燧原科技

Enflame

Hot Chips 33

ZOMI

Ryan Liu / Chuang Feng

August 2021

AI 芯片

Talk Overview

1. AI 计算体系
   - 深度学习计算模式
   - 计算体系与矩阵运算
2. AI 芯片基础
   - 通用处理器 CPU
   - 从数据看 CPU 计算
   - 通用图形处理器 GPU
   - AI专用处理器 NPU/TPU
   - 计算体系架构的黄金10年

1. **华为昇腾 NPU**
   - 达芬奇架构
   - 昇腾AI处理器
2. **谷歌 TPU**
   - TPU 核心脉动阵列
   - TPU 系列架构
3. **特斯拉 DOJO**
   - DOJO 架构
4. **国内外其他AI芯片**
   - AI芯片的思考

Talk Overview

I. 国内其他 AI 芯片

- 壁仞 芯片剖析
- 寒武纪 芯片剖析
- 燧原科技 芯片剖析
- AI 芯片架构的思考

目录 Context

I. 国内其他 AI 芯片

- 壁仞 芯片剖析
- 寒武纪 芯片剖析
- 燧原科技 芯片剖析
- AI 芯片架构的思考

- 什么是燧原
- 燧原产品形态
- 燧原 DTU1.0 芯片架构
- 对燧原思考

1. 什么是燧原

燧原科技：2021年初完成18亿融资，累积融资超30亿

中国最大AI芯片问世

格芯 12nm FinFET 工艺打造

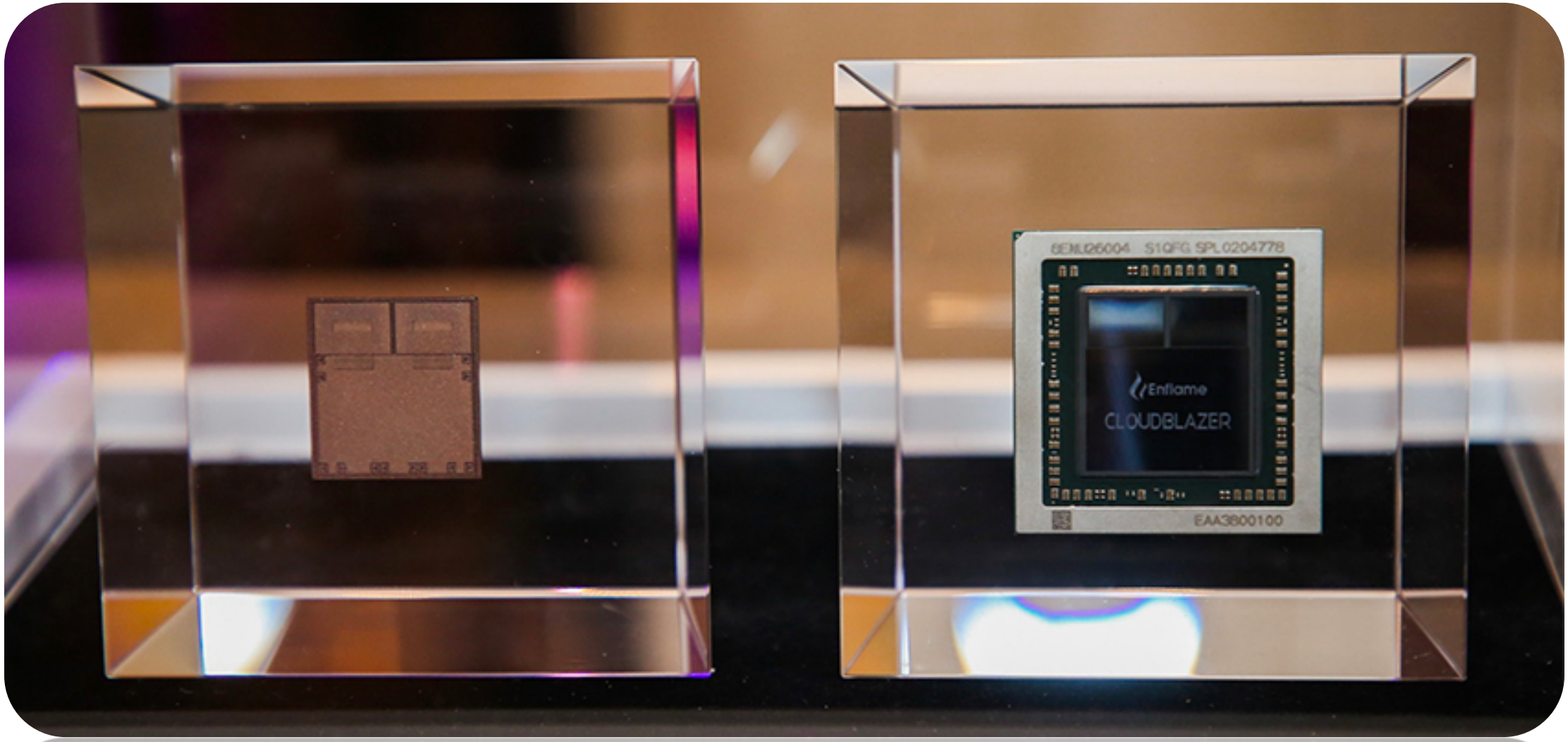
尺寸方面，为57.5毫米×57.5毫米

达到了日月光企业的2.5D封装极限

燧原科技：2021年初完成18亿融资，累积融资超30亿

芯片流片	产品发布	T10/T11 1X Pref/W	T20/T21 4X Perf/W	T30/T31 14X Perf/W
2019.5.30	2019.12.11	2020	2021	2023

