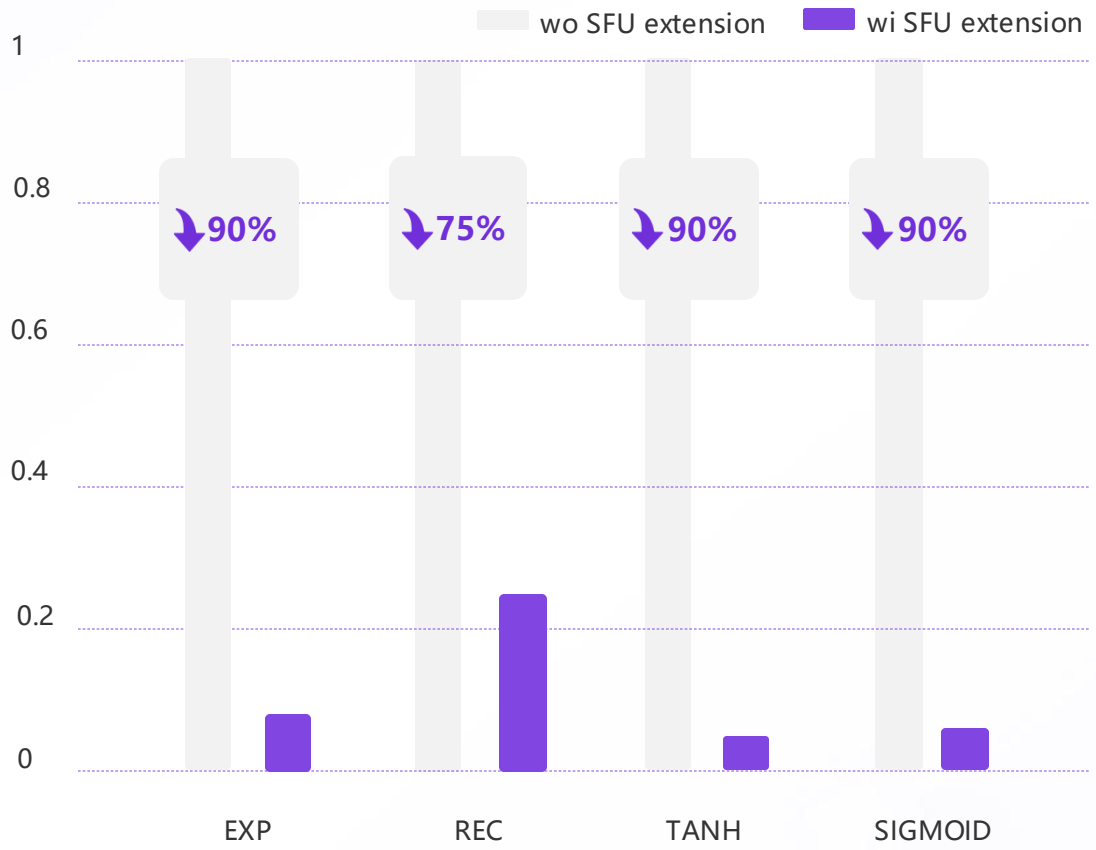
Special Function Unit

性能提升

算力弹性

可变 VLEN

灵活实现



Acceleration effect of different functions with Xuantie SFU vector extension instructions

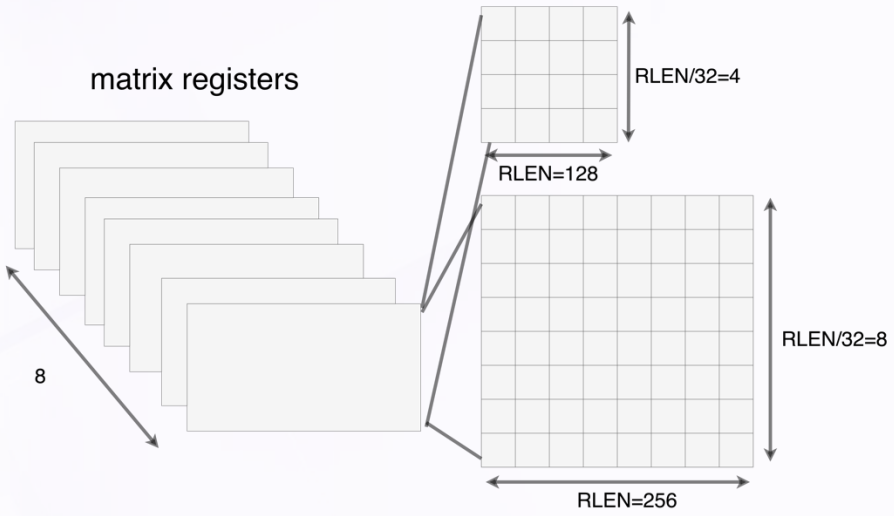
数据采用最高性能配置，SFU 单元面积占用小于 vector 标准模块的 15%

