壁仞芯片剖析

壁仞科技 智绘全球

RENDER THE WORLD WITH INTELLIGENCE

AI 芯片 ZOMI

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

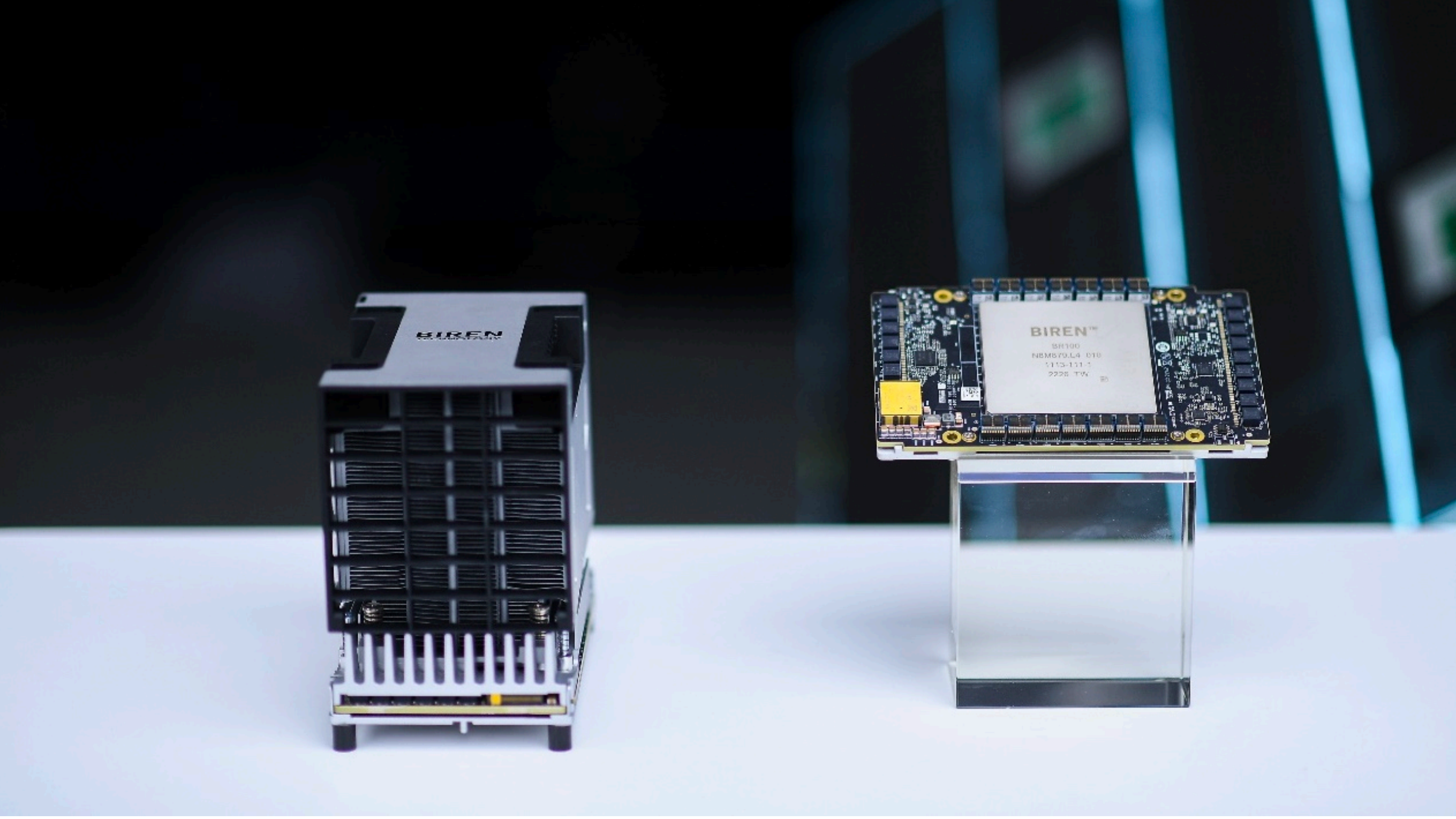
- 壁仞 芯片剖析
- 寒武纪 芯片剖析
- 燧原科技 芯片剖析
- AI 芯片架构的思考

