

# 壁仞BR100架构



AI 芯片



ZOMI

壁仞科技 智绘全球

RENDER THE WORLD WITH INTELLIGENCE

# Talk Overview

## I. 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

## I. 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
- 麟原科技 芯片剖析
- AI 芯片架构的思考

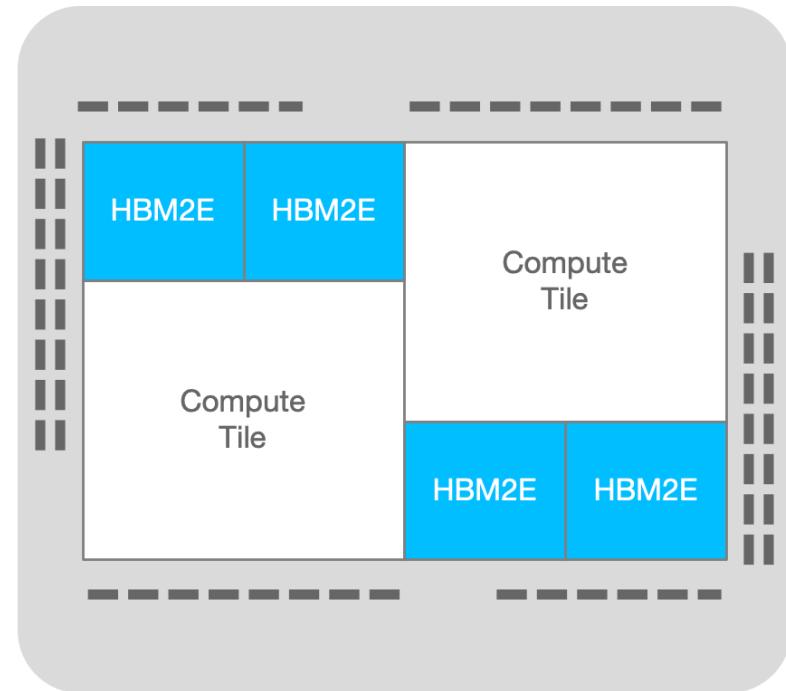
# 目录 Context

## I. 国内其他 AI 芯片

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

- 什么是壁仞
- 壁仞产品形态
- 壁仞软件平台
- RBI00 芯片架构细节
- 对壁仞思考

# 产品形态



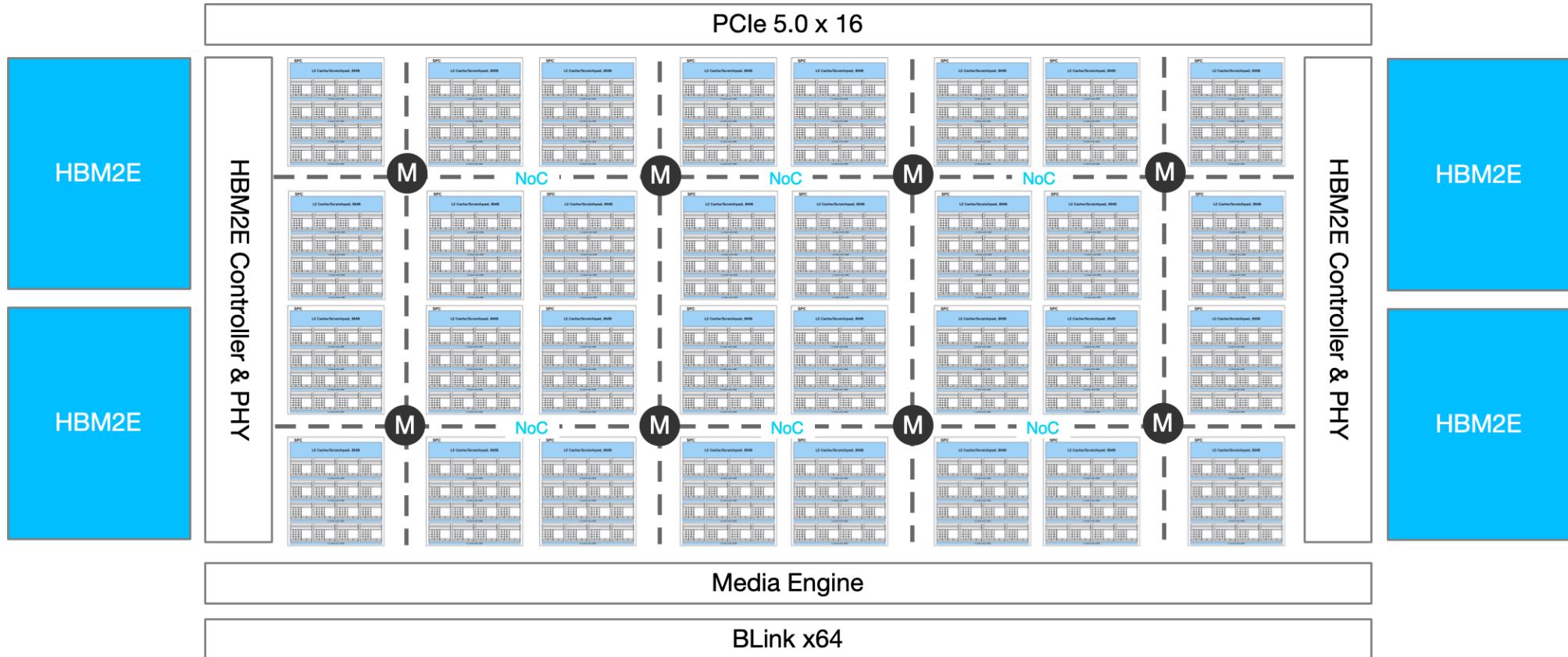
# 壁砾™ 100 vs NV A100/H100

	BIREN BR100	NV A100	NV H100
制造工艺	TSMC N7 @ 77B 1074mm <sup>2</sup>	TSMC N7 @ 54.2B 828mm <sup>2</sup>	TSMC N4@80B 814mm <sup>2</sup>
算力	2048 TOPS @ INT8 <b>1024 TFLOPS @ BF16</b> 512 TFLOPS @ TF32+ 256 TFLOPS @ FP32 支持FP16、INT32、INT16等类型	624 TOPS @ INT8 <b>312 TFLOPS @ BF16</b> <b>312 TFLOPS @ FP16</b> 156 TFLOPS @ TF32 支持 FP32、FP64 等类型	4000 TOPS @ INT8 <b>2000 TFLOPS @ BF16</b> <b>2000 TFLOPS @ FP16</b> <b>1000 TFLOPS @ TF32</b> 支持 FP32、FP64 等类型
多实例GPU	SVI	MIG(7个， 每个10G)	MIG(7个， 每个10G)
架构	壁立	Ampere	Hopper
通用计算核心数	8192 Steam Processing	6912 CUDA Core	15872 CUDA Core
AI计算核心数	512 T-Core	432 Tensor Core	528 Tensor Core
缓存	300MB	40MB L2	50MB L2
内存	64GB HBM	80GB HBM	80GB HBM
互联	Blink 512GB/s	NVLink 600GB/s	NVLink 900GB/s
功耗	550W	400W	700W
接口	PCIe Gen5	SXM5	SXM5
发布(量产)	2022(2025?)	2020(2021)	2022(2023)

