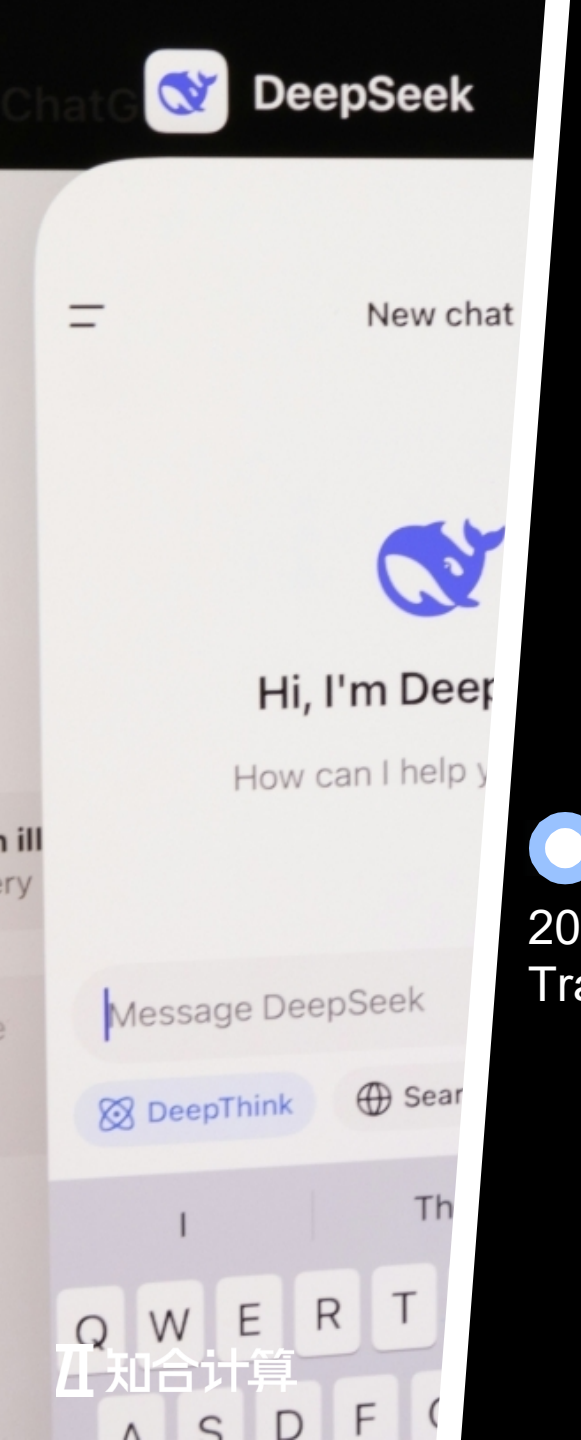
知合计算 ZHIHE COMPUTING | MAKE COMPUTING MORE EFFICIENT 让计算更高效