2代芯片



2. 燧原产品形态

云端AI产品路线图

云端训练

板卡：云燧T10/T11

板卡：云燧T20/T21

2019.09/2019.12

2020.12

2021.07

2021.12

板卡：云燧i10

板卡：云燧i20

云端推理

产品矩阵

1. **封装**：中国最大的计算芯片
2. **计算**：TF32精度峰值算力160TFLOPS
3. **数据**：植入完全可编程的数据流
4. **存储**：率先支持HBM2E先进存储
5. **互联**：高速互联支撑算力扩展

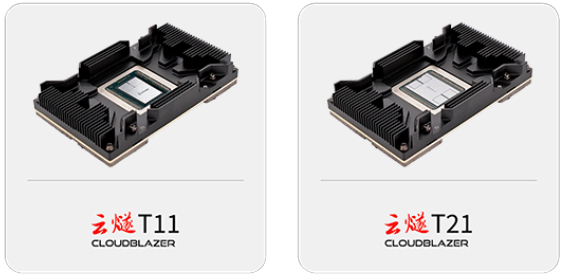
i20 云端推理产品

	云燧i20	云燧i10	NVIDIA T4	NVIDIA L4
制造工艺	GloFo N12	GloFo N12	TSMC 12nm	TSMC 5nm
单精度性能(FP32)	32 TFLOPS	17.6 TFLOPS	8.1 TFLOPS	30.3 TFLOPS
单精度性能(TF32)	128 TFLOPS	/	/	120 TFLOPS
半精度性能(FP16/BF16)	128 TFLOPS	70.4 TFLOPS	65 FP16 TFLOPS	242 TFLOPS
整型性能(INT8)	256 TOPS	470.4 TFLOPS	130 INT8 TOPS	485 TOPS
内存	16GB HBM2E	16GB HBM2	16 GB GDDR6	24GB
内存带宽	819 GB/s	512 GB/s	320 GB/s	300GB/s
IO 接口	150W	150W	70W	72W
外形规格	FH 3/4L 单槽位 PCIe卡	FHFL 单槽位 PCIe卡	Gen3 x16 PCIe	Gen4 x16 PCIe
发布(量产)	2021	2020	2018	2023

T20 云端训练产品

	云燧T21	云燧T20	NVIDIA A100
制造工艺	GloFo 12nm	GloFo 12nm	TSMC 7nm
单精度性能(FP32)	32 TFLOPS	32 TFLOPS	156 TFLOPS
单精度性能(TF32)	128 TFLOPS	128 TFLOPS	156 TFLOPS
半精度性能(FP16/BF16)	128 TFLOPS	128 TFLOPS	312 TFLOPS
整型性能(INT8)	256 TOPS	256 TOPS	624 TOPS
内存	32GB HBM2E	32GB HBM2E	80GB HBM
内存带宽	1.6TB/s	1.6TB/s	1.55 TB/s
IO 接口	400W	300W	400W
外形规格	OAM模块	FHFL 双槽位 PCIe卡	SXM5
发布(量产)	2021	2020	2020