XUANTIE

02 矩阵 AI 扩展

Attached Matrix 扩展指令集

玄铁矩阵扩展



AI Domain Specific

Small set of matrix multiplication instructions
AI/ML data types int4/int8/bf16/fp16 and so on
Multi-precision and mixed-precision computation

Scalability

RLEN scales from 128 to 1024+ , Peak performance from 0.128 Tops to 32 Tops
Binary portability, no recompile or rewrite for different RLEN
Various matrix shapes

Attached Facility

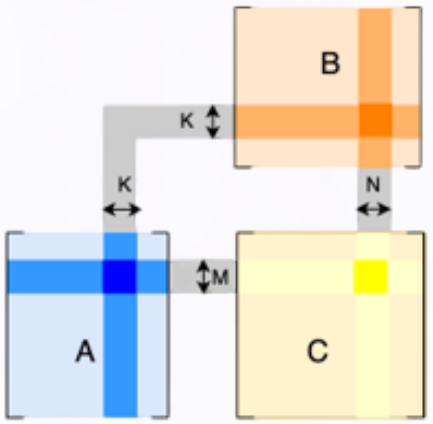
Decoupled from vector extension at programming model level

Extensibility for Future

Future data types(fp4 fp8 binary)
Other matrix operations and features (pointwise sparsity)