大模型在RISC-V架构上的技术创新与应用

知合计算 解决方案总监

黄怡皓



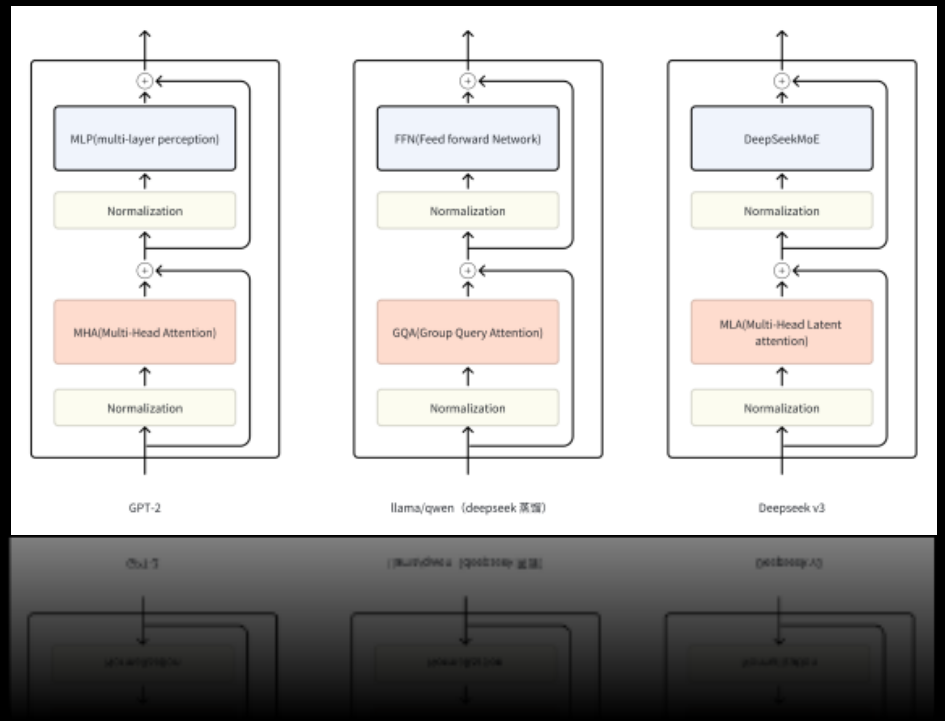
大模型持续创新

底层架构仍是Transformer

- 2017年6月 Transformer
- 2020年5月 GPT-3
- 2022年11月 ChatGPT
- 2023年2月 LLaMA
- 2023年12月 Qwen2
- 2025年1月 DeepSeek-R1

模型百花齐放

核心算子逐渐趋同



模型	GPT-2	LLaMA/Qwen	DeepSeek-v3
FFN 结构	MLP（由两个矩阵乘组成）	FFN（由3 个矩阵乘组成）	MOE（3 个矩阵乘为一组，选择性的激活若干组）
Attention 结构	MHA（基础的多头注意力机制）	GQA（按组共享的注意力机制，优化 kvcache 大小）	MLA（低秩分解的注意力机制，目的也是优化 kvcache 大小）
代码行数	200	500	800
算子种类	22	21	21

模型算力

需求集中

DeepSeek 7B 模型中，

核心算子共**11个**

Matmul 计算量

占比约 **95%**

算子	说明
embedding	词嵌入
rms_norm	均方根归一化
matmul	**矩阵乘**
reshape	调整张量形状
rope	位置编码相关
llm_pos	存cache、取cache
transpose	转置
mul	乘法
softmax	softmax
add	加法
silu	激活函数