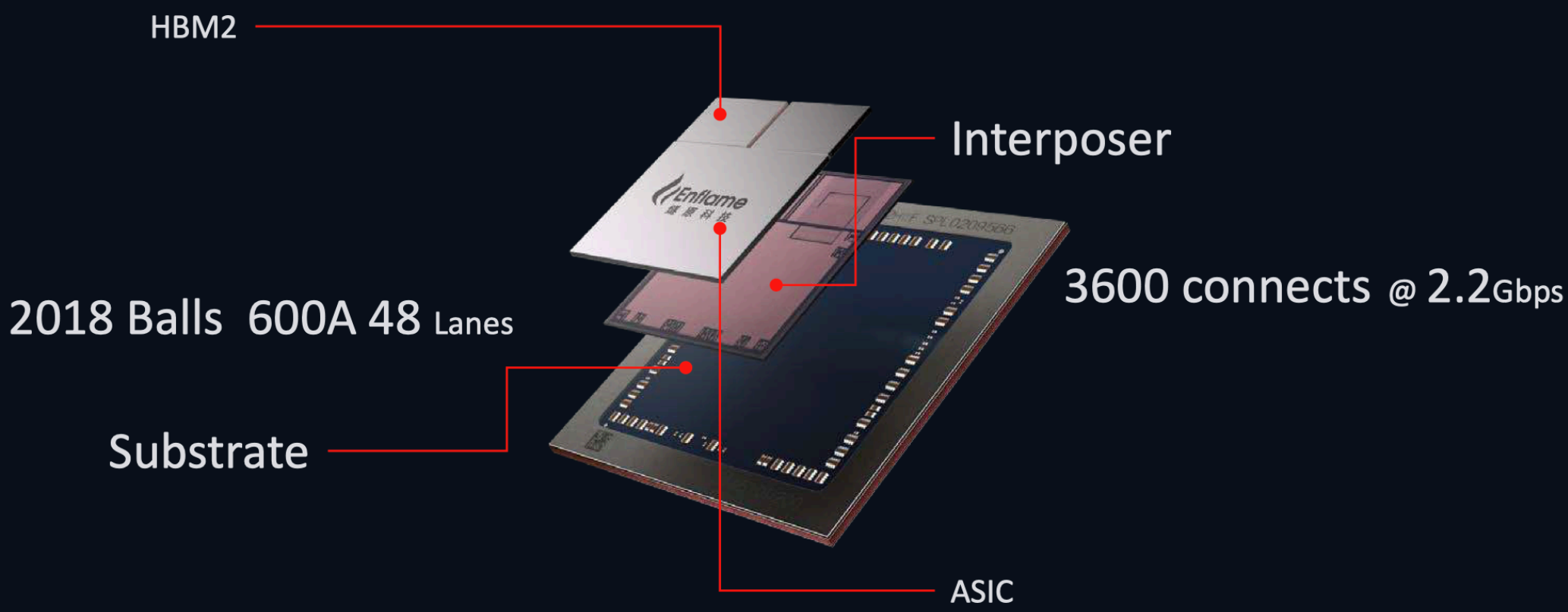
训练POD基于“云燧”训练卡产品系列



推理POD基于“云燧”推理卡产品系列

3. 燧原芯片架构细节

DTU 1.0 封装


HBM2
Interposer
2018 Balls 600A 48 Lanes
3600 connects @ 2.2Gbps
Substrate
ASIC

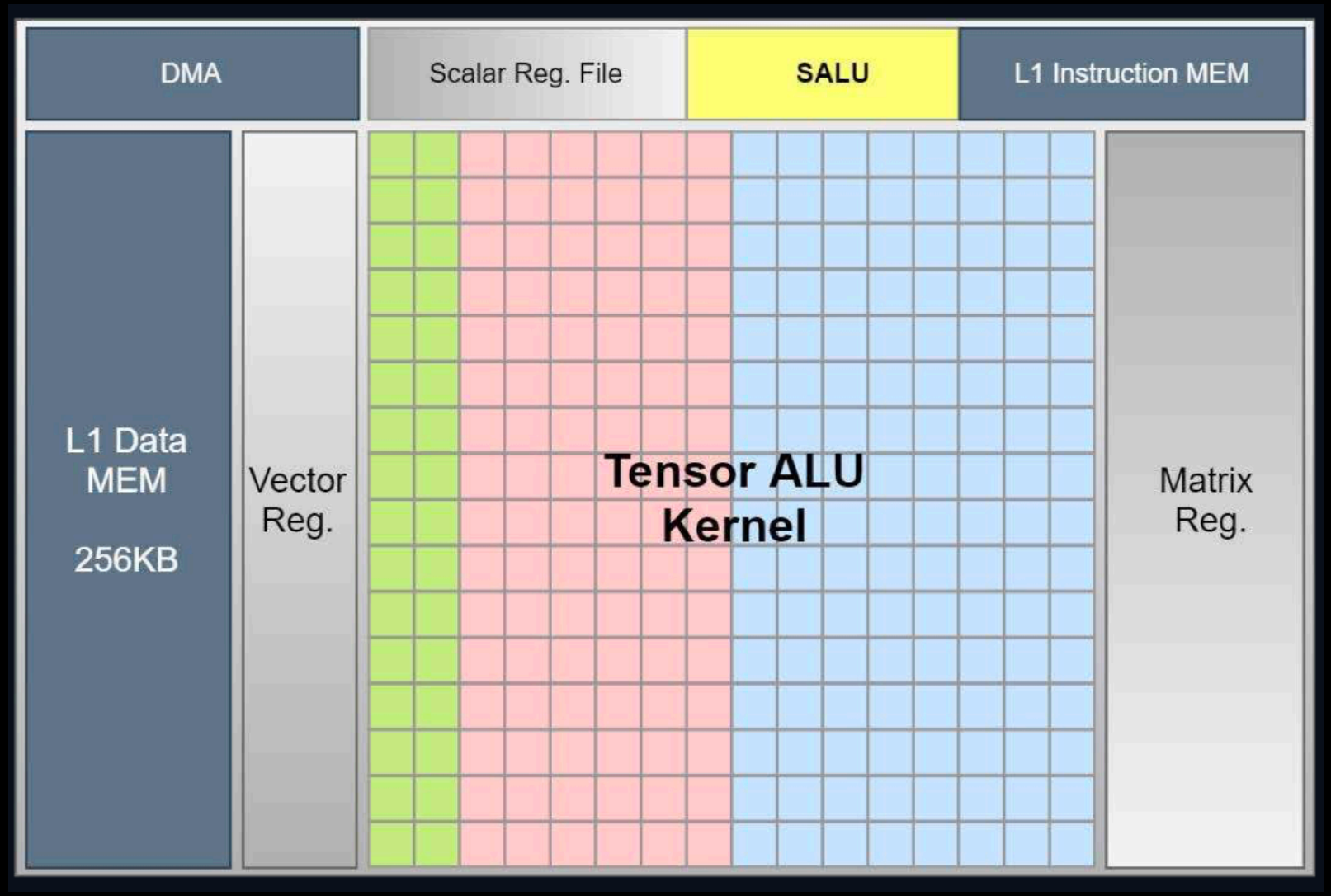