# 4. RBI00 芯片

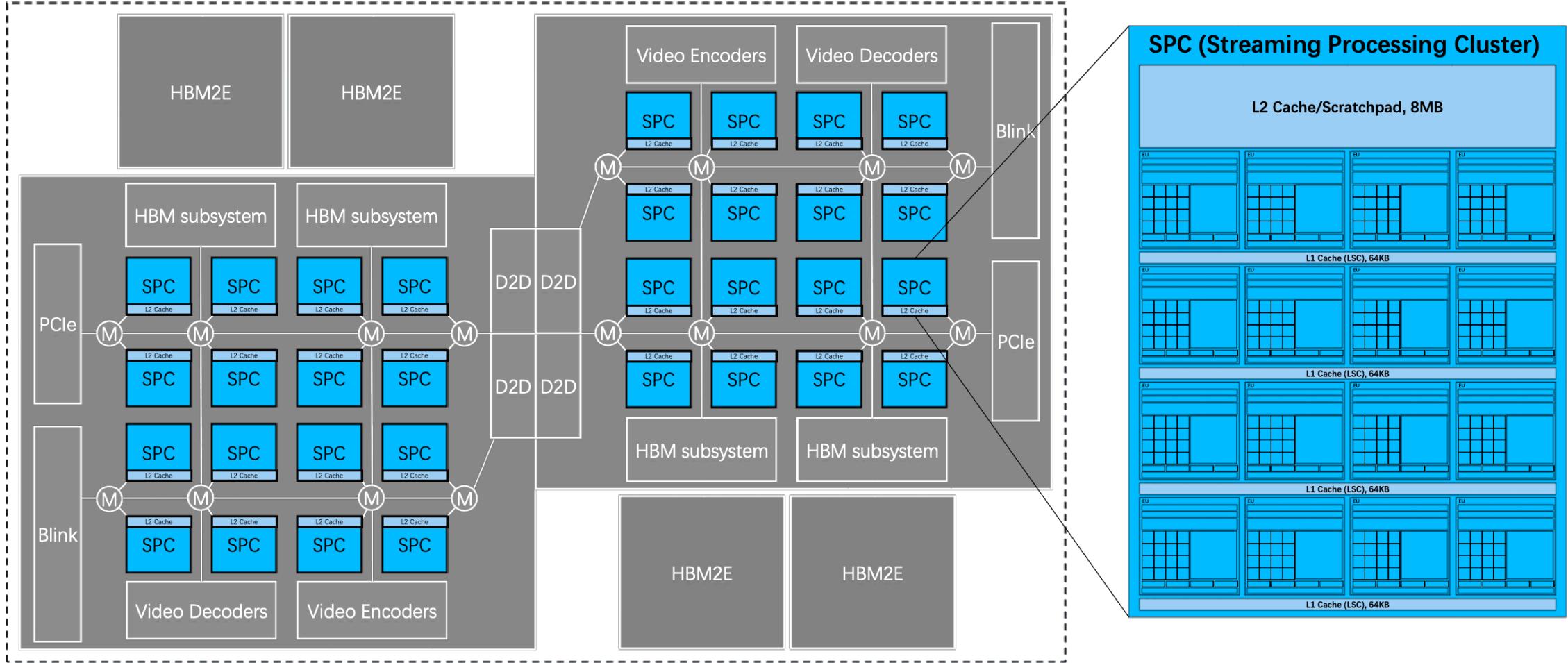
## 架构细节

# BR100系列 芯片架构

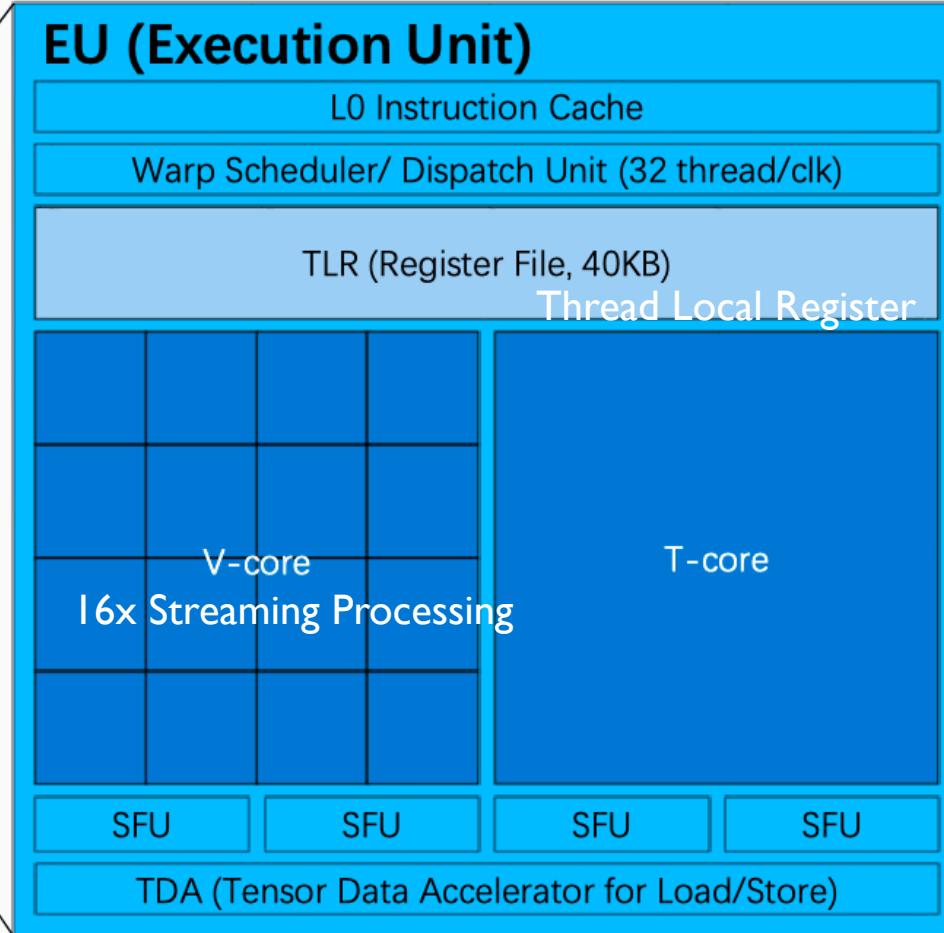
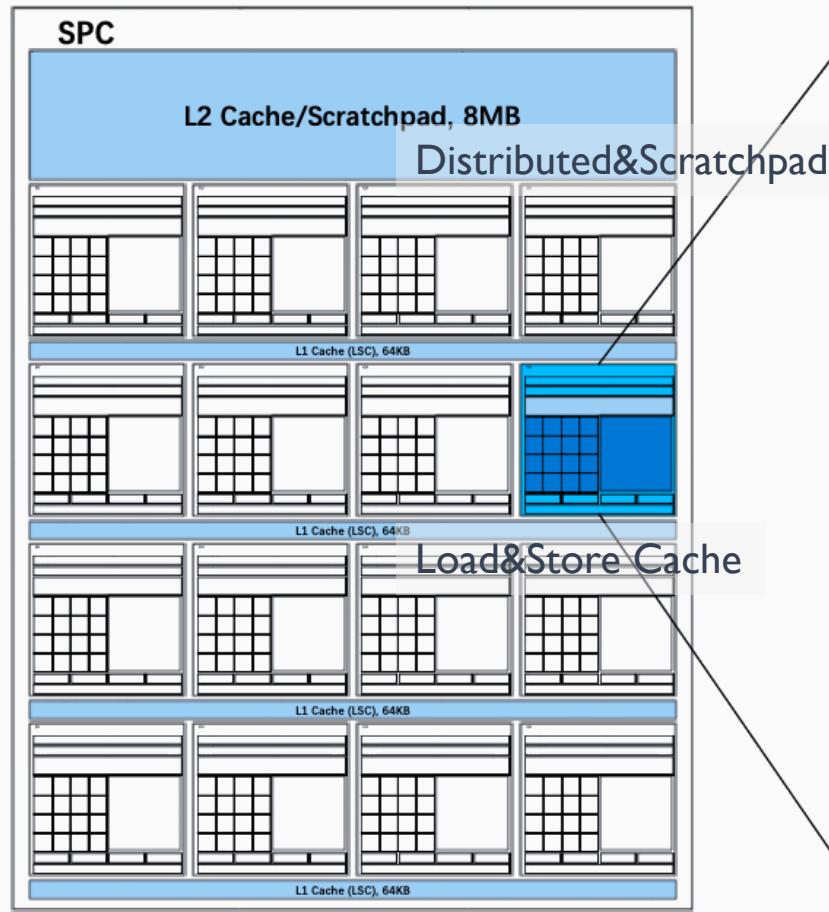


