StarFive
赛昉科技

乱序 RVV：动态调度提升AI 计算任务效率

崔进

2025/7/18

StarFive 赛昉科技

Agenda

1. AI计算的新挑战
2. RVV在AI计算中的优势
3. 乱序RVV在AI计算中的优势
4. 典型计算任务性能实验
5. 乱序RVV核实例介绍

AI计算的新挑战

当前，人工智能正经历从专用模型向通用大模型、从云端向边缘的双重演进。这一过程对计算架构提出了三大挑战：

- **模型多样性**：从 CNN、Transformer 到新兴的多模态模型，计算模式差异显著
- **部署碎片化**：从超低功耗 IoT 设备到高性能数据中心，硬件需求千差万别
- **生态封闭性**：传统 AI 加速方案依赖专有架构，导致开发成本高、迁移困难

RVV在AI计算中的优势

作为首个真正开放的向量指令集标准，RVV 具有两大核心优势：

参数化设计

向量长度（VLEN）、寄存器组大小等均可配置，使其能够高效支持不同规模的 AI 计算

指令集兼容性

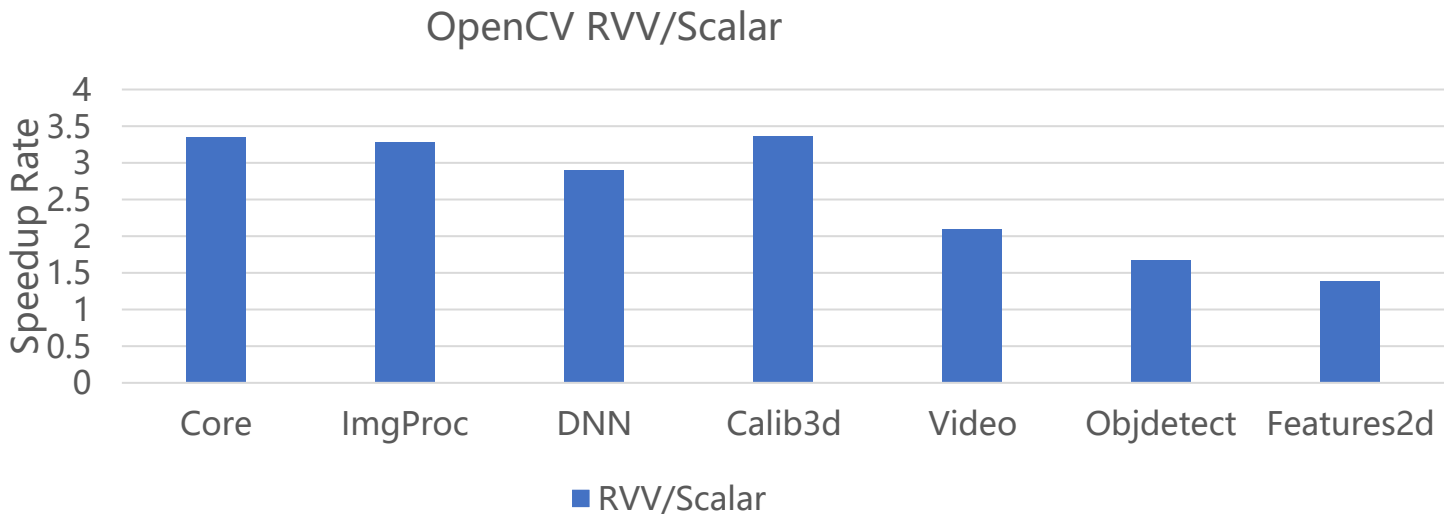
可以在不同的硬件平台上运行同样的软件，极大的减少了软件移植等开销，对于DSA亦是如此