RISC-V助推

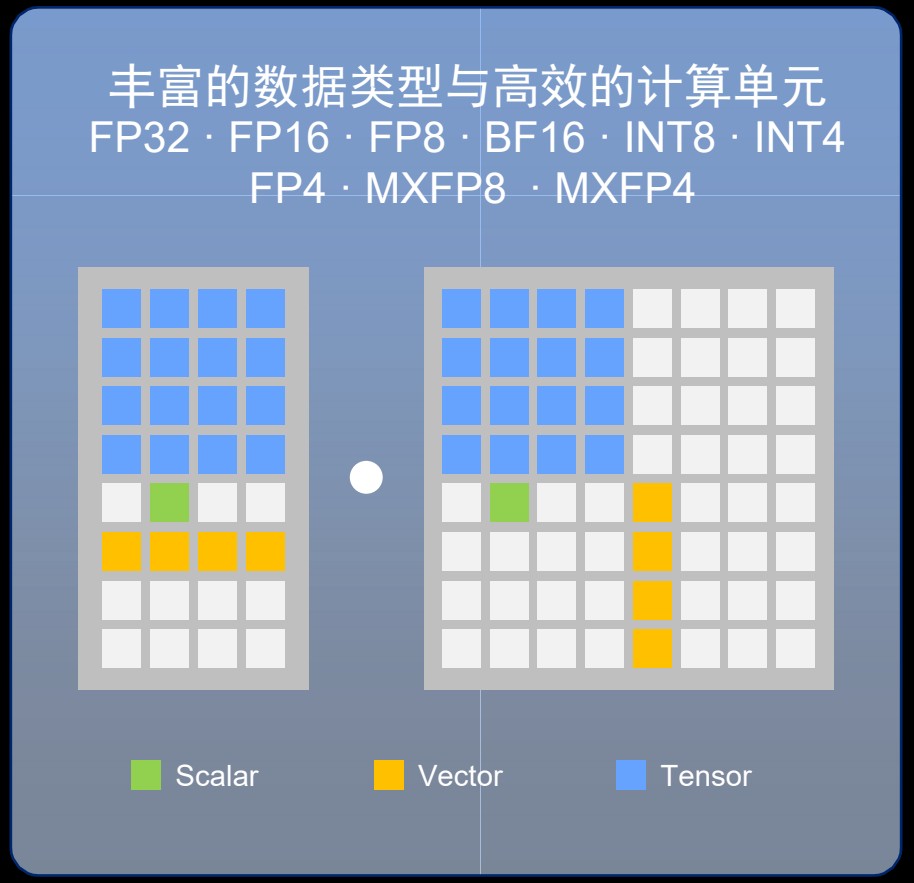
统一计算平台

AME 指令

完美适配 **Matmul** 算子

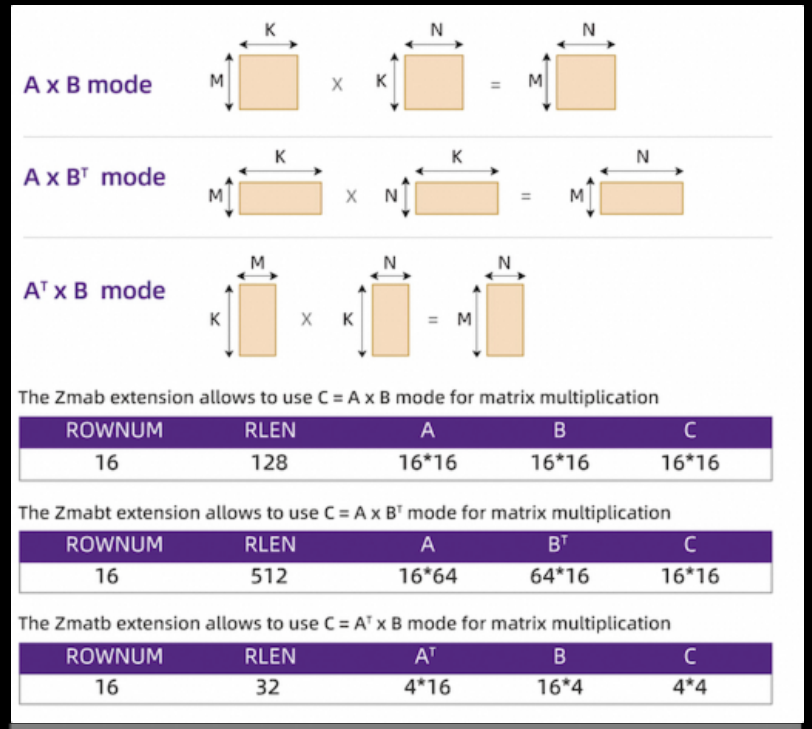
除 3 个算子外，均可采用AME优化

支持全面的 AI 数据格式



指令集创新

RISC-V AME指令介绍



The Zmab extension allows to use C = A x B mode for matrix multiplication

ROWNUM	RLEN	A	B	C
16	128	16*16	16*16	16*16

The Zmabt extension allows to use C = A x Bᵀ mode for matrix multiplication

ROWNUM	RLEN	A	Bᵀ	C
16	512	16*64	64*16	16*16

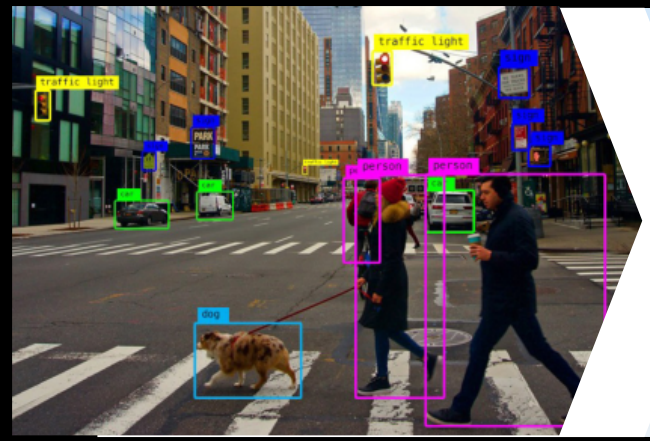
The Zmatb extension allows to use C = Aᵀ x B mode for matrix multiplication