- PCIe5.0 新一代主机接口
- BLink™ 点对点全互连技术
- NoC 网格式多播片上互连
- SPC 流式处理器簇
- L2 Cache 分布式共享缓存
- HBM2E 内存系统

# BR100系列 芯片架构



# BR100系列 SPC架构



- 4x64KB L1 Cache
- 8MB L2 Cache
- 16xEU

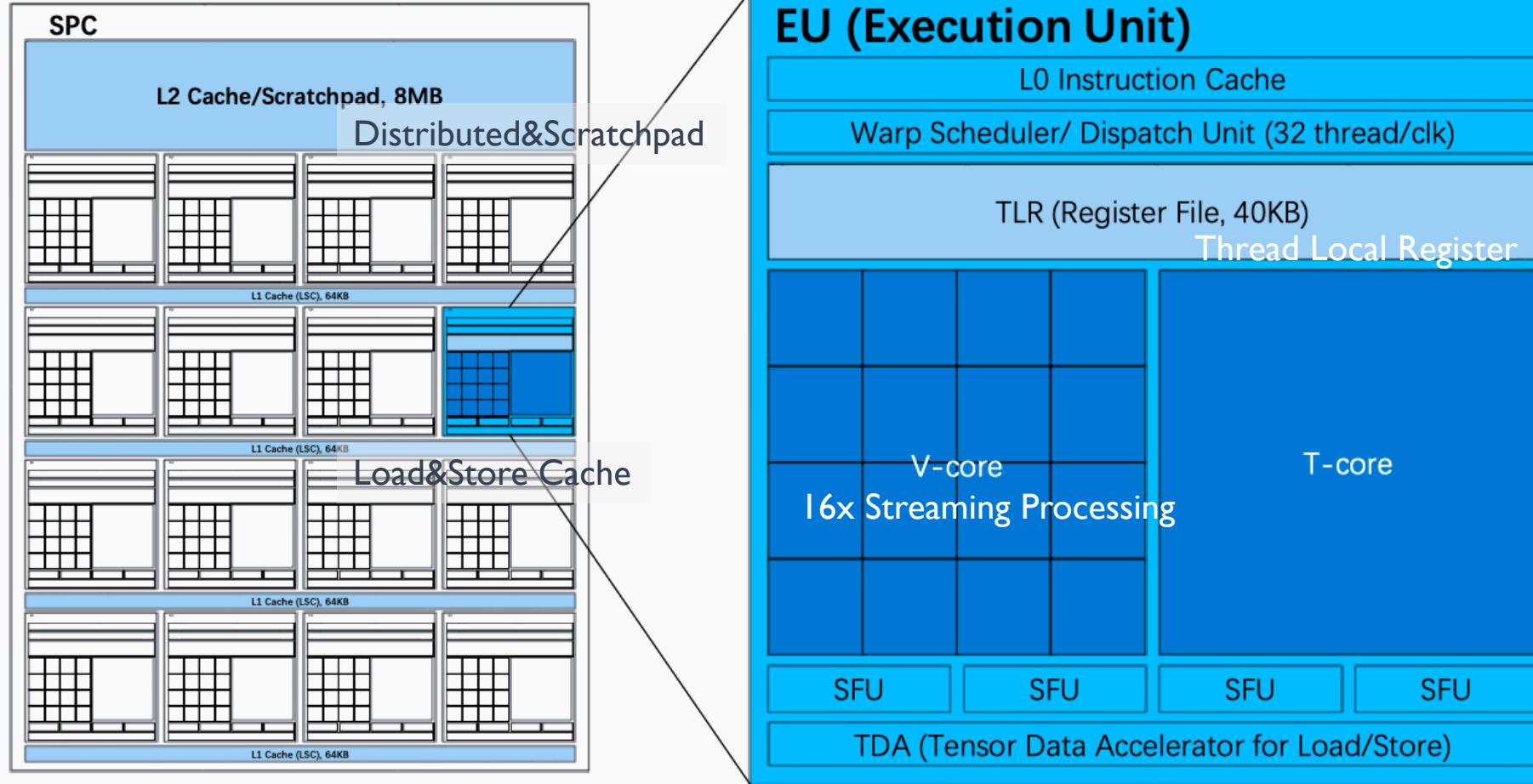
# NVidia A100 架构



- NV GPU L2 cache —般在芯片中间，或者芯片边上，在 Memory Controller 旁边。

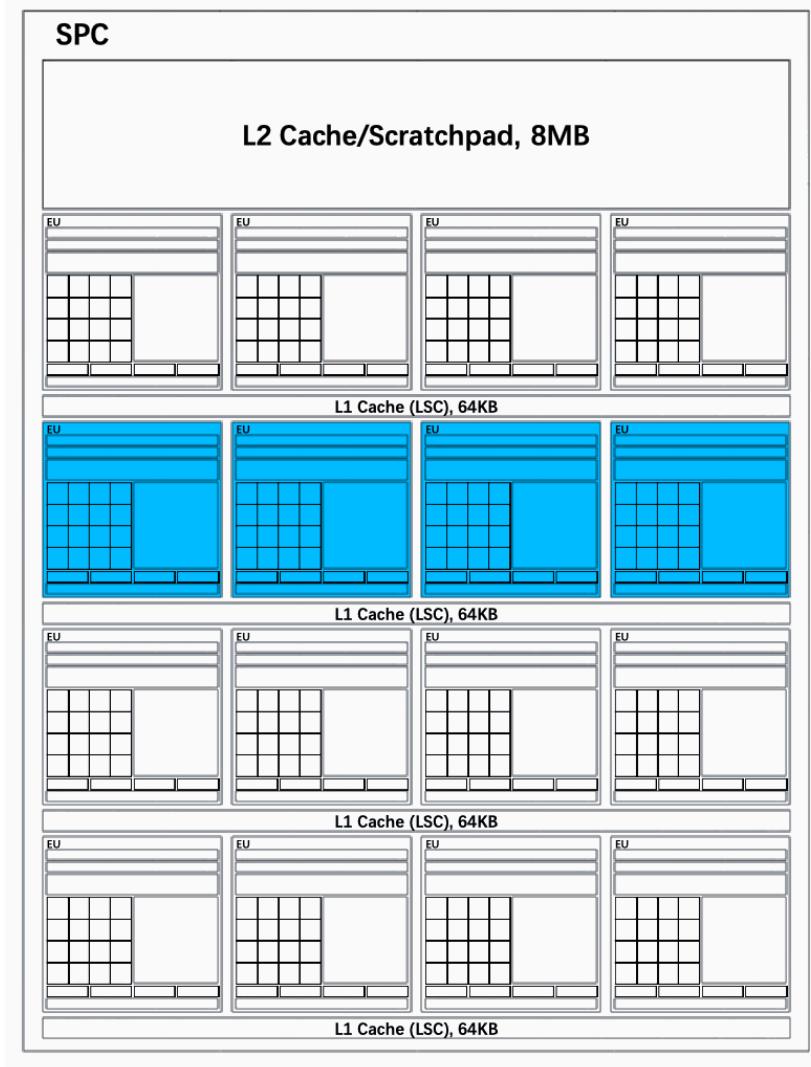
Figure 6. GA100 Full GPU with 128 SMs (A100 Tensor Core GPU has 108 SMs)