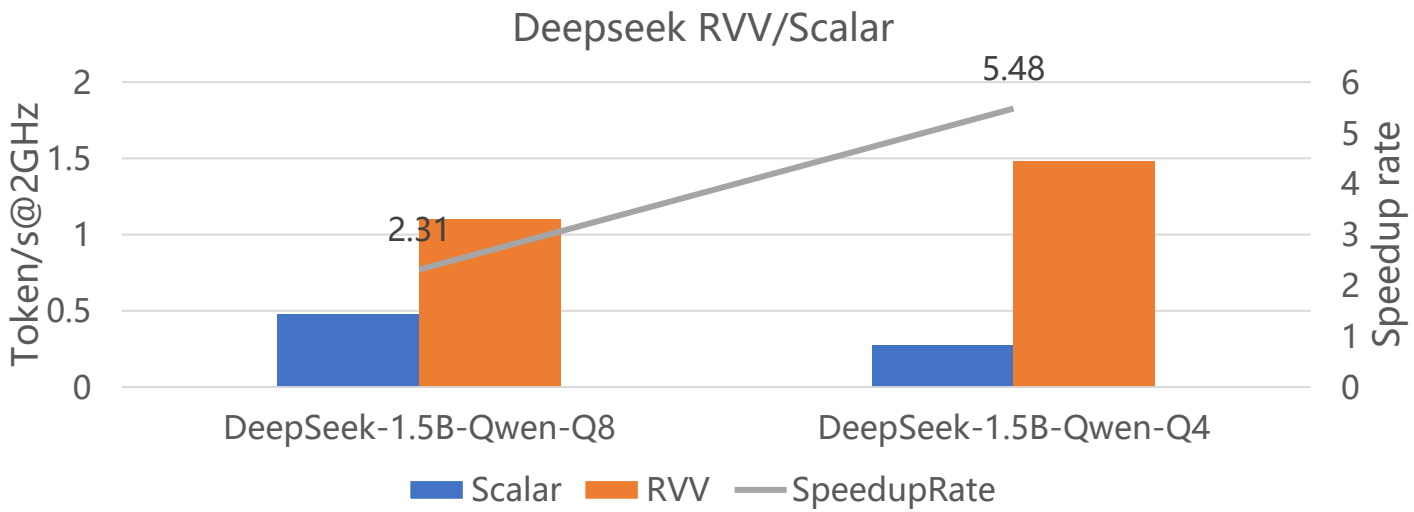
RVV在AI计算中的优势

RVV加速比实验

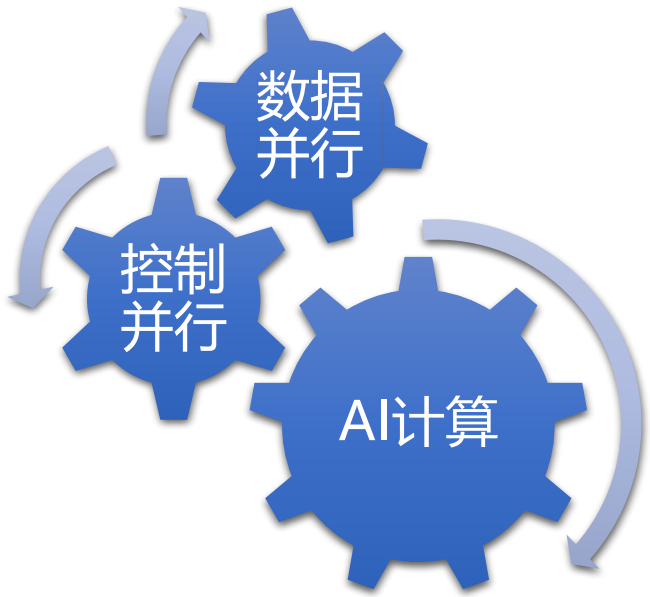
RVV对OpenCV的平均加速2.6倍

RVV对Deepseek-1.5B-Qwen-Q8和Q4分别加速2.3，5.5倍





乱序RVV在AI计算中的优势



AI 计算的本质是**数据并行和控制并行**的混合负载，乱序RVV的乱序执行通过**动态调度**将两者解耦，突破顺序执行的 “指令墙” 和 “内存墙”。

乱序适用场景

高指令级并行场景	· 如矩阵运算密集的 Transformer 层
内存访问密集型任务	• 如参数服务器架构的分布式训练
动态控制流场景	· 如强化学习决策、自适应推理