目录 Context

1. 什么是壁仞

壁仞科技：目标成为世界领先的智能计算系统企业

- 累计融资超50亿元人民币。创全球同行业初创企业融资纪录，成为成长最快的独角兽企业之一，顶尖投资机构 不仅给予资金支持，更带来了丰富的产业和战略资源。


BIREN
BIREN™

2. 壁仞产品形态

BR100系列 通用GPU芯片

1024T@BF16

超高性能与能效比

7nm与Chiplet

先进制程与封装工艺

PCIe5.0与CXL

先进接口系统

BLink™

先进互连系统

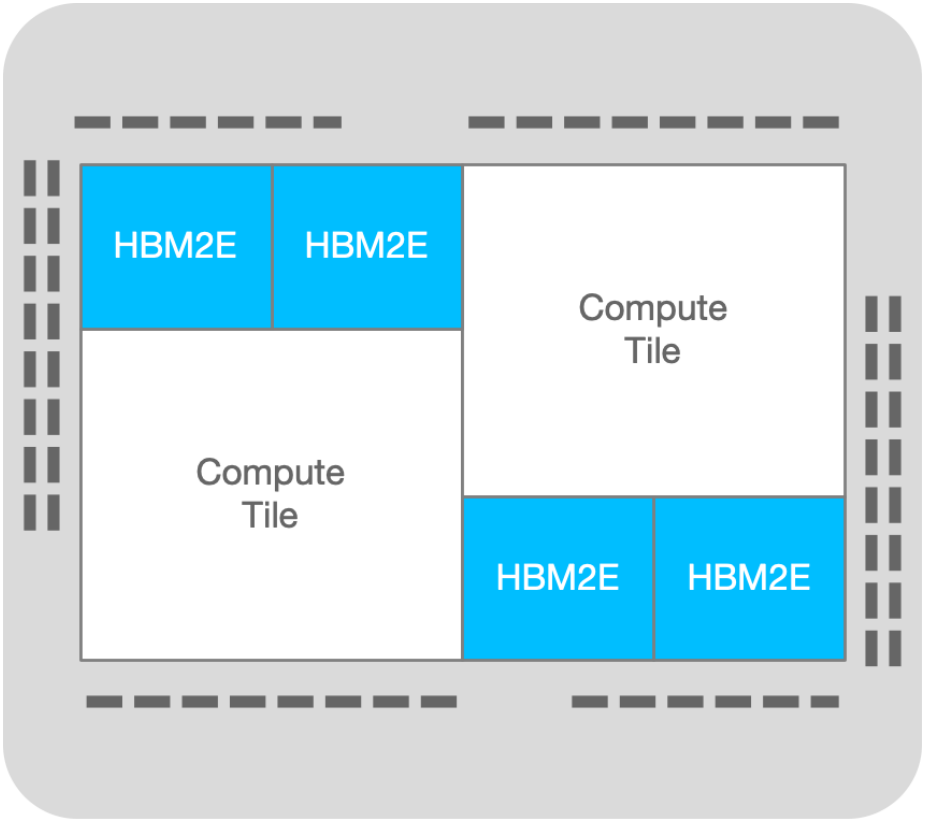
浮点与整数型

多数据精度支持

最高支持8份

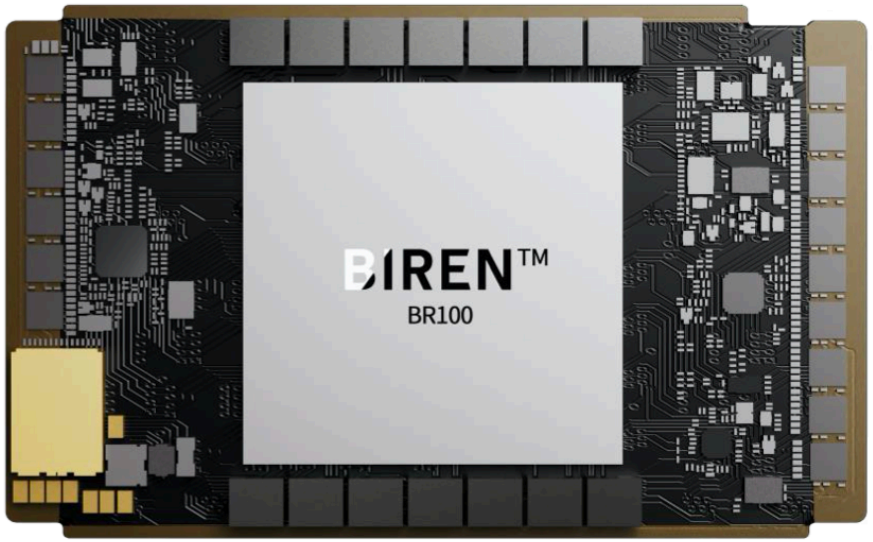
安全虚拟实例（SVI）

双 Die 设计



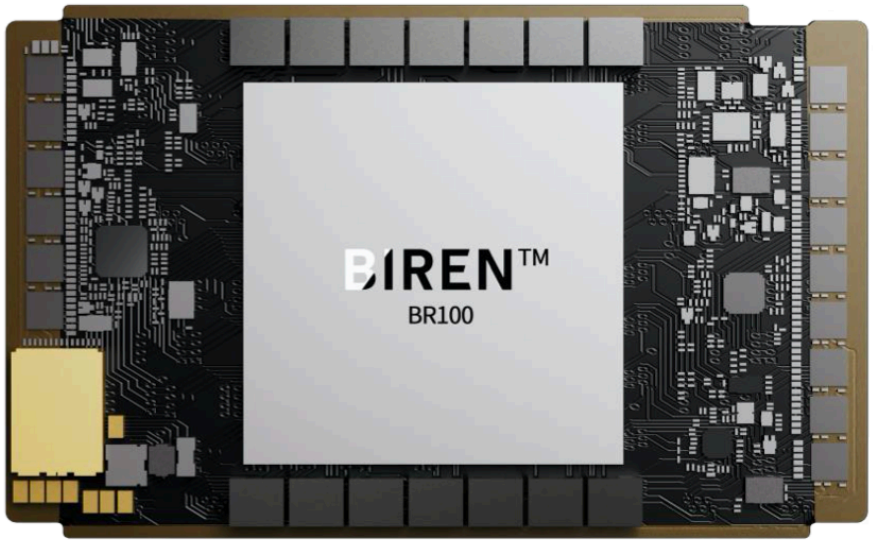
- 打破掩模版尺寸限制，在芯片上集成更多晶体管
- 一次流片可以封装多个产品