# BR100系列 SPC架构

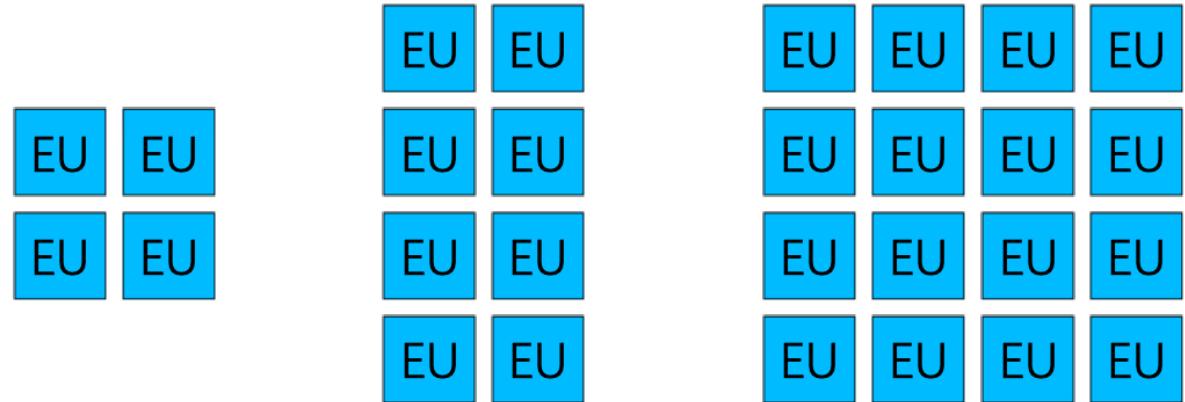


- BR100设计为分布式缓存，在计算核附近，能够共享 cache，并通过片上 NoC 网络将其连在一起。
- **好处：**让数据和计算单元很近，与此同时又在芯片层面实现数据共享（近存计算）；