矩阵扩展ISA

XUANTIE

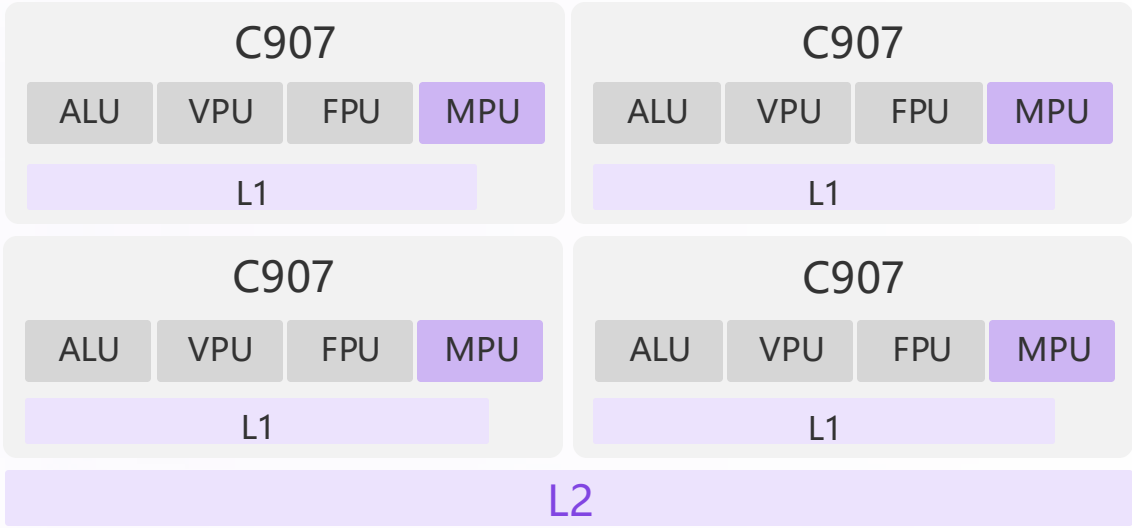


Matrix MACC	fmmacc.<h/s/d> fwmmacc.<h/s> mmacc.<s/u>.<s/u>.<b/h>	Floating point matrix multiply and accumulate Floating point matrix multiply and accumulate(widen) Integer matrix multiply and accumulate(4x widen)
Memory Access	mld.<b/h/w/d> mst.<b/h/w/d> mldm/mstm	Matrix load to matrix registers Matrix store from matrix registers Load/store whole matrix register
Matrix Operations	madd/msub.<s/d>.<mm/mv/mx>.<x/i> mshift.<s/d>.<mm/mv/mx>.<x/i> mn4clip.<s/d>.<mm/mv/mx>.<x/i> mmul.<s/d>.<mm/mv/mx>.<x/i>	Matrix add/sub matrix/vector/scalar Matrix shift matrix/vector/scalar Matrix clip Matrix product matrix/vector/scalar