- 模具更小，产量更高，成本更低
- Die to Die 提供 896GB/s 高速互连带宽
- 相比单芯片设计有 30% 性能和 20% 产能提升

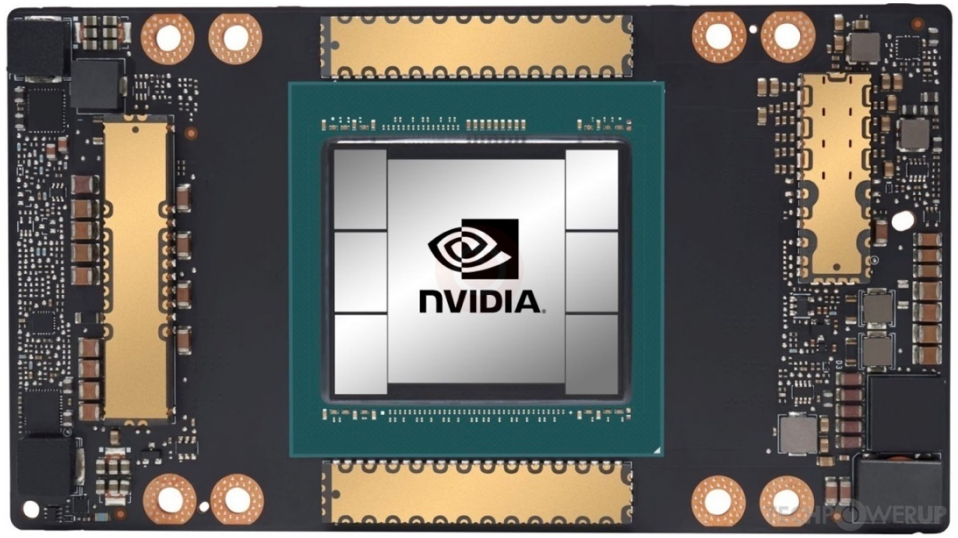
壁砺™ 100



芯片	BR100
制程	1074mm^2 @7nm (77Billion)
峰值性能	2048 TOPS @ INT8 1024 TFLOPS @ BF16 512 TFLOPS @ TF32+ 256 TFLOPS @ FP32 支持FP16, INT32, INT16等精度类型
系统接口、带宽、互连协议	PCIe 5.0 X16，128GB/s，支持CXL
内存容量、接口位宽、带宽	64GB HBM2E，4096bit，1.64TB/s
互连	512 GB/s BLink™
安全虚拟实例	最高8份
编解码(FHD@30fps)	64路HEVC/H.264编码 512路HEVC/H.264解码
功耗	550W
产品形态	OAM模组