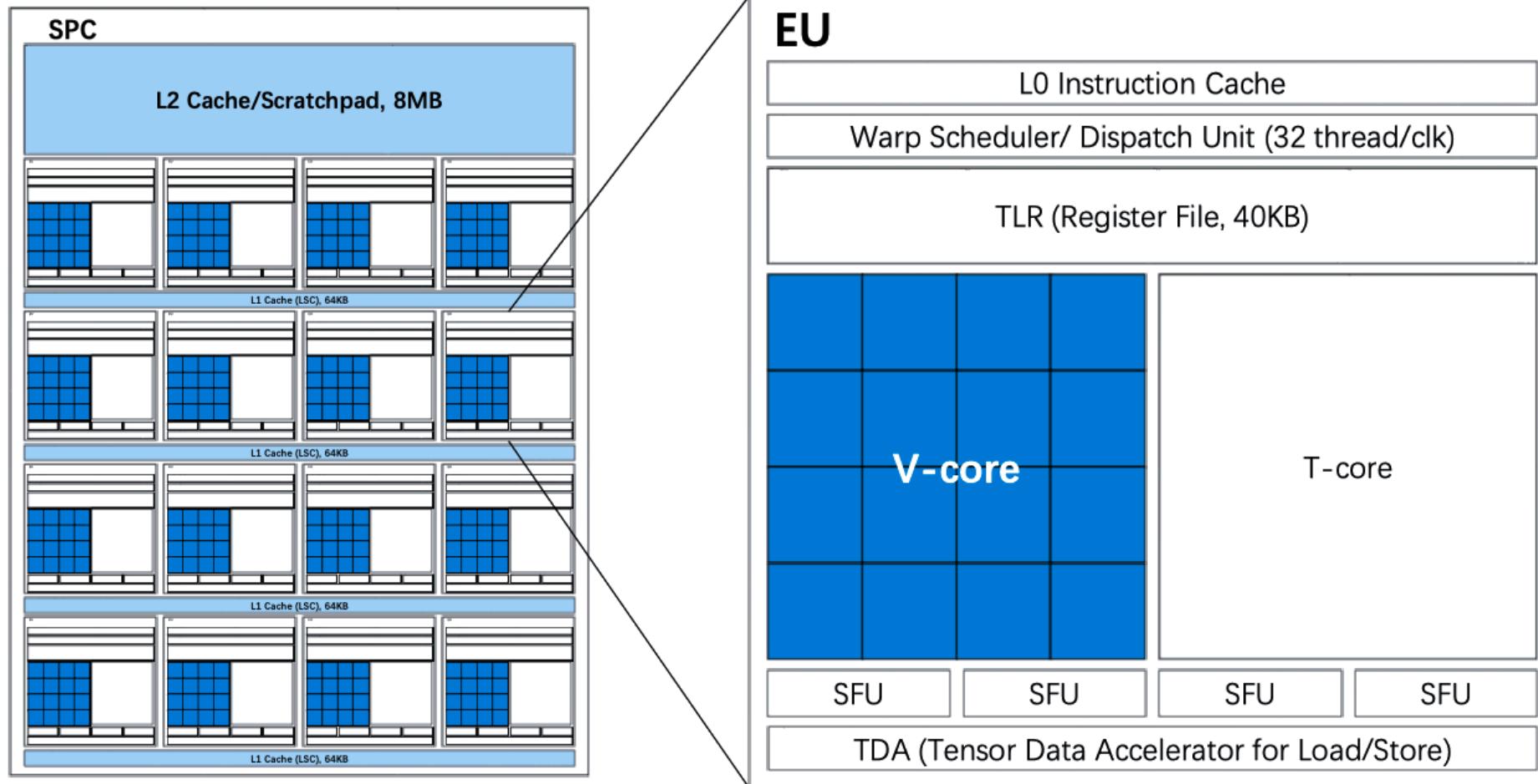
# BR100系列 CU 结构



- 多个 EU 组成一个 CU ( 计算单元 )
- 每个 CU 可包含 4/8/16 EU ( 执行单元 )
- CU 内线程组可以进行同步

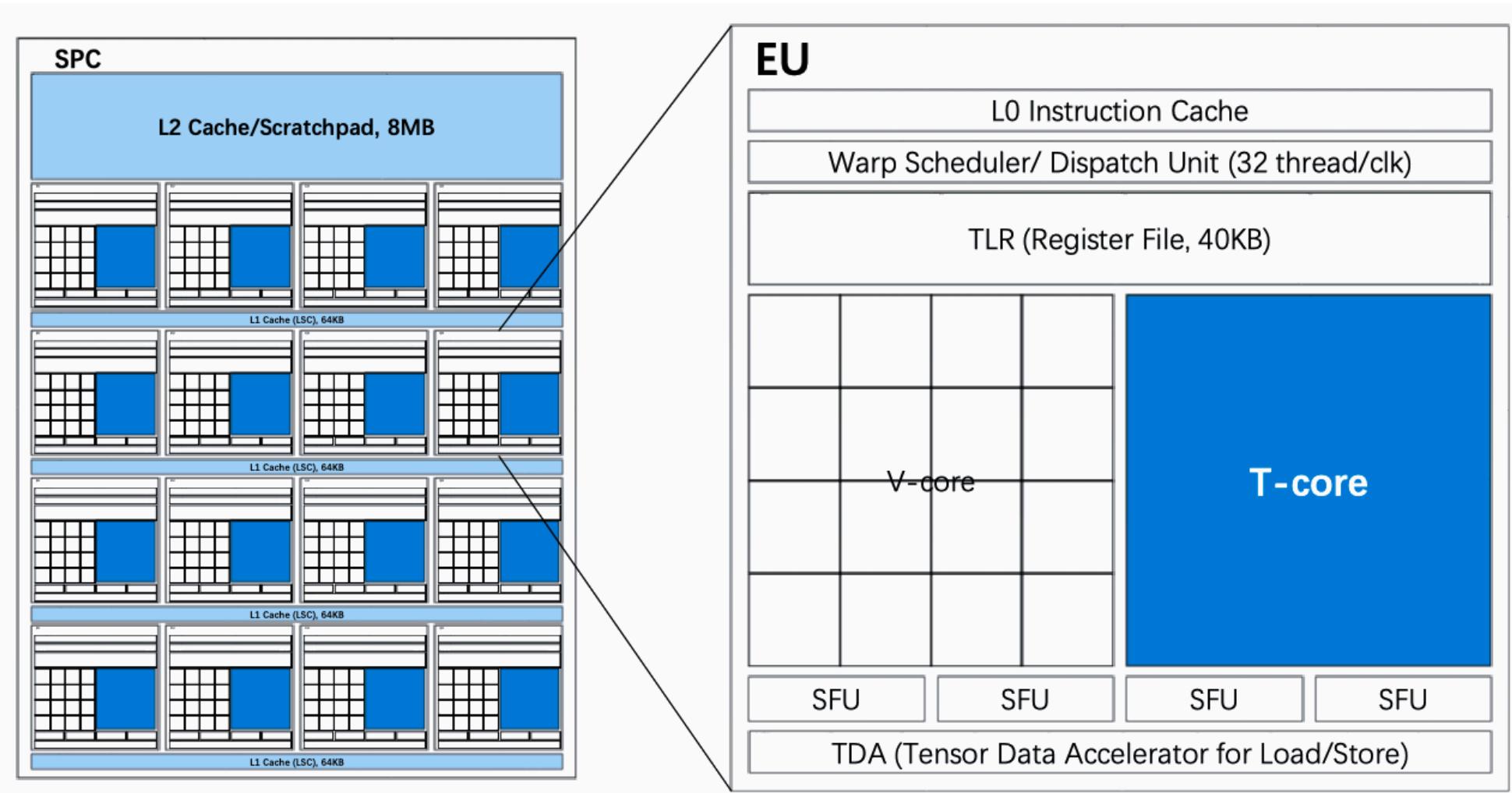


# V-Core(Vector Core) , 通用SIMT计算单元



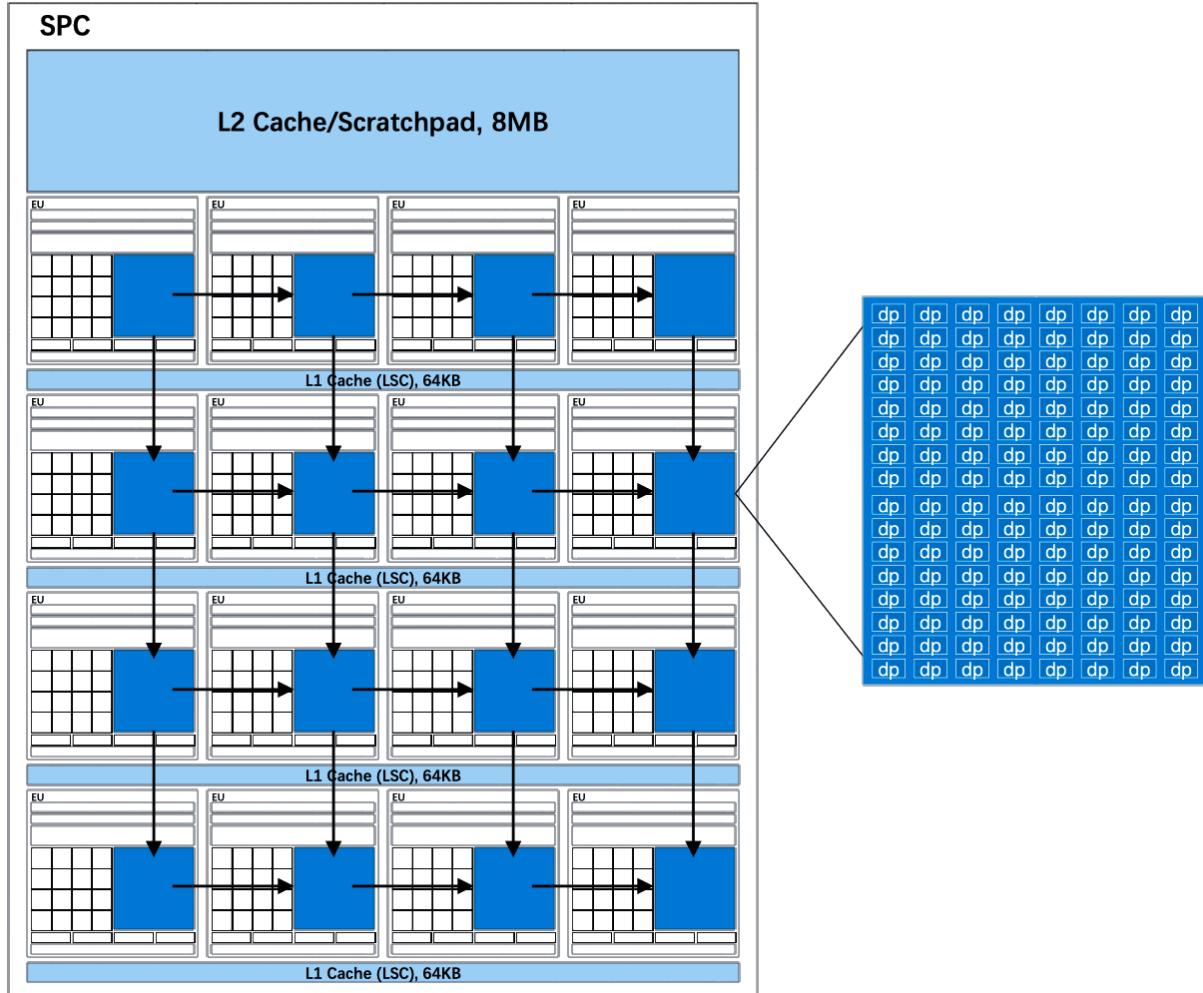
- Batch Norm
- ReLU
- Softmax
- ...
- 128K threads run on 32 SPCs
- Cooperative Wa rps