DTU 1.0 SOC

- 32X AI AI核
- 4X 计算簇
- 40个数据传输引擎

- 4 路高速互连
- 2X HBM2 32G
- 512GB/s bandwidth

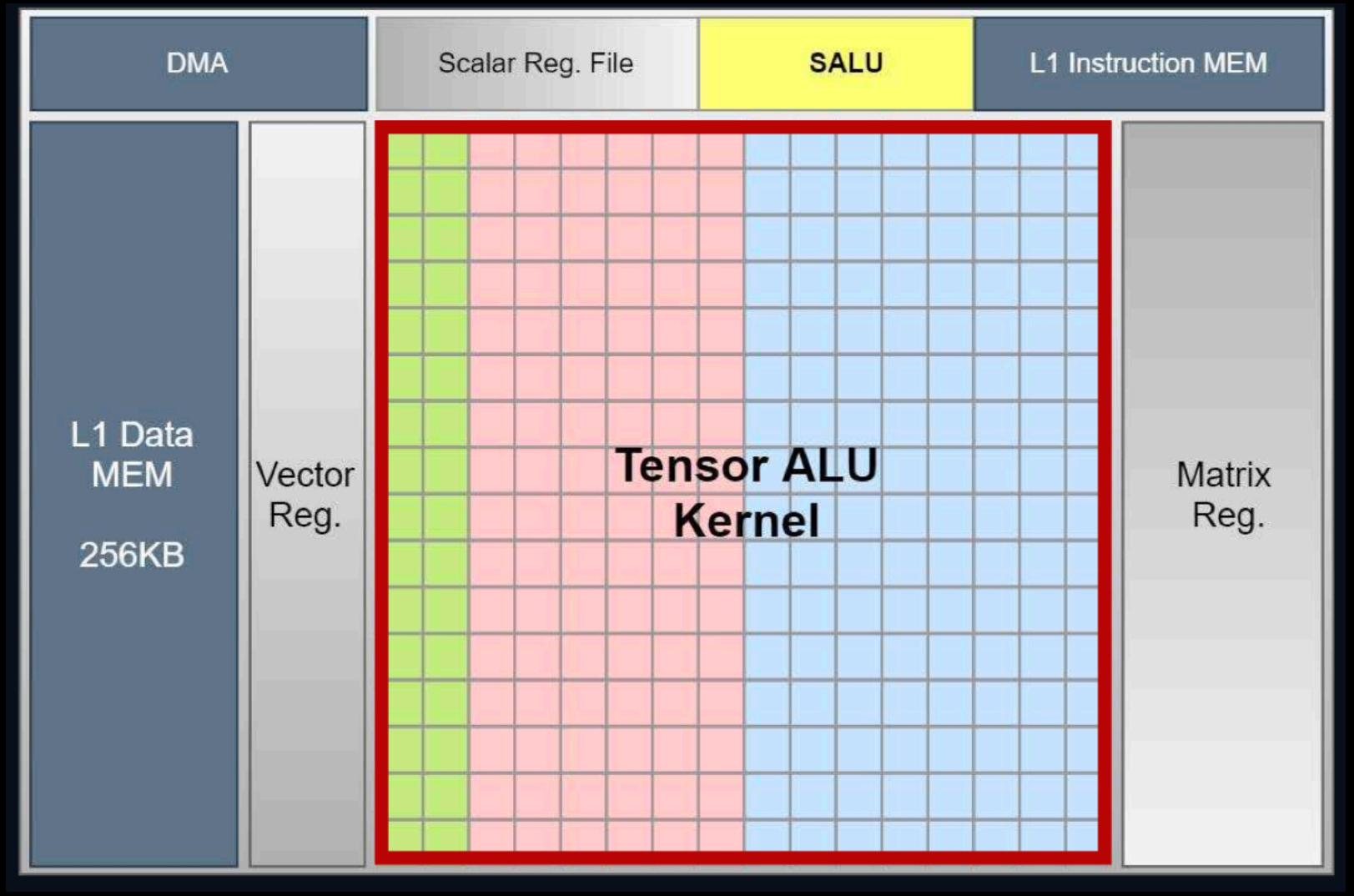
GCU-CARE 1.0

- 20 TFLOPS@FP32
- Bus width 1024-bit
- Full precision support
- Fully programmable VLIW