壁砺™ 100 vs NV A100



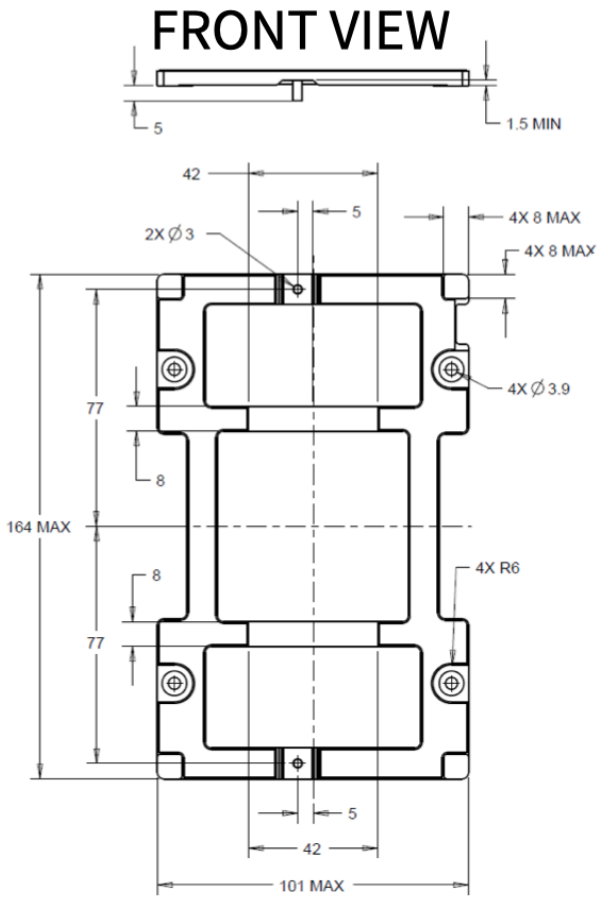


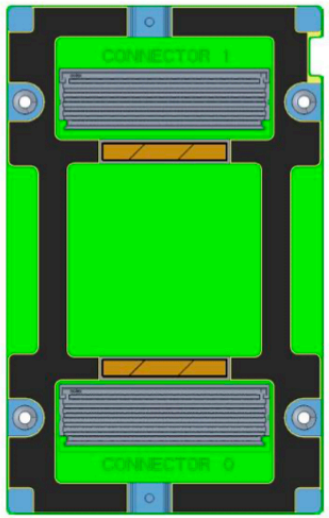
壁砺™100P

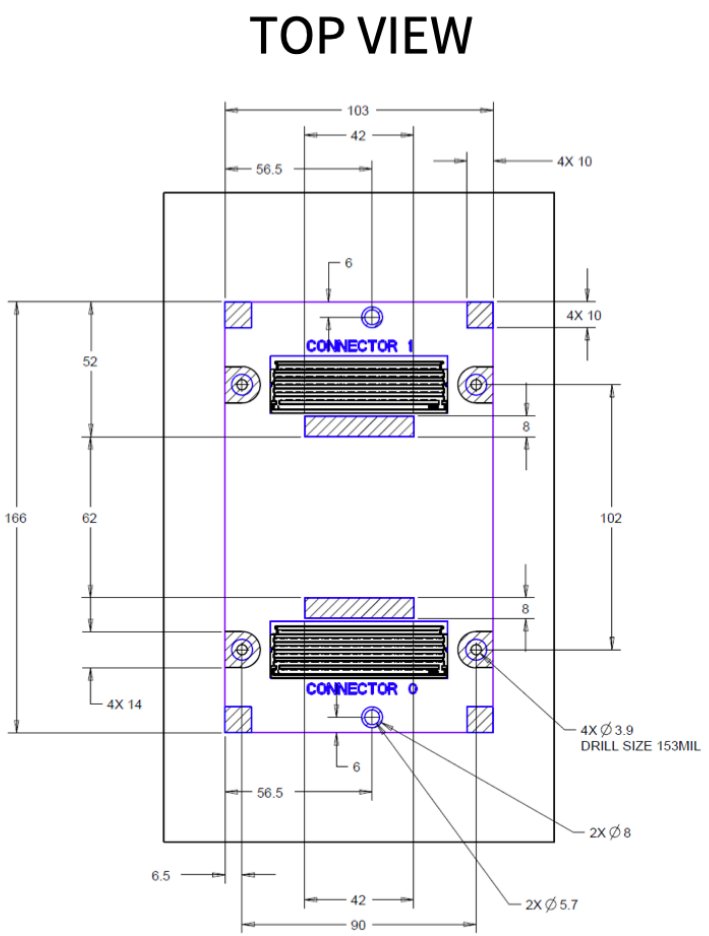
壁砺™100P产品形态为OAM模组

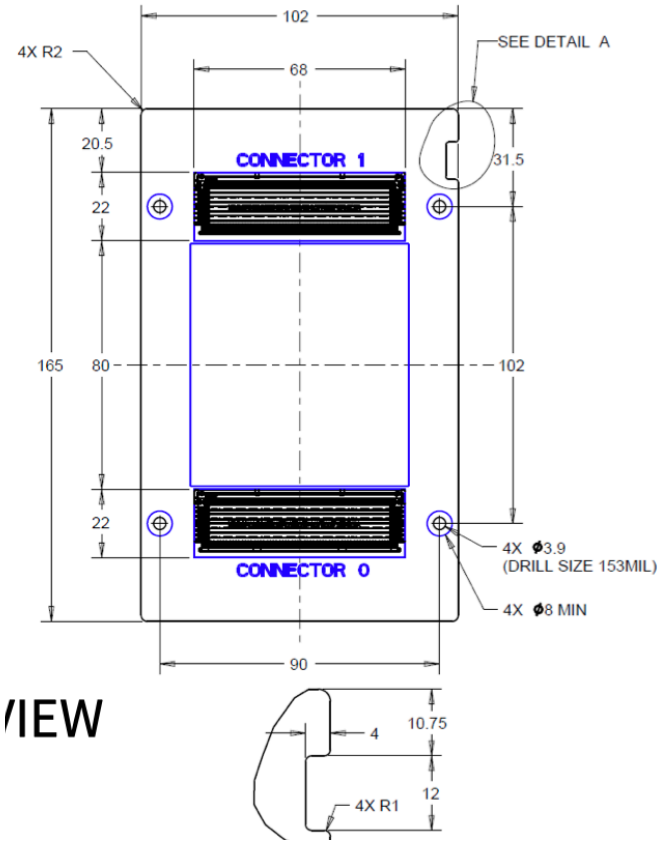
- **OAM**：OCP Accelerator Module(OAM)，OCP 开放加速模组。
- **目的**：1）提供一个基本框架，使不同芯片供应商的OAM可以在同一系统中使用。2）提供一个完整的参考设计，使重新设计最小化。
- **内容**：提供底盘平台、散热器、夹层、定位销、螺钉安装孔、底版、EMI垫圈、PCB 版等标准。