# T-Core(Tensor Core) , 专用 AI 计算单元



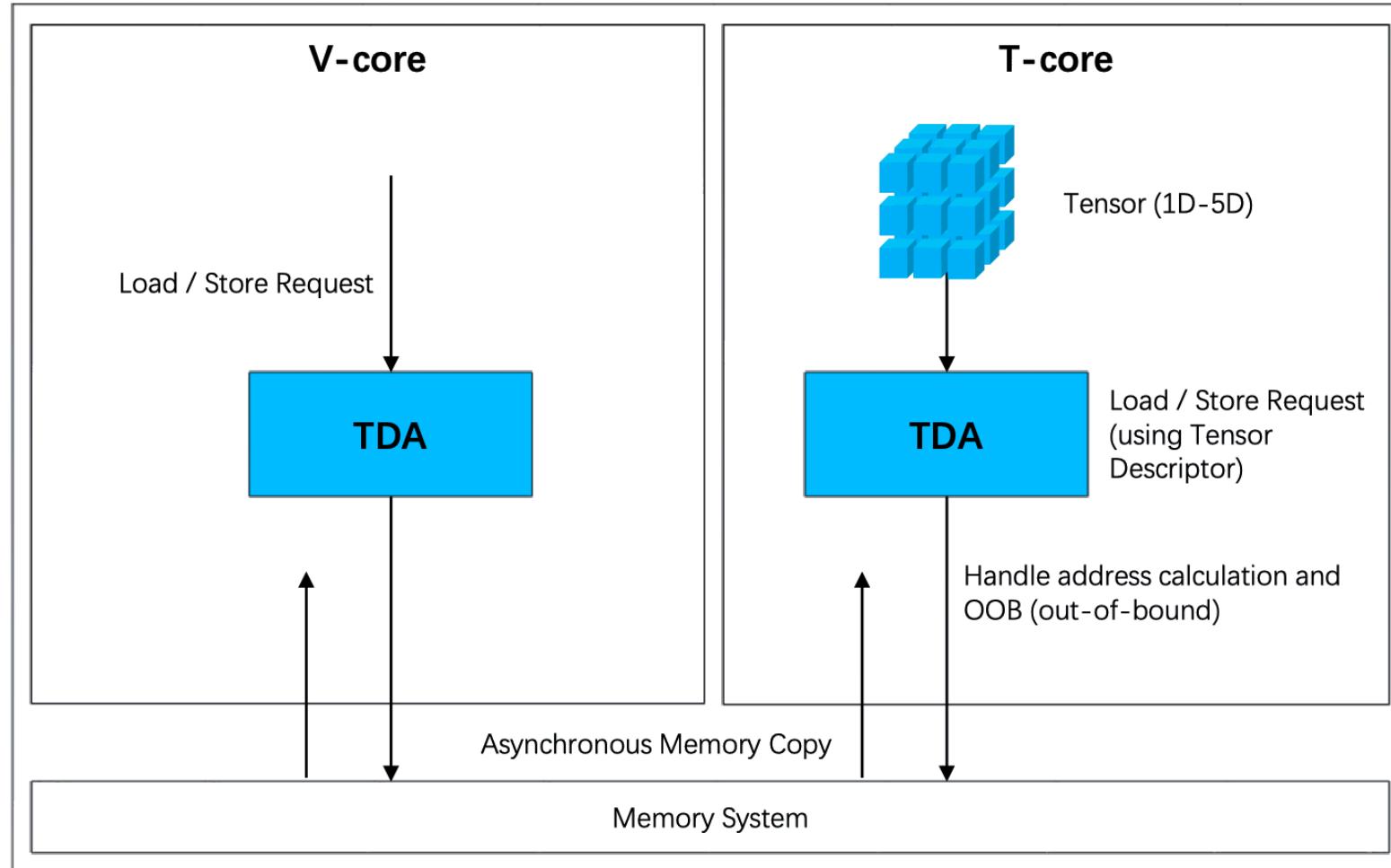
- MM
- Conv
- ...

# GEMM 计算架构



- 大尺寸的GEMM能增加数据 Reuse , 提升数据复用度，减少外存带宽诉求和功耗
- 16 个 T-Core 组成一个 2D 的脉冲阵列
- 每个 T-Core 提供 2 组  $8 \times 8$  外积计算
- 一个 SPC 相当于提供  $64 \times 64$  的MM计算
- 缺点：GEMM不是越大越好

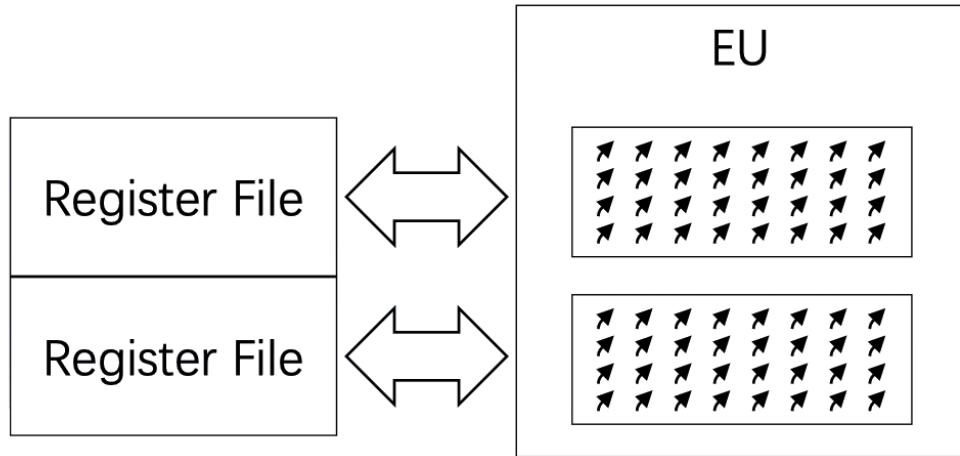
# Tensor Data Accelerator (TDA) 张量数据加速



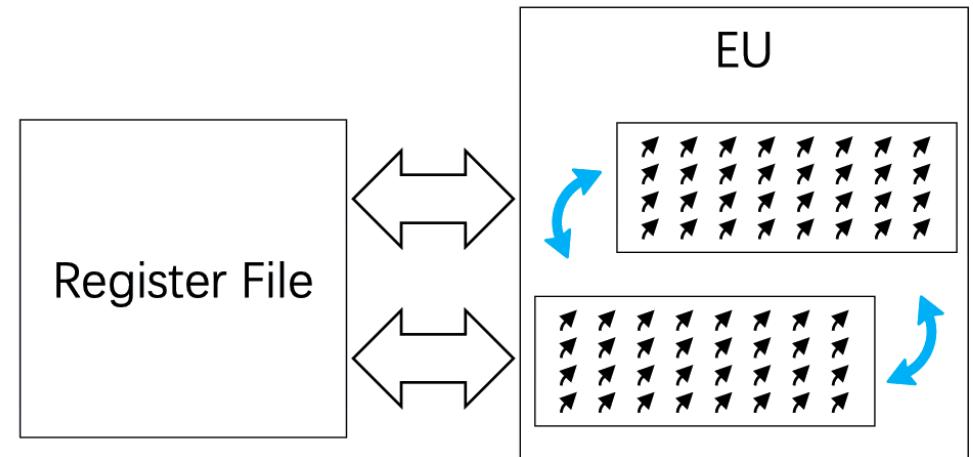
- **TDA** : Tensor Data Accelerator(TDA) , 张量数据存取 L/S 加速专用硬件。
- **作用** : 负责计算单元的数据存取工作 , 实现数据地址计算、数据同步。让计算和数据搬动实现硬件异步。