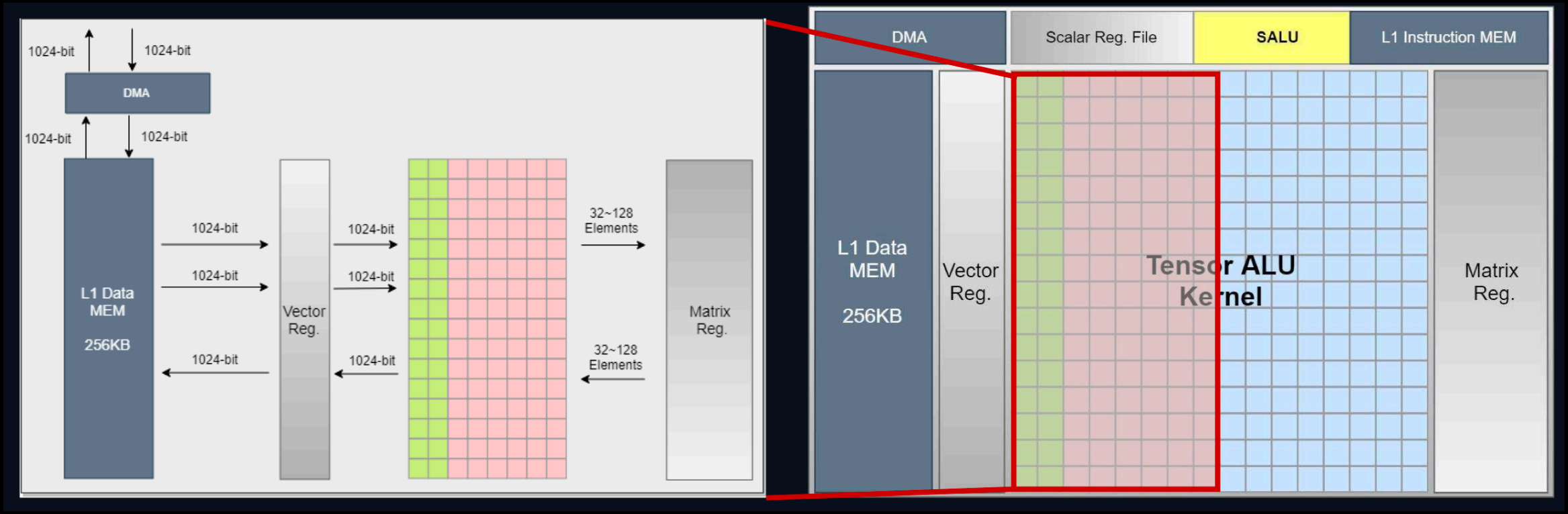
GCU-CARE 1.0

- 256 个 Tensor kernel
- Each kernel 核能力
  - Supports 1x 32-bit MAC
  - Supports 4x 16-bit/8-bit MAC
  - Supports full precision
  - Supports mixed precision
- Reuse data and bandwidth between kernels