OAM，OCP 开放加速模组

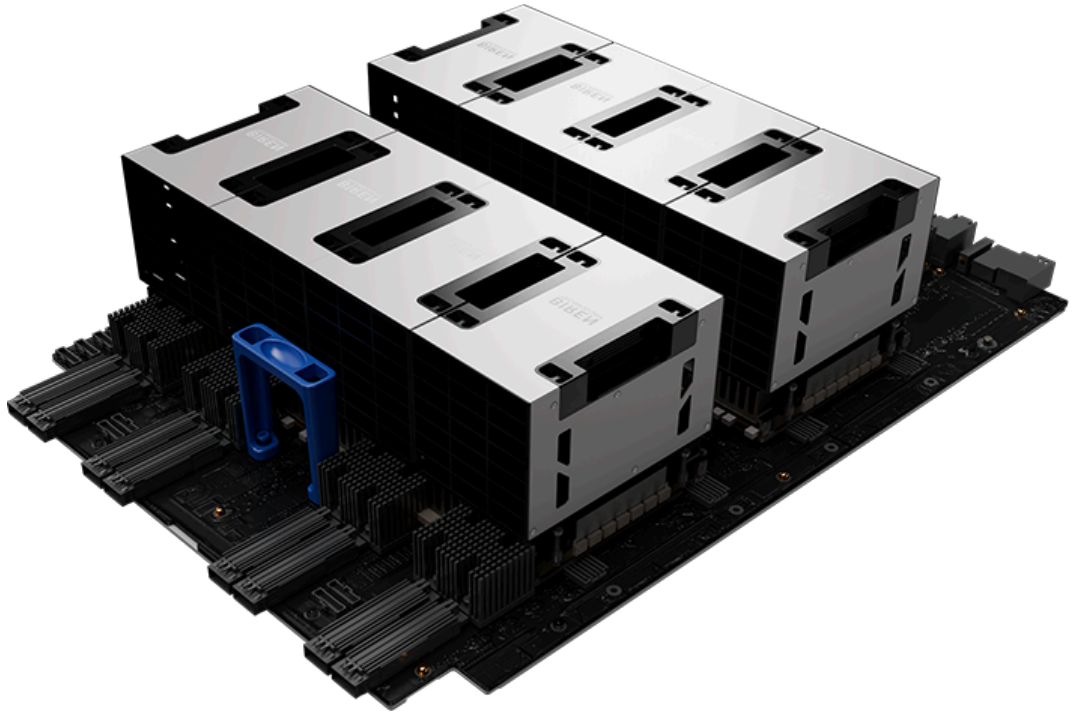








壁砺™100 UBB



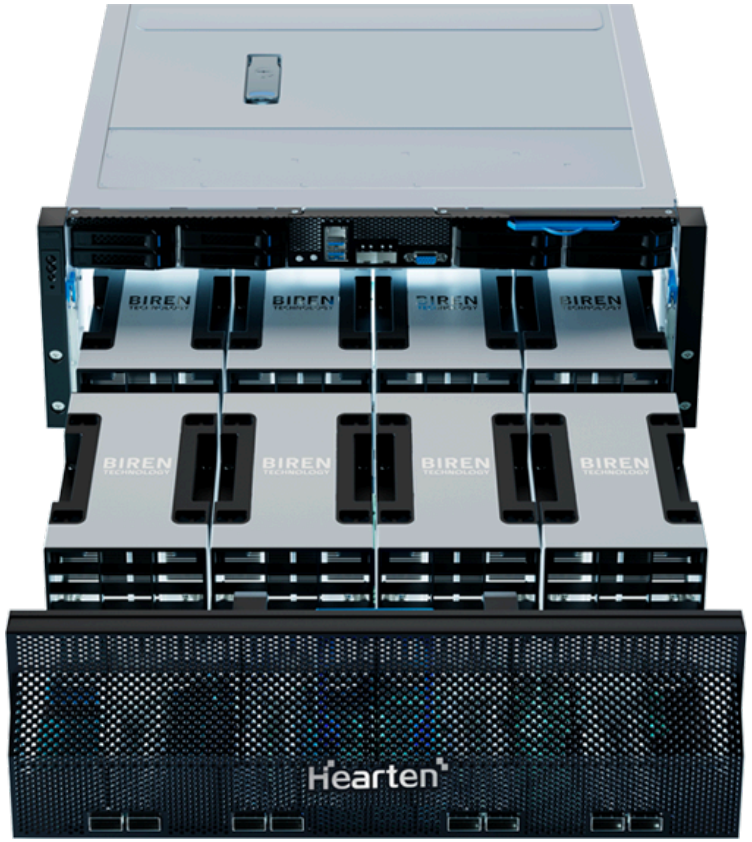
模组基于OCP UBB v1.0标准开发