# V-Core Warp 线程控制

标准的控制模式



C-Warp/Kernel 协作控制模式

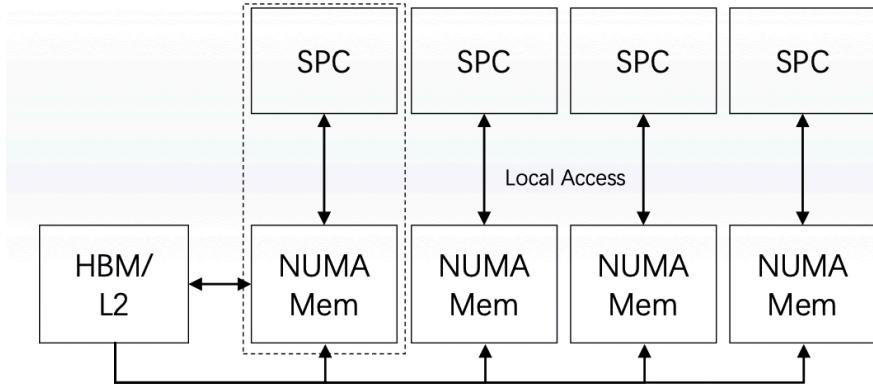


- Warps对寄存器文件进行分区，控制线程执行相同代码

- Warps 执行不同的代码，在 EU 中可以灵活在寄存器文件中交换数据

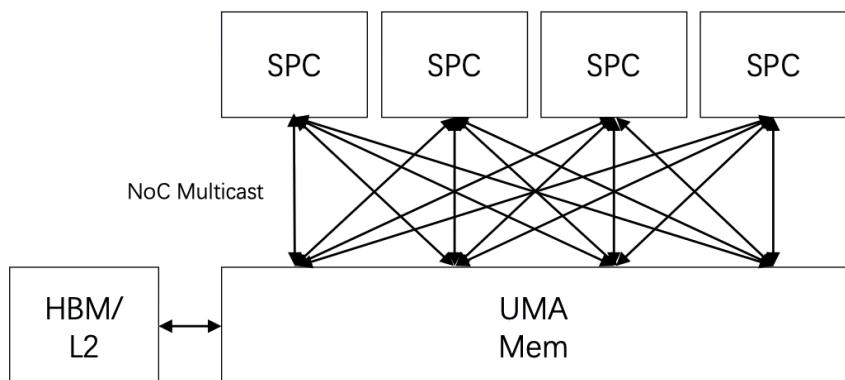
通过Register File来直接传递数据，也就减少了数据的搬移

# NUMA/UMA 访存机制



## NUMA

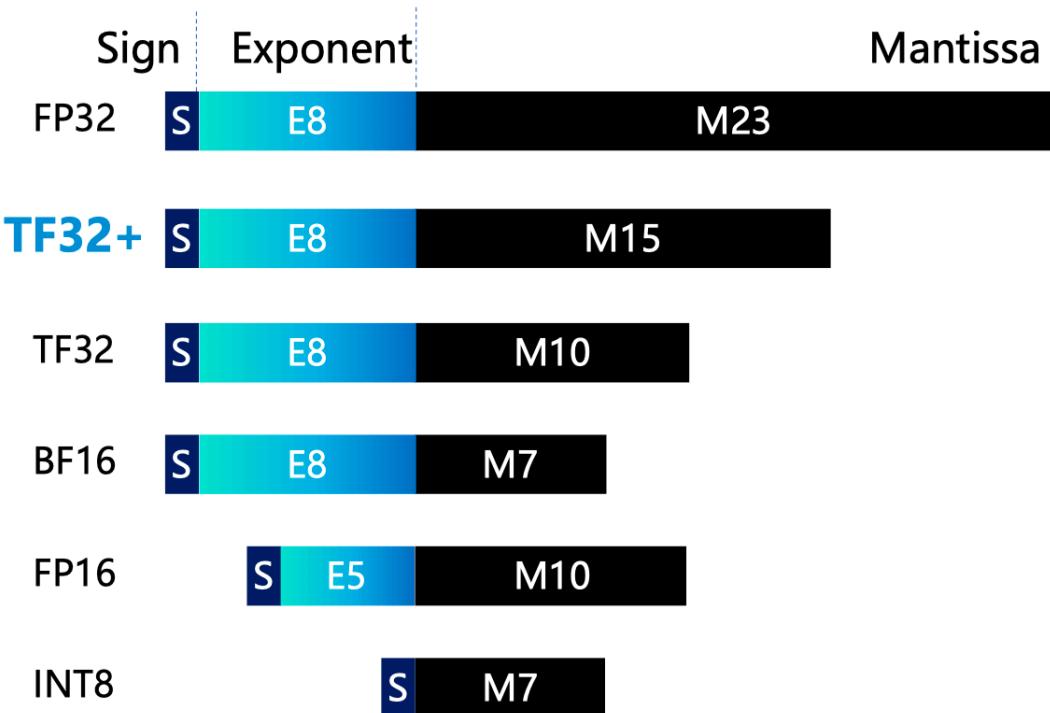
- 激活函数具有局部性，由SPC本地缓存读取，提升访问带宽
- 通过高带宽实现局部数据读取，减少 Memory Bound



## UMA

- 权重数据由SPC共享
- DoC多播加速权重数据读取，减小搬运

# 数据格式



- E8M15：指数位与FP32形同，位数比TF32多5位
- 相比 TF32 拥有更高数据精度
- 相比V100 TF32 有更高吞吐
- 软件系统自动切换数据格式
- 满足更多 AI 训练场景

# 5. 对壁仞的思考



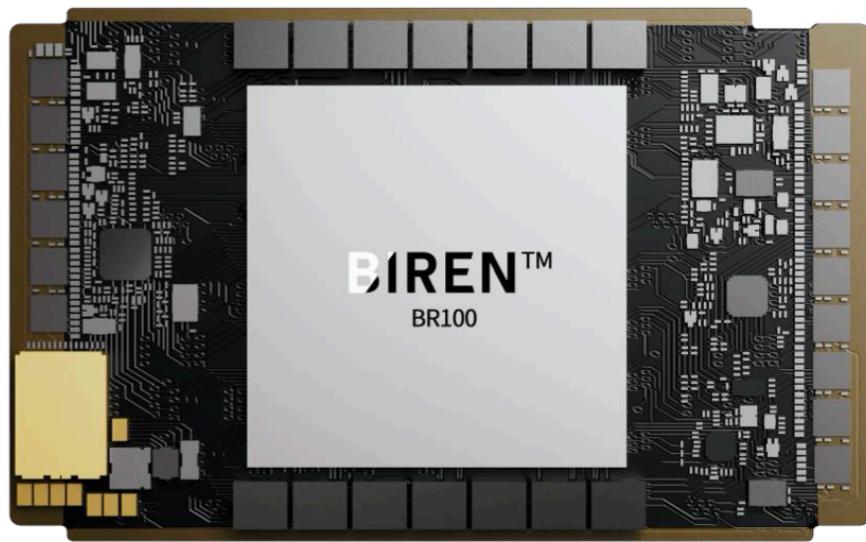
# 对格式的思考

- **数据格式**：目前峰值算力主要来源是 1024 TFLOPS @ BF16 & 512 TFLOPS @ TF32+，跟主流计算芯片 FP16 & FP32 数据格式不完全一致，对部分模型需要重新预训练或者二次训练，算力的增长与成本的增加如何平衡？