Move	mmov.mm/mmov.mv.x/mmov.mv.i mmov<b/h/w/d>.x.m mdup<b/h/w/d>.m.x/mmov<b/h/w/d>.m.x	Move between matrix registers Move form matrix register to scalar register Move form scalar register to matrix register
Matrix Config	mcfgi/mcfg	Matrix tail configuration
Others	release zero	Release matrix register Zero matrix register

矩阵扩展硬件实现

XUANTIE

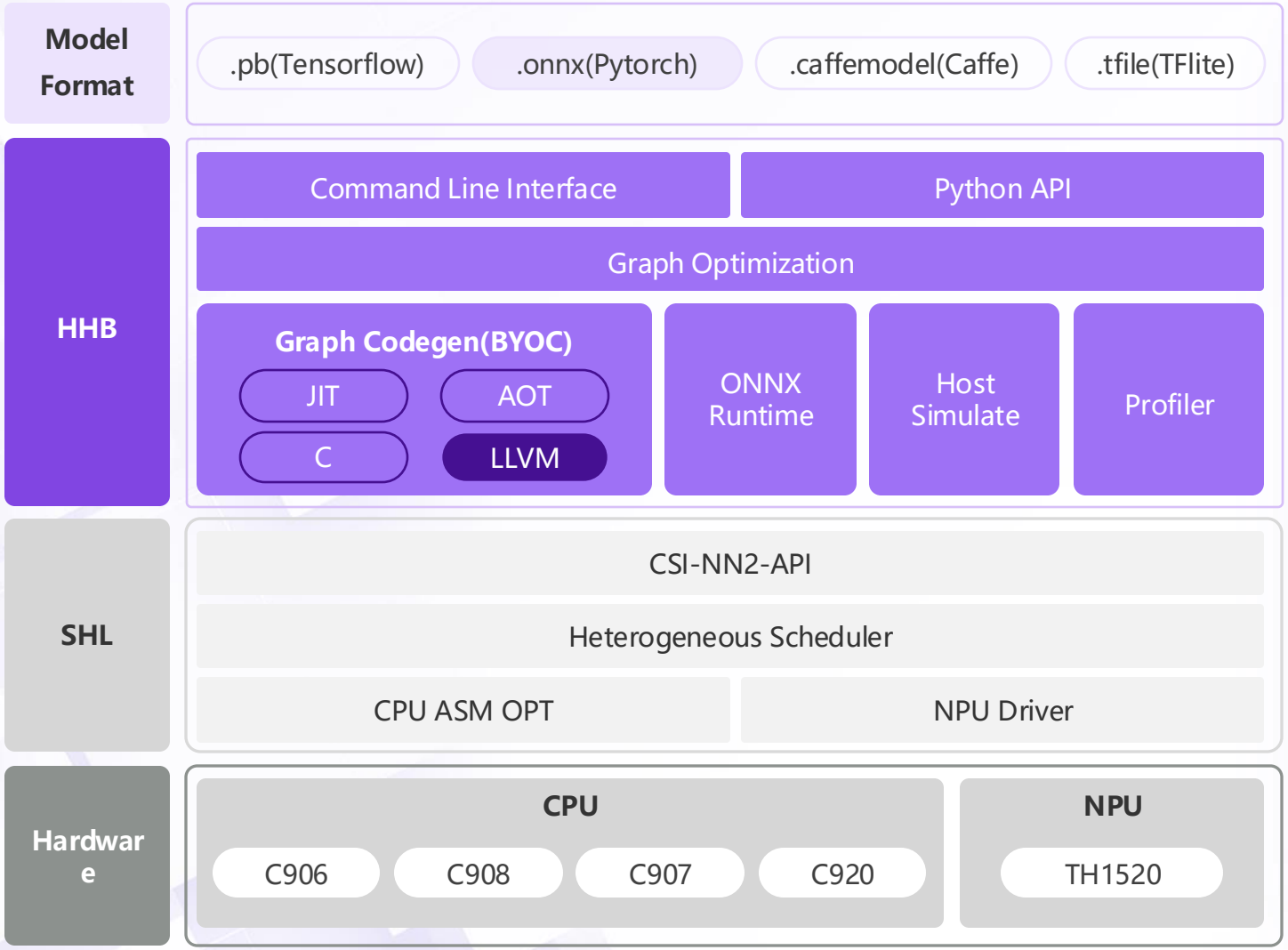
- 支持 Attached Matrix Extension ISA
- 独立Matrix执行单元，和向量单元并行，共用片上2级Cache系统，协同加速AI典型应用
- 矩阵大小可伸缩，MLEN 512～2048
- 支持多种数据类型，INT8/INT4/FP16/BF16
- 乘累加单元脉动阵列设计，增强AI核心运算能效比
- 浮点运算和整型运算资源复用，提升资源利用率
- 存储层次结构优化，先进预取算法，满足算法访存带宽需求