乱序RVV在AI计算中的优势

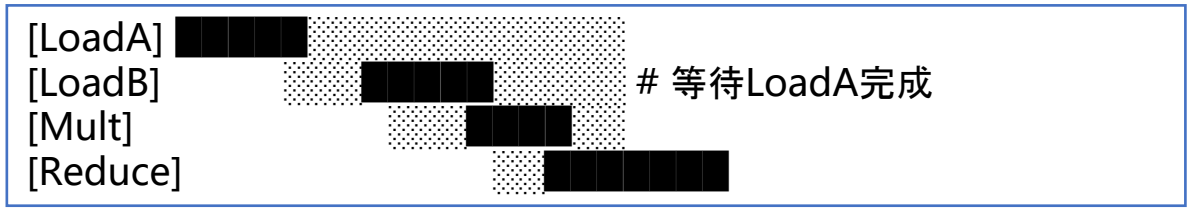
- 指令级并行
  - 非阻塞内存访问
  - 寄存器重命名
  - 乱序发射窗
- 内存访问效率提升
  - 隐藏延迟
  - 优化内存
- 控制流与资源调度优化
  - 分支预测错误率降低
  - 动态资源分配提升利用率
  - 能效比提升

计算任务：向量点积C = Σ(A[i] * B[i])

乱序RVV减少37%气泡 → 减少25%总延迟

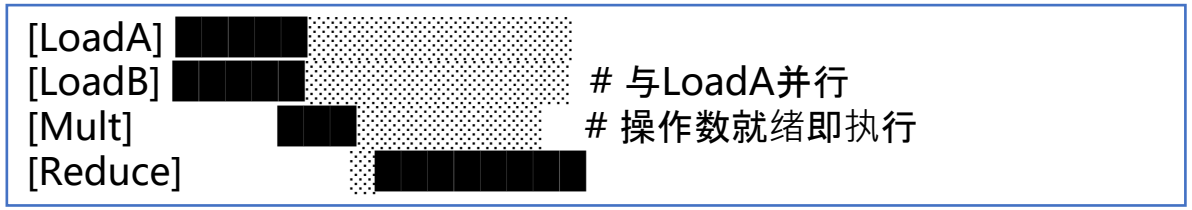
顺序RVV

```
vle32.v v8, (a1)          # 加载A → 阻塞5周期（内存延迟）
vle32.v v16, (a2)         # 加载B → 阻塞5周期（依赖v8完成？）
vfmul.vv v24, v8, v16     # A*B → 依赖v8/v16就绪（阻塞3周期）
vfredsum.vs v0, v24, v0 # 规约累加 → 依赖v24（阻塞7周期）
```



乱序RVV

```
vle32.v v8, (a1)          # 发射加载A（不等待）
vle32.v v16, (a2)         # 立即发射加载B（地址独立，无依赖）
vfmul.vv v24, v8, v16     # 发射乘法（乱序执行引擎自动等待操作数）
vfredsum.vs v0, v24, v0
```



▉ = 执行中，⁙ = 气泡（闲置）
假设访存延迟=5周期，乘法=3周期，规约累计=7周期

典型计算任务性能实验

		cycles		
		顺序RVV	乱序RVV	性能变化
NCNN	retinaface	30128678	26264898	14.71%
	mobilenet2	177047635	155845959	13.60%
	yolo7	222210627	201500295	10.28%
	yolo4	272077999	255856213	6.34%
	mobilenet	229263130	210928672	8.69%
	squeezenet	138588391	122670885	12.98%
	blazeface	24763745	19728965	25.52%
	shufflenet	57232159	49578193	15.44%
opencv_demo	AddingImages	640640295	532370924	20.34%
	median_blur	1348372734	1290392279	4.49%
	mobilenet_ssd	727945574	559124099	30.19%
	yolov5	2124447648	1771466810	19.93%
opencv_imgpro	demosaicingEA/8	153973013	154130746	-0.10%
	cvtColor8u/330	657242821	654715143	0.39%
openblas	amax	1126	1126	0.00%
	daxpy	1749	1686	3.74%
	gemv	166407	166407	0.00%
	copy	71778	58499	22.70%
	nrm2	5946	5946	0.00%