搭载8张壁砺™100P通用GPU

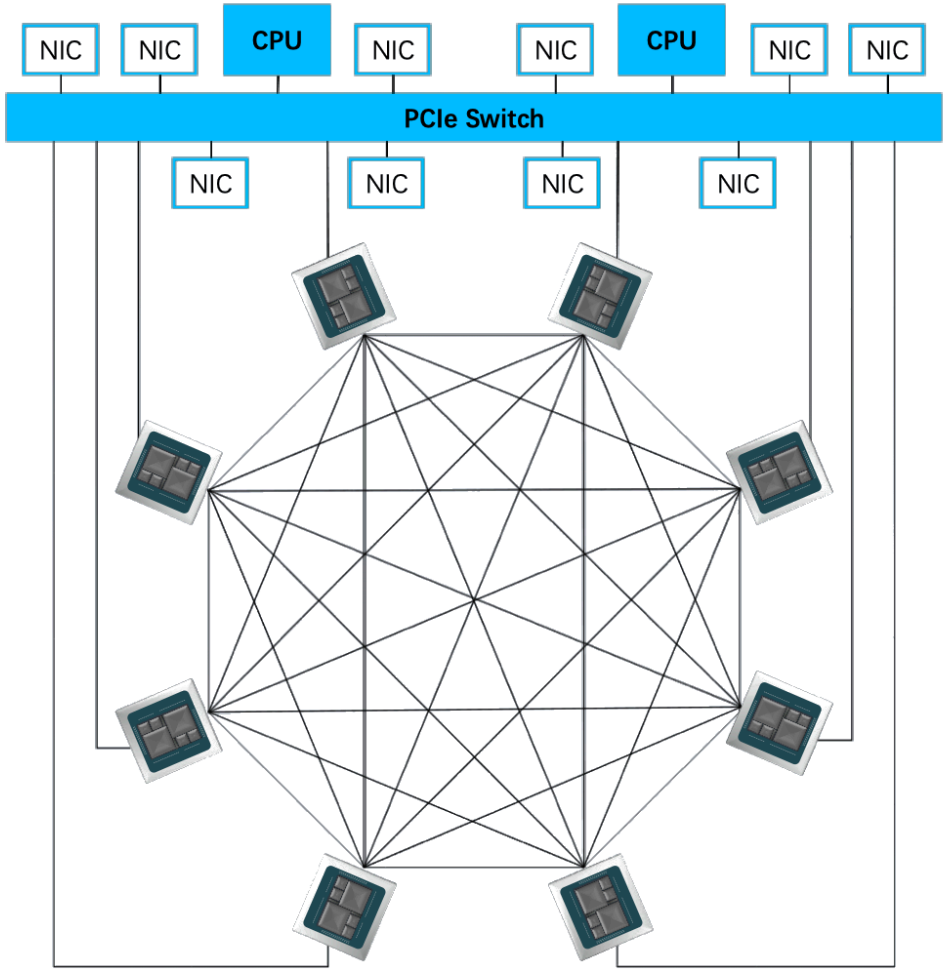
- OCP：An Open Accelerator Infrastructure(OCP) Project，开放计算项目。
- UBB：Universal Baseboard (UBB)，统一基板。