可配置　Peak performance（Gops/Ghz）

RLEN	int4	int8	fp16/bf16
128	256	128	64
256(H)	512	256	128
256	1024	512	256

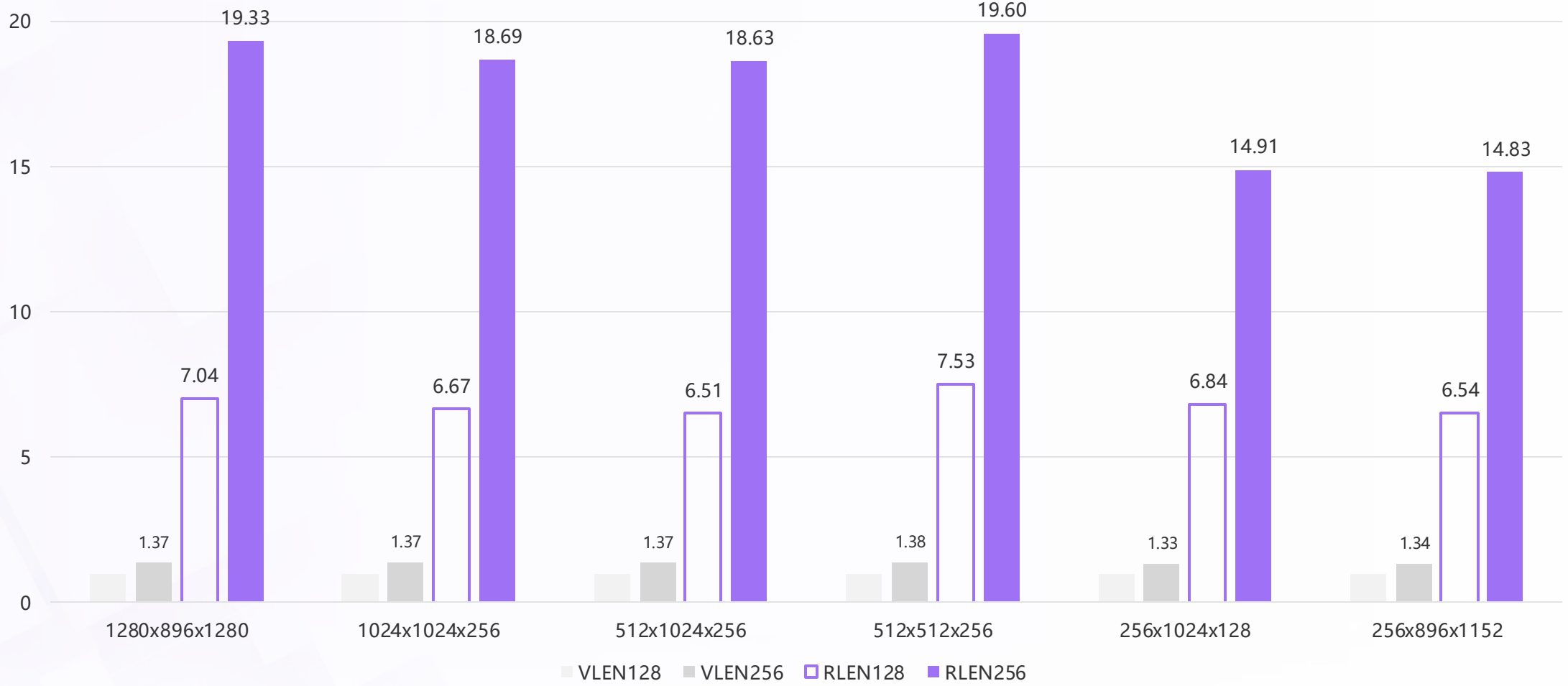
矩阵扩展软件实现



- Dynamic Shape 原生支持动态 shape
- Unified IR abstraction 统一IR抽象，适配不同的部署机制
- Meta-schedule 支持算子自动调优
- 支持 RVV/RVM 指令自动生成
- 仿真器/工具链更新