顺序RVV vs 乱序RVV

相同的VLEN，DLEN长度，乱序RVV对比顺序RVV，NCNN，OpenCV，OpenBLAS的典型任务性能提升 6.34%-30.19%

乱序RVV核实例介绍

Dubhe-83 特性：

RV64GCBVH，兼容RVA23;

10级以上流水线，三发射;

Scalar和Vector均为乱序执行;

支持Vector1.0 和 Vector crypto扩展;

VLEN =DLEN=256-bit;

支持Data-type: INT8/16/32/64,FP16/32/64,BF16;

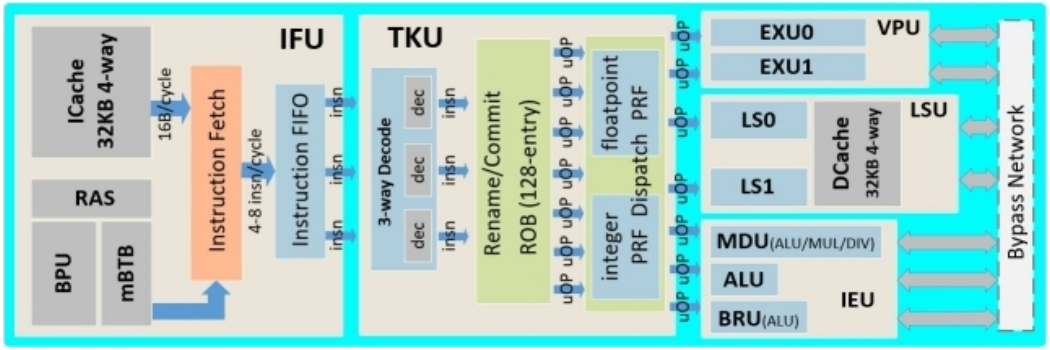
SPECint2006：8.5/GHz

Dubhe-83 设计亮点：

Vector宏指令拆分为DLEN长度的微指令，任何微指令的执行均为乱序;

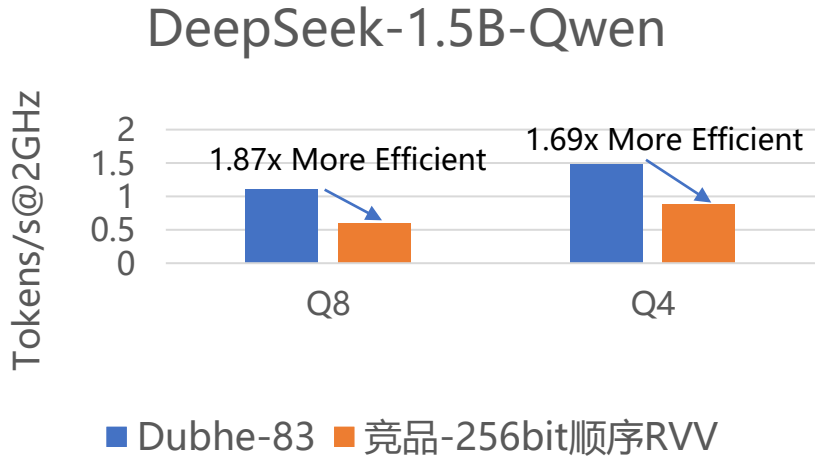
VPU和FPU共享执行单元和物理寄存器，可以达到更好的能效比;