ROWNUM	RLEN	Aᵀ	B	C
16	32	4*16	16*4	4*4

category	instructions	Operand Type	Accumulator Type
Float	mfmacc.h	fp16	fp16
Float	mfmacc.s	fp32	fp32
Float	mfmacc.d	fp64	fp64
Float	mfmacc.s.h	fp16	fp32
Float	mfmacc.d.s	fp32	fp64
Float	mfmacc.s.bf16	bf16	fp32
Float	mfmacc.<h/bf16/s>.<e4/e5>	fp8(e4m3/e5m2)	fp16/bf16/fp32
Int	mmacc.w.b	int8	int32
Int	mmaccu.w.b	uint8	int32
Int	mmaccsu.w.b	(su)int8	int32
Int	mmaccus.w.b	(us)int8	int32
Int	mmacc.w.hb	int4	int32
Int	mmaccu.w.hb	uint4	int32
Int	mmaccsu.w.hb	(su)int4	int32
Int	mmaccus.w.hb	(us)int4	int32

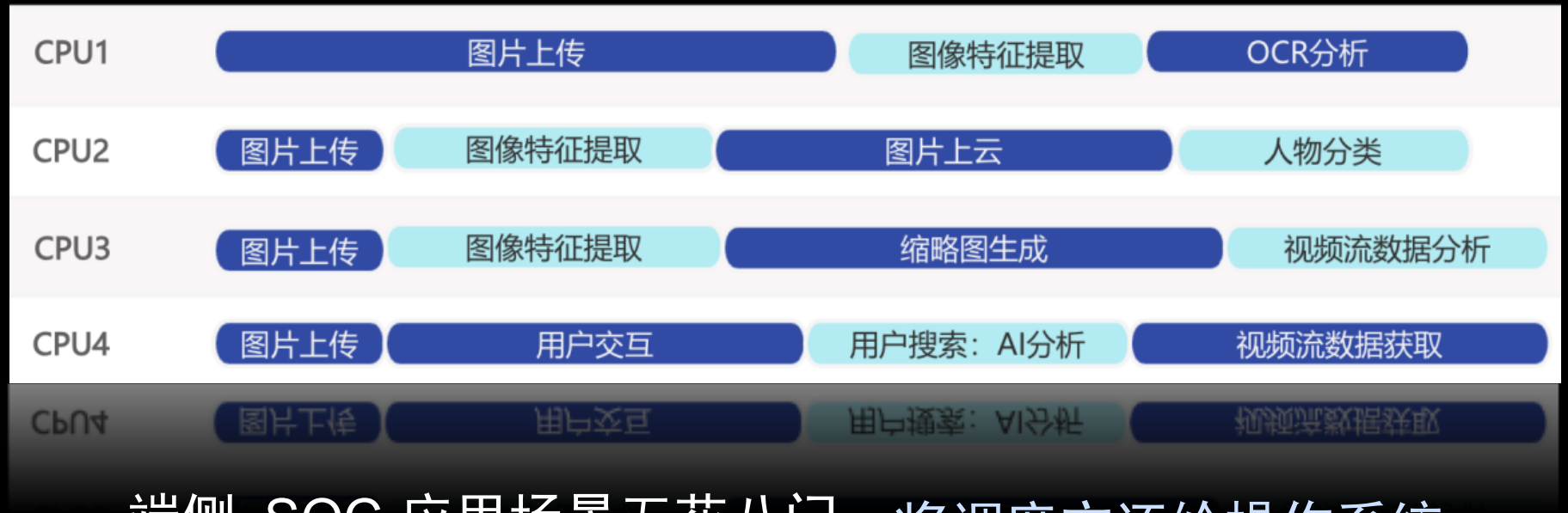
AI应用发展趋势

从识别走向认知



通推一体

实现通用计算与AI增强的高效融合



端侧 SOC 应用场景五花八门，将调度交还给操作系统

“通推一体”CPU产品A210

已可尝鲜

阿基米德系列

通推一体CPU

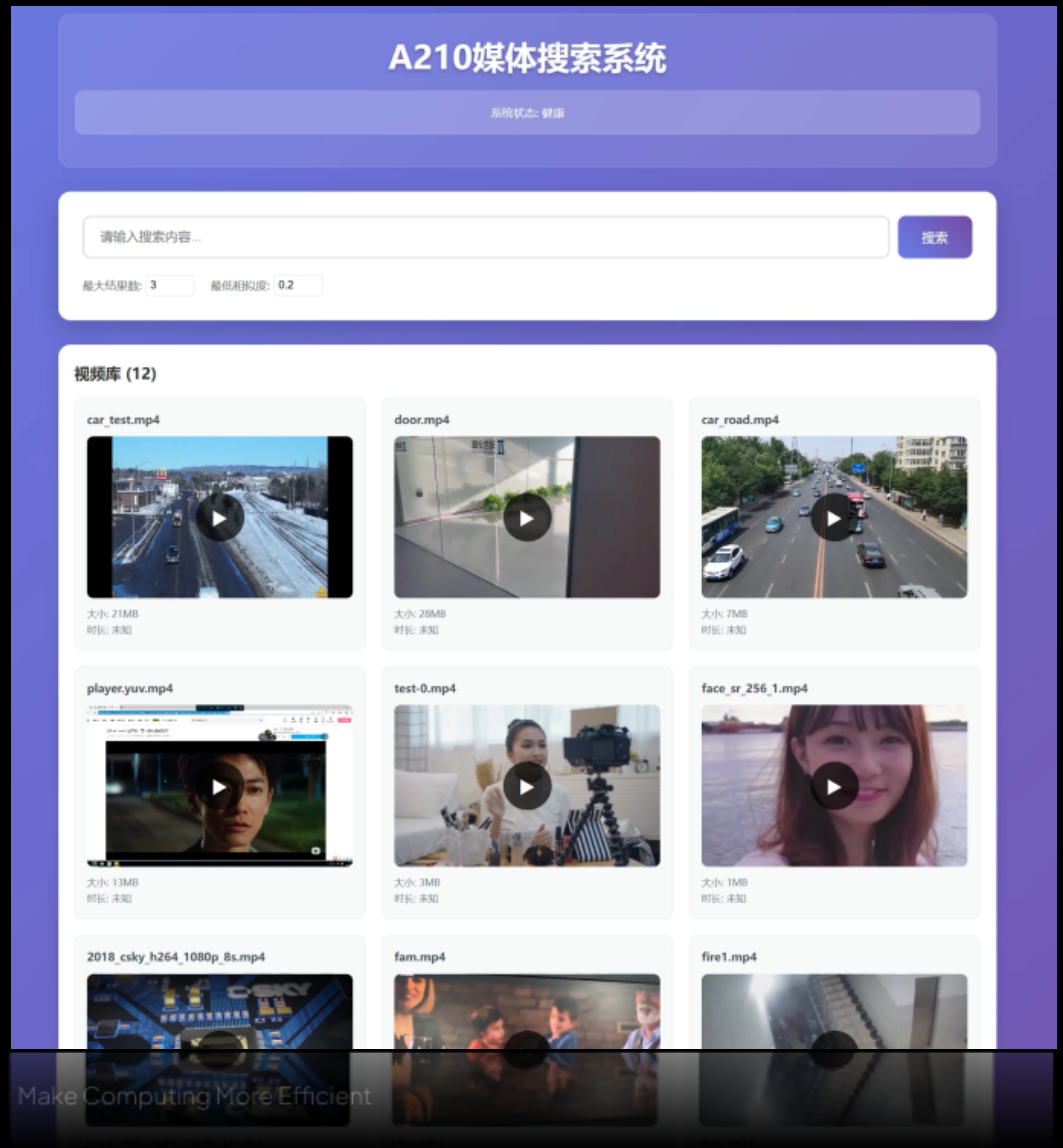
· UCA统一计算架构 ·
· 统一内存 ·
· 统一算子 ·

8核CPU · 12 TOPS算力

敬请期待

A210应用场景案例 点餐系统

A210应用场景案例 模糊搜索



通用计算需要高效

AI推理更需要高效

- 大模型算子统一，为RISC-V提供“生态红利”
- 开放架构助推算子优化实现
- AI能力跃迁：从“识别”走向“认知”
- 通用计算与AI计算开始融合

知合计算 ZHIHE COMPUTING
MAKE COMPUTING MORE EFFICIENT
让计算更高效

谢谢聆听