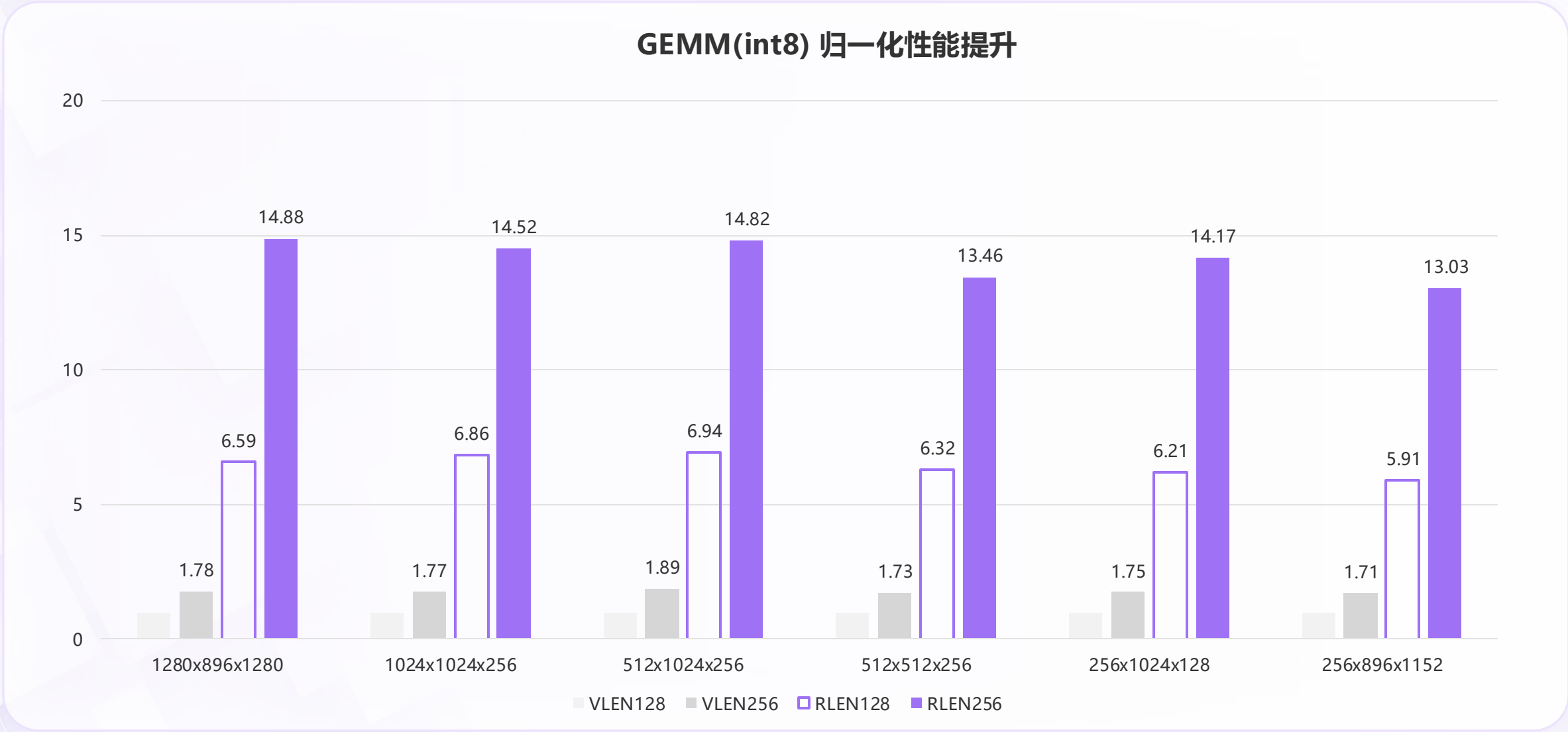
GEMM 性能优化——Based on C907

XUANTIE



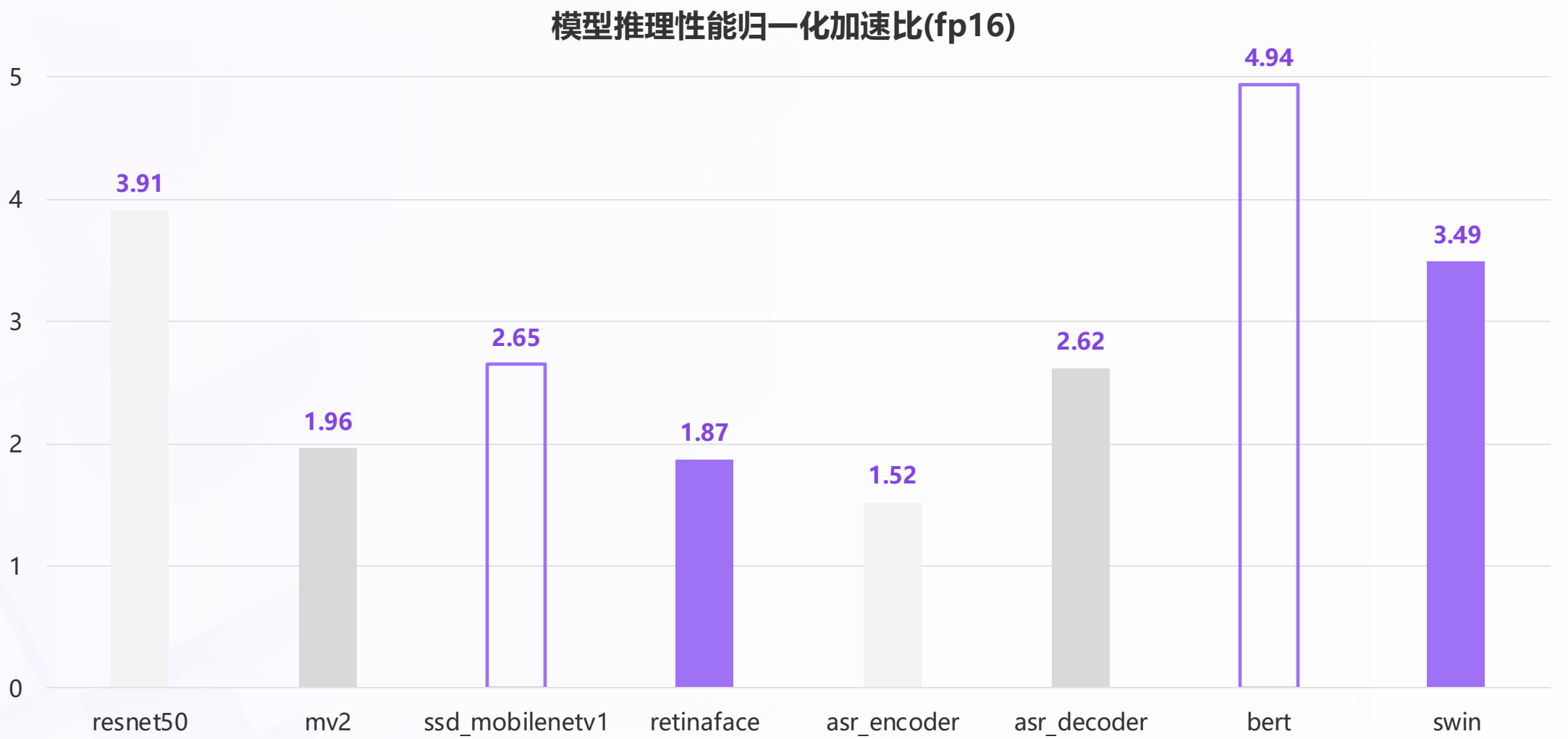
GEMM 性能优化——Based on C907

XUANTIE



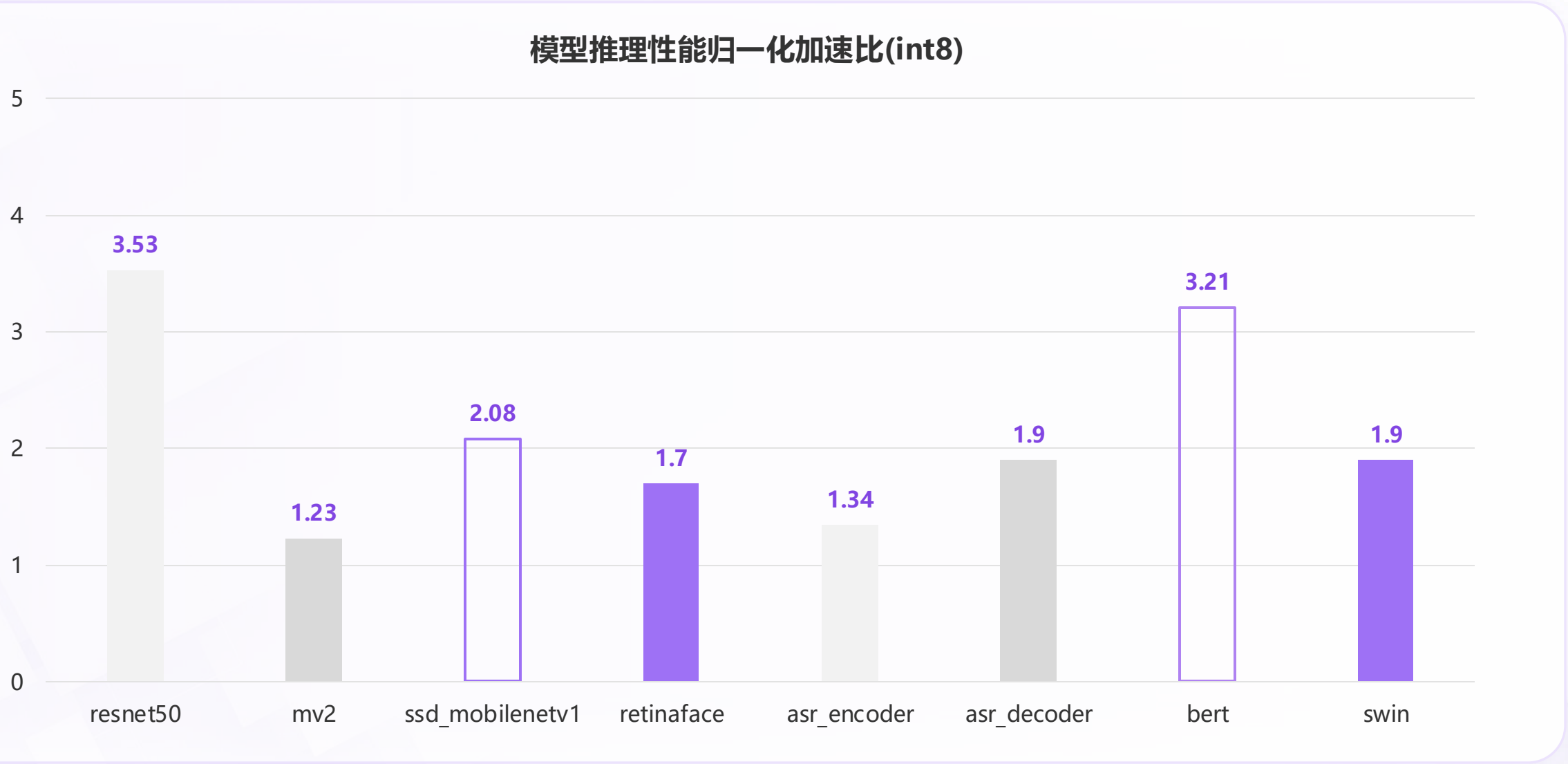
推理性能优化——Based on C907

XUANTIE



推理性能优化——Based on C907

XUANTIE



Matrix ISA Update Preview

XUANTIE

长指令编码格式

- 操作数信息增加
- 寄存器类型增加
- 长立即数增加
- 寄存器数目增加
- 减少配置操作
- 提升指令性能

稀疏操作支持

- 结构化稀疏
- 增加索引方式
- 提升运算算力
- 优化内存访问

数据格式支持

- 新数据类型支持
- 混合精度支持

基础特性更新

- 灵活矩阵尺寸
- 运算模式更新
- LLM 支持
- 配置方式更新