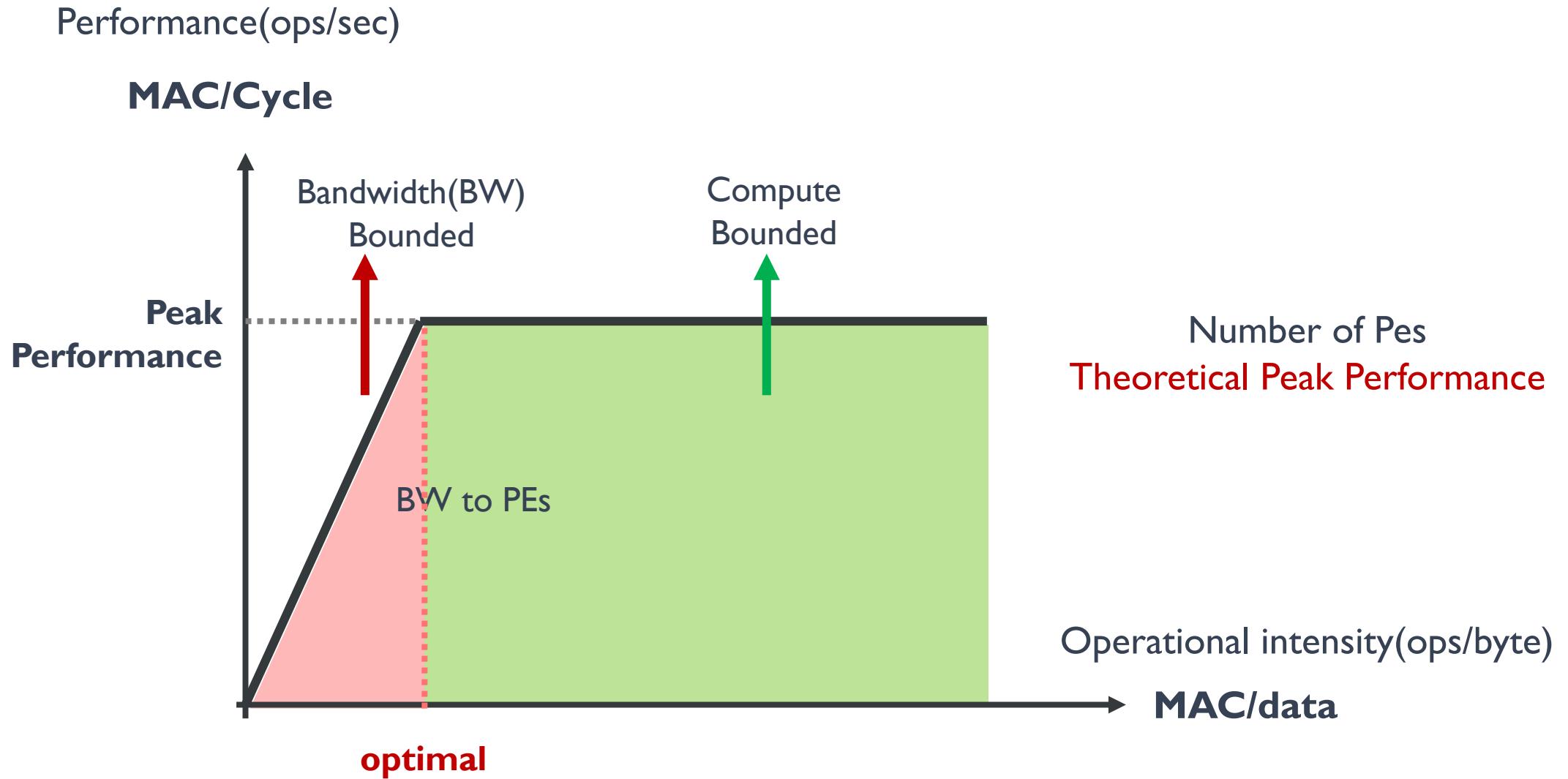


# 对计算的思考

- **通用计算**：通用计算能力是任何一款 GPU 芯片的根本，目前的测评过于强调 AI 能力，针对通用计算、HPC、科学计算等通用计算场景的软件兼容性、性能会呈现什么样的形式？
- **计算密集度**：对AI计算进行流处理器优化，芯片电路面积上，会牺牲通用计算能力来换取 AI 计算峰值。最终会不会沦为类GPU架构下AI芯片？

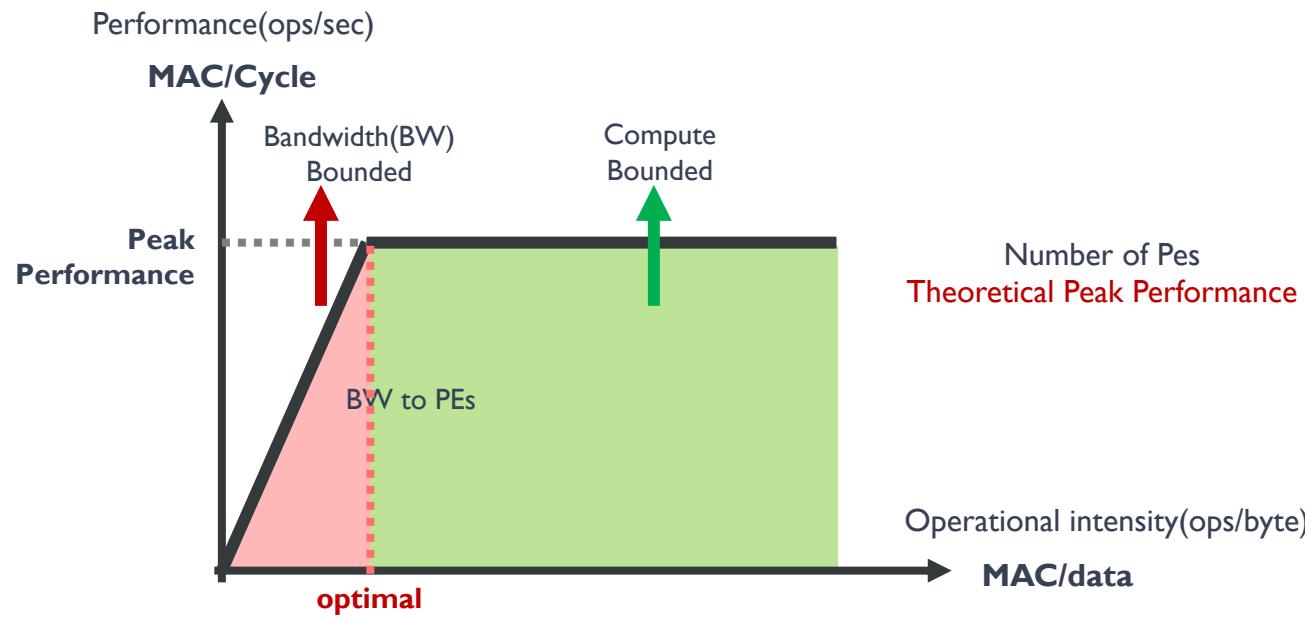


# 对利用率的思考



# 对利用率的思考

- **Give more than need** : 内部算力带宽超过外部数据带宽，作为重数据的AI应用利用率猜测会维持一个较高水平，但是对于通用计算场景会不会导致有效算力的利用率急剧下降？
- **数据流近存计算** : 高速的片内带宽、较大的 SRAM 决定壁仞的芯片架构类似于数据流近存架构，主要面向类 GPU 计算模式的 AI 计算应用，软件栈相对于 CUDA 类似。带宽和计算平衡点在哪里？



# 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

**Copyright©**

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. May change the information at any time without notice.