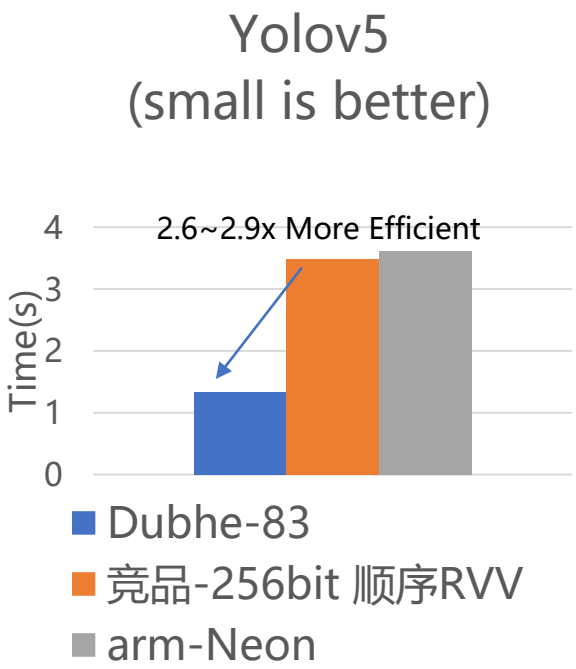
Scalar和Vector共享LSU

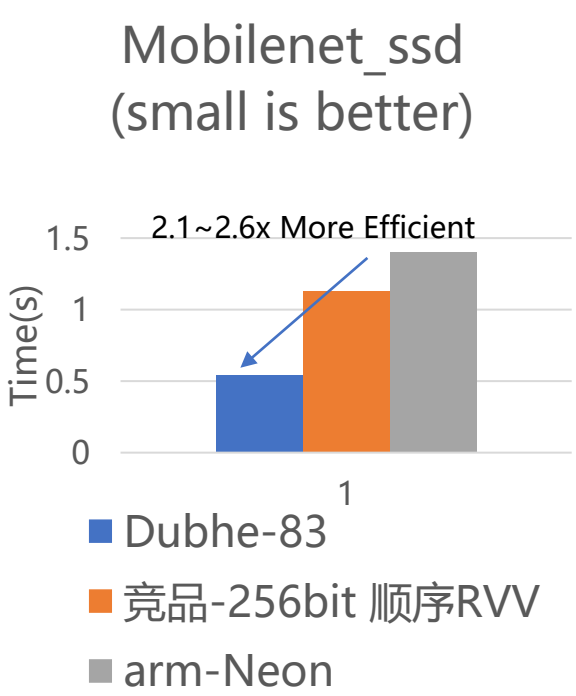


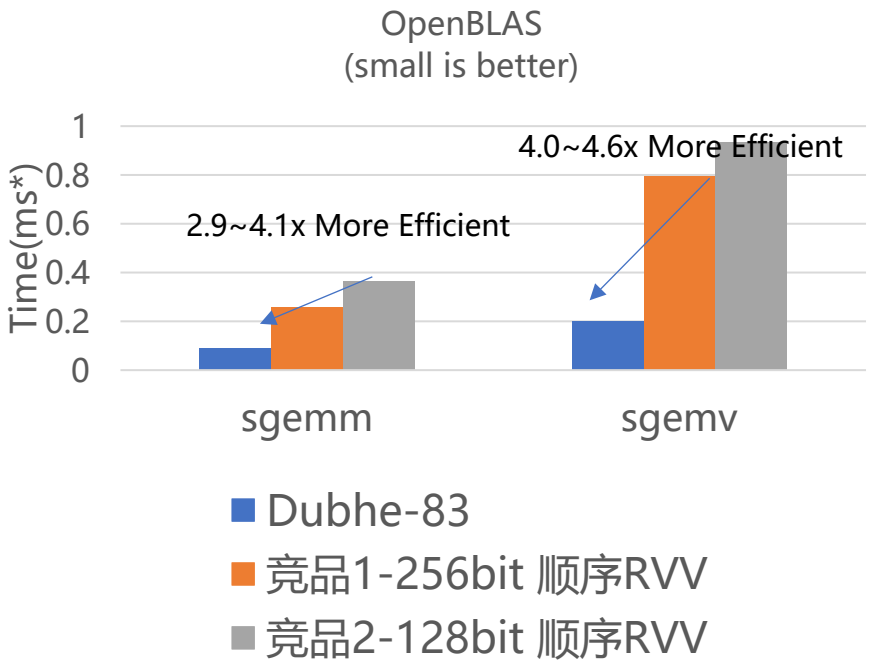
Dubhe-83：Vector Performance

对比DeepSeek,OpenCV和OpenBLAS在不同平台上的运行效率，Dubhe-83 乱序RVV的性能优于其他竞品，充分发挥了乱序执行微架构的优势。









StarFive
赛昉科技

微信公众号

RVspace社区

www.starfivetech.com

sales@starfivetech.com
marketing@starfivetech.com

021-50478300

StarFive
赛昉科技

以RISC-V创新为客户创造价值