- 基础算子更新

XUANTIE

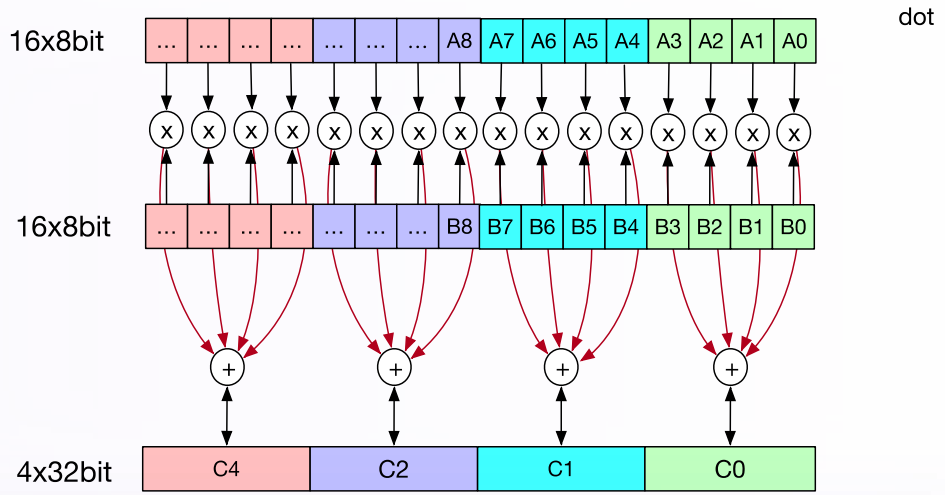
03 总结

展望和鸣谢

面向AI应用的RISC-V指令

XUANTIE

- RISC-V Vector Extension
- XuanTie AI dot turbo Extension
- XuanTie Vector SFU Extension
- XuanTie Attached Matrix Extension



XuanTie Cores	RVV Vector
XuanTie Cores	RVV Vector Attached Matrix Extension
XuanTie Cores	RVV Vector Vector SFU Extension
XuanTie Cores	RVV Vector AI dot Turbo Vector SFU Extension Attached Matrix Extension

合作展望

XUANTIE

	2024	2025
RISC-V Attached Matrix Task Group	● 发起创立，Acting Chair仇径 ● 完成 Workload 分析	● 发布 Preview Specification ● 完成 POC 和Freeze
RACE matrix 指令集工作组	● 发起创立，副组长仇径 ● 完成 matrix 指令集草案	● 合作伙伴完成硅验证
工委会指令集工作组	● 发起创立 ● 指令集工作部部长单位 ● 完成 matrix 指令集标准制定	● 标准化 matrix 指令集扩展

鸣谢

XUANTIE

**诚挚感谢，排名不分先后*

开芯院异构系统组

上海交通大学
数字与系统芯片研究设计中心

复旦大学微电子学院

RISC-V工委会指令集工作部

中科院计算所先进中心
中科院软件研究所

RISC-V+AI算力生态委员会
(RACE)

重庆大学集成电路学院

All GitHub contributors

关心 Matrix 开源指令集的朋友

XUANTIE

2024 第4届 RISC-V 中国峰会

Thank you

玄铁公众号

玄铁中文站

玄铁海外站