- **目的**：OCP 系统 和 UBB 标准设计，让开源技术更广泛的推广到产业链，使产业配套更加简化。

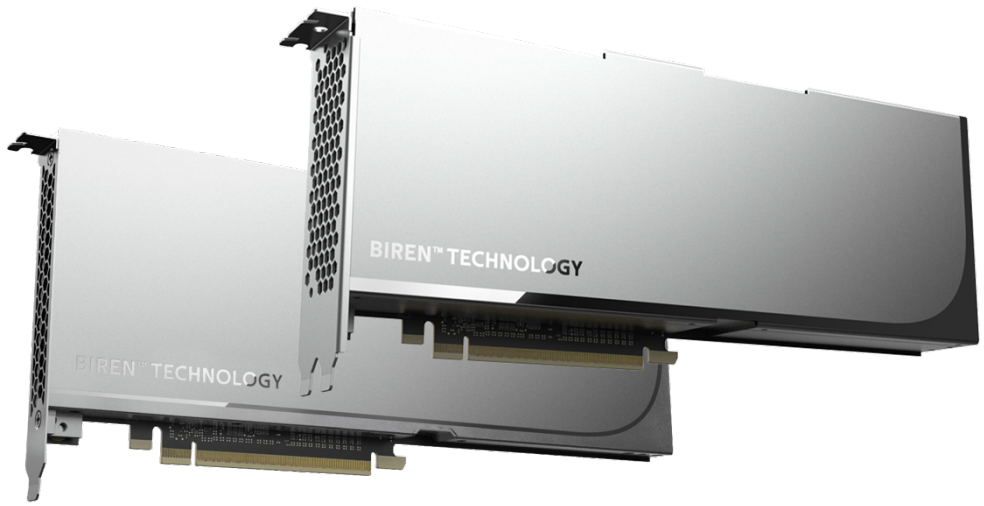
海玄服务器



搭载 8 个壁砺™ 100P OAM 模组



壁砺™ 104



芯片	BR104
制程	7nm
峰值性能	1024 TOPS @ INT8 512 TFLOPS @ BF16 256 TFLOPS @ TF32+ 128 TFLOPS @ FP32 支持FP16, INT32, INT16等精度类型
系统接口、带宽、互连协议	PCIe 5.0 X16，128GB/s，支持CXL
内存容量、接口位宽、带宽	32GB HBM2E，2048bit，819GB/s
互连	192 GB/s BLink™
安全虚拟实例	最高4份
编解码(FHD@30fps)	32路HEVC/H.264编码 256路HEVC/H.264解码
功耗	300W
产品形态	全高全长，双槽位PCIe卡

壁砺™ 100 vs 壁砺™ 104

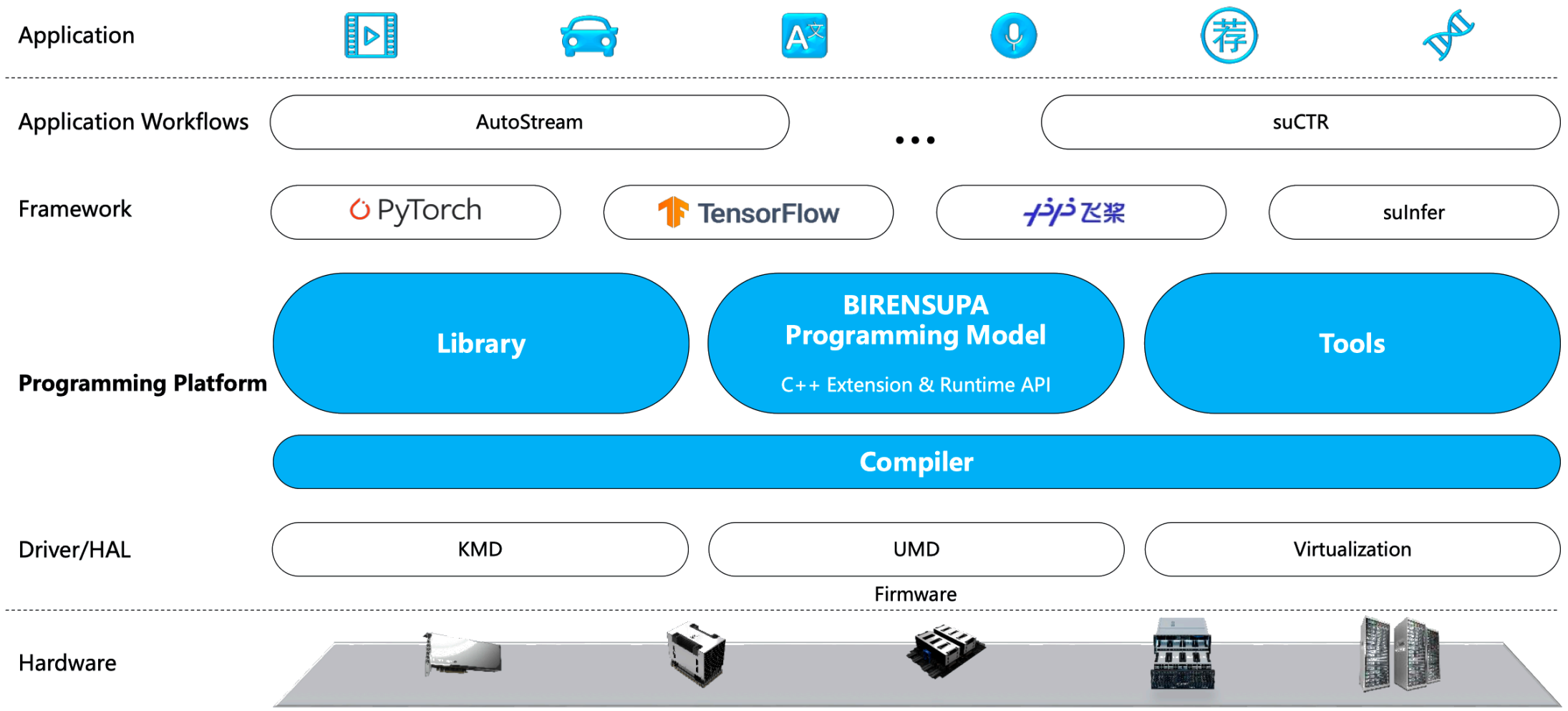
芯片	BR100	BR104
制程	7nm	7nm
峰值性能	2048 TOPS @ INT8 1024 TFLOPS @ BF16 512 TFLOPS @ TF32+ 256 TFLOPS @ FP32 支持FP16, INT32, INT16等精度类型	1024 TOPS @ INT8 512 TFLOPS @ BF16 256 TFLOPS @ TF32+ 128 TFLOPS @ FP32 支持FP16, INT32, INT16等精度类型
系统接口、带宽、互连协议	PCIe 5.0 X16，128GB/s，支持CXL	PCIe 5.0 X16，128GB/s，支持CXL
内存容量、接口位宽、带宽	64GB HBM2E，4096bit，1.64TB/s	32GB HBM2E，2048bit，819GB/s
互连	512 GB/s BLink™	192 GB/s BLink™
安全虚拟实例	最高8份	最高4份
编解码(FHD@30fps)	64路HEVC/H.264编码 512路HEVC/H.264解码	32路HEVC/H.264编码 256路HEVC/H.264解码
功耗	550W	300W
产品形态	OAM模组	全高全长，双槽位PCIe卡