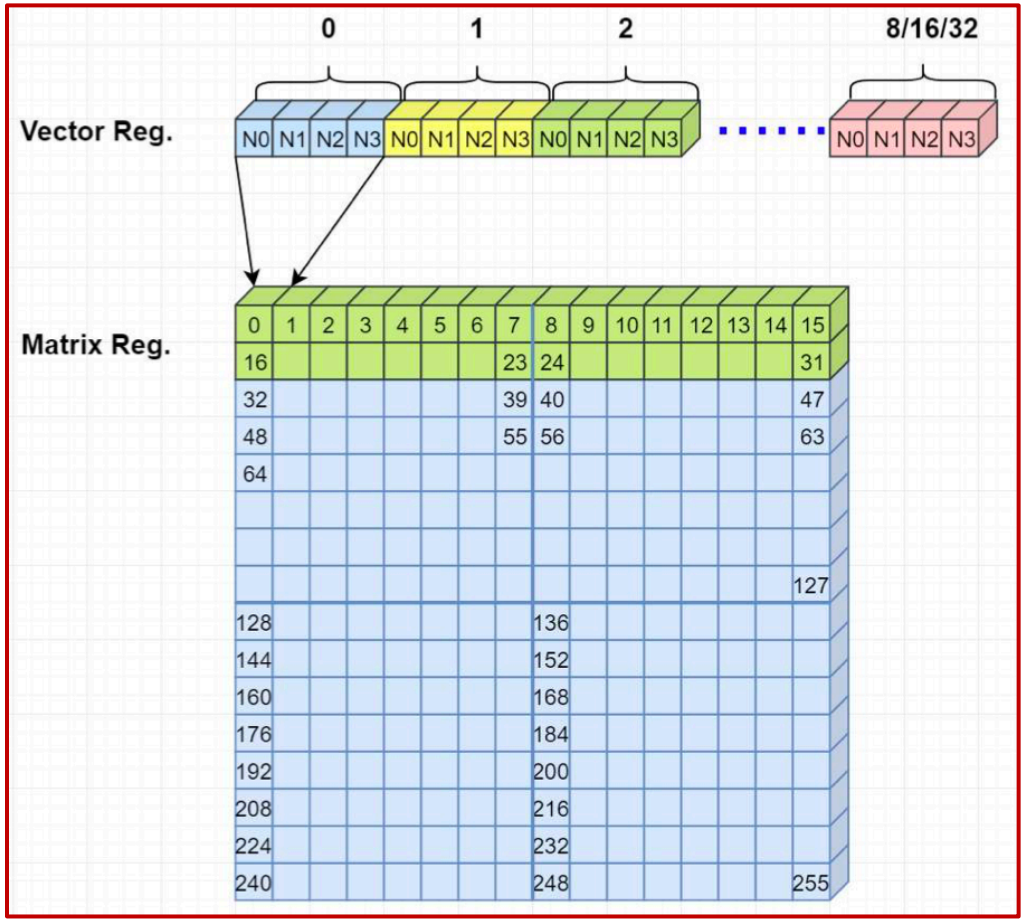


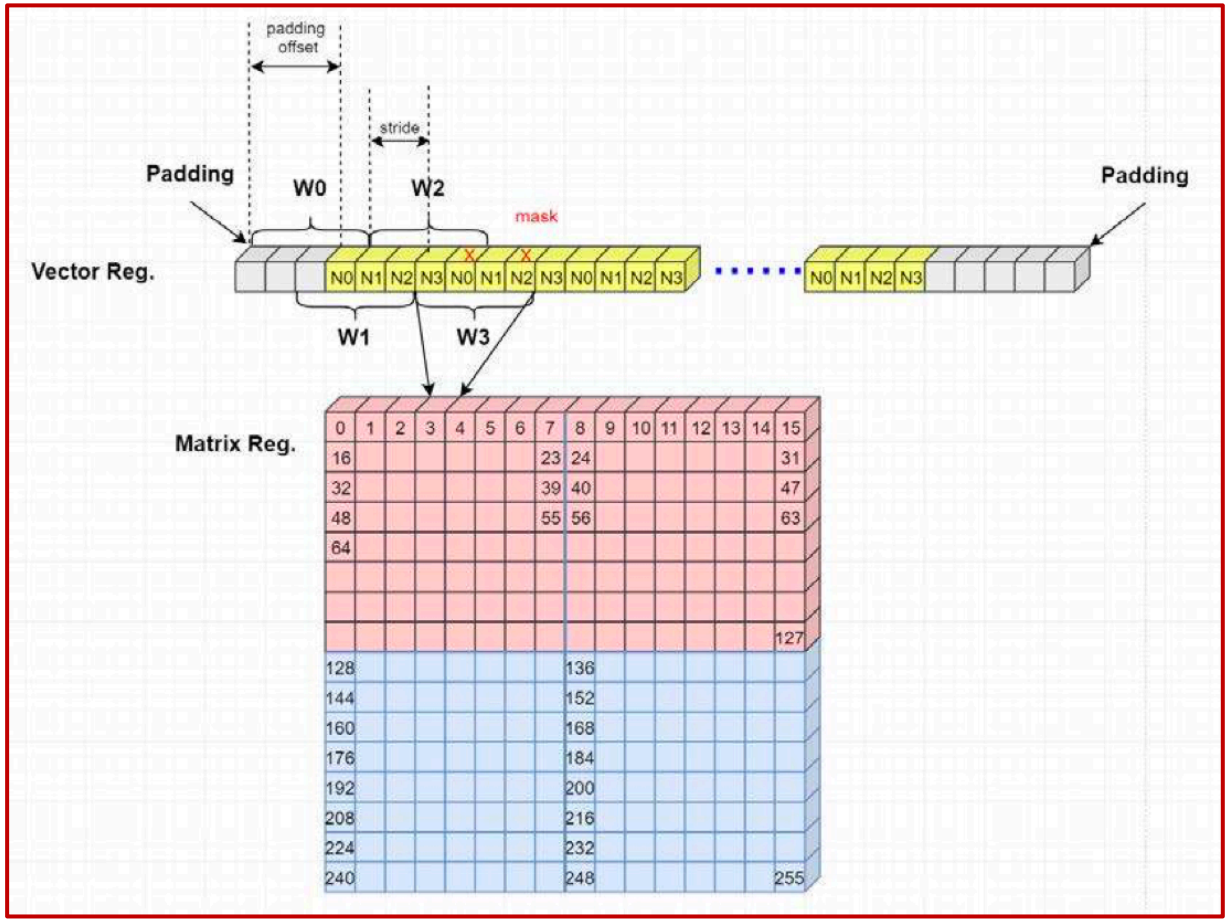
32 Kernels for 1024-bit General Vector

32/64/128 Kernels for Vector MAC

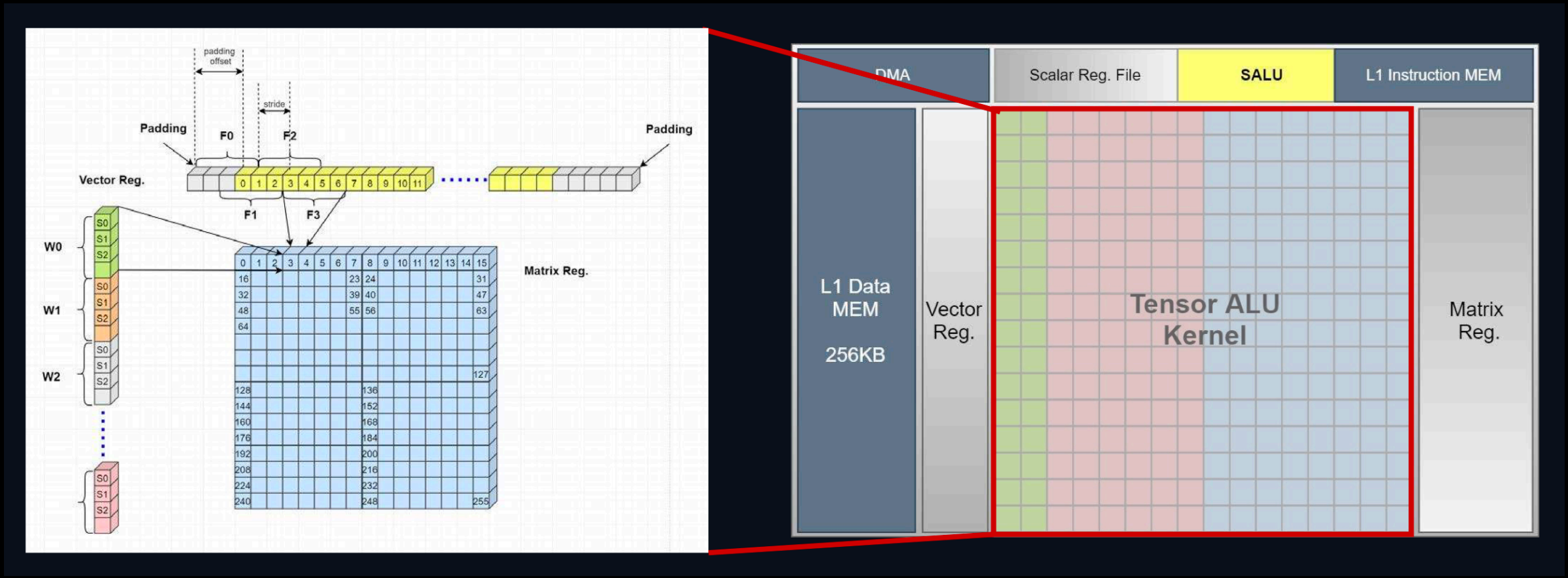


Vector Sum and Vector Pooling

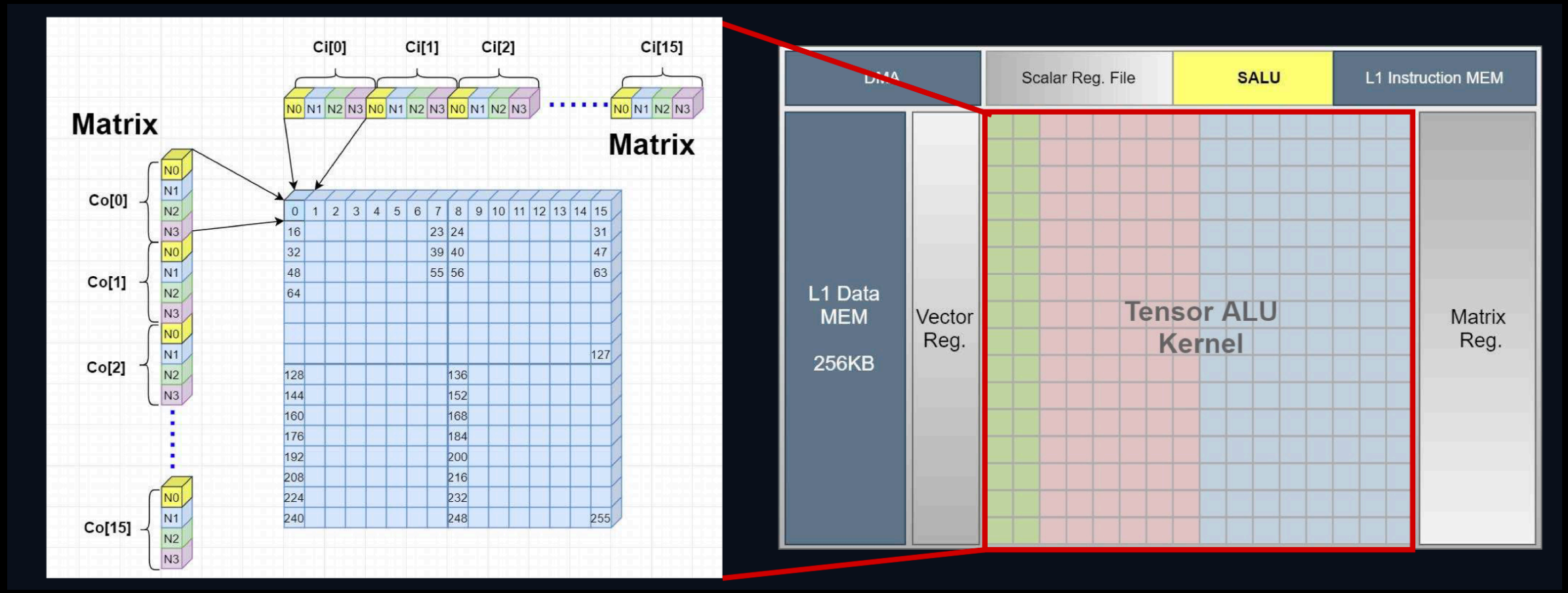




256 Kernels for Conv Operations



256 Kernels for GEMM Operations



邃思2.0芯片: 基于12nm的全新GCU架构

计算

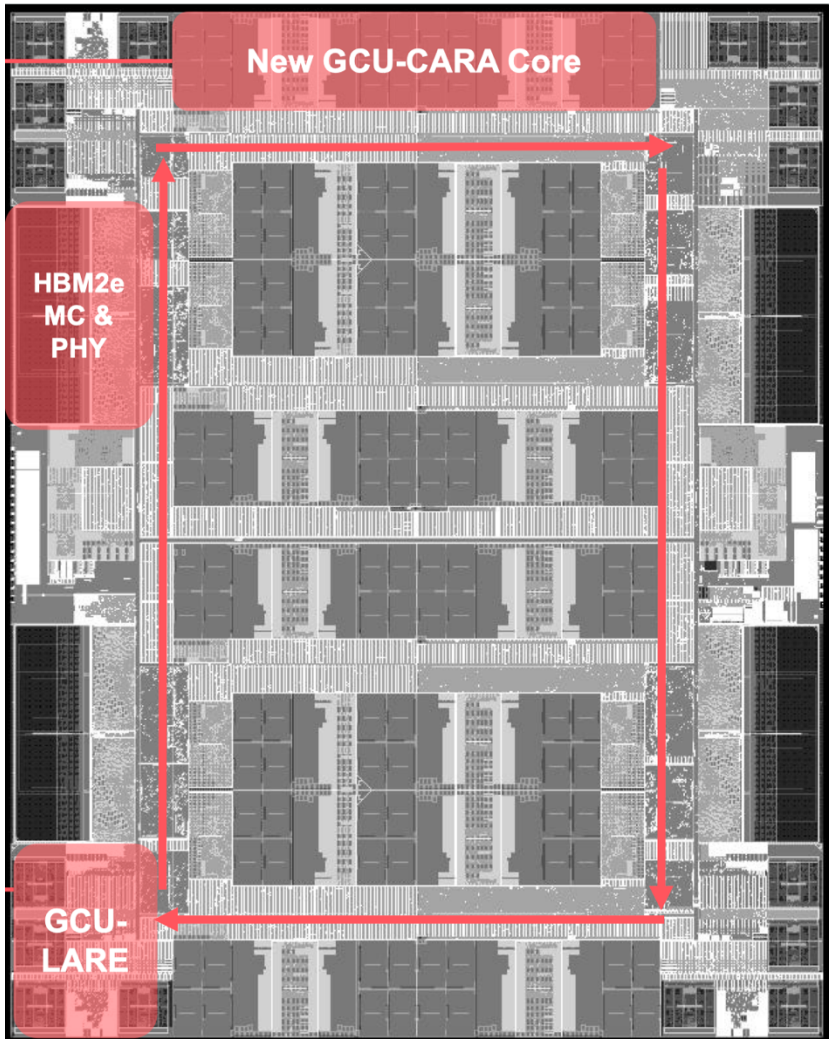
- 新算力架构，支持TF32数据格式
- 面积增加7%, 性能增强60%
- 增强多线程张量计算和超越函数，提高数据计算并行度

存储

- 支持HBM2E, 存储容量最高64GB
- 增大片上L2容量，增强多级存储间的访问类型

带宽和互联

- 片上带宽提升 1.6x~3.2x
- 互联带宽提升 50%



4. 思考

思考

1. **竞争力**：产品形态主要集中在云端推理，软件栈没有提（快速构建过程中），性能对标NV上一代T4；云端训练基本上没有优势，作为 DSA 架构对标 NV GPGPU 架构 7 年前 P100 仍有差距。竞争力在哪？
2. **先进性**：hotchip33 会议公开了 DTU 1.0 的大致架构，都2023年了，看不到新的内容。更多把2年上一代将要退市产品的架构拿出来show，诚意不够，先进性落后。如何追赶成为业界佼佼者？还是甘于做腾讯云的NV替代方？

Reference 引用&参考

1. https://zhuanlan.zhihu.com/p/551888300 陈巍谈芯：最新发布的壁仞GPU BR100参数深度对比和优势分析
2. https://www.eet-china.com/news/202208100913.html 详解壁仞刚刚发布的GPU
3. https://zhidx.com/p/341643.html 国产最强通用GPU来了
4. https://www.geekpark.net/news/306540 详解壁仞刚刚发布的 GPU
5. https://www.zhihu.com/question/547728200 如何评价壁仞科技发布的最大算力GPGPU BR100

BUILDING A BETTER CONNECTED WORLD

THANK YOU