壁砺™ 100 vs NV A100/H100

	BIREN BR100	NV A100	NV H100
制造工艺	TSMC N7 @ 77B 1074mm^2	TSMC N7 @ 54.2B 828mm^2	TSMC N4@80B 814mm^2
算力	2048 TOPS @ INT8 1024 TFLOPS @ BF16 512 TFLOPS @ TF32+ 256 TFLOPS @ FP32 支持FP16, INT32, INT16等类型	624 TOPS @ INT8 312 TFLOPS @ BF16 312 TFLOPS @ FP16 156 TFLOPS @ TF32 支持 FP32、FP64 等类型	4000 TOPS @ INT8 2000 TFLOPS @ BF16 2000 TFLOPS @ FP16 1000 TFLOPS @ TF32 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)

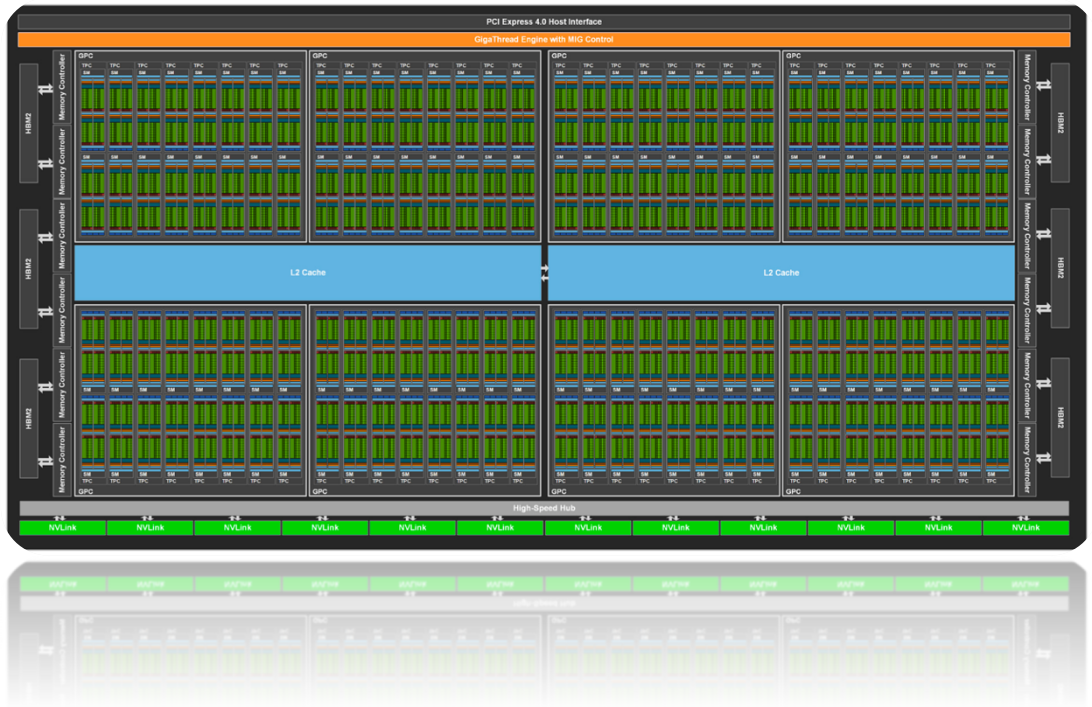
3. 软件平台

BIRENSUPA 端到端全覆盖软件平台 — 整体系统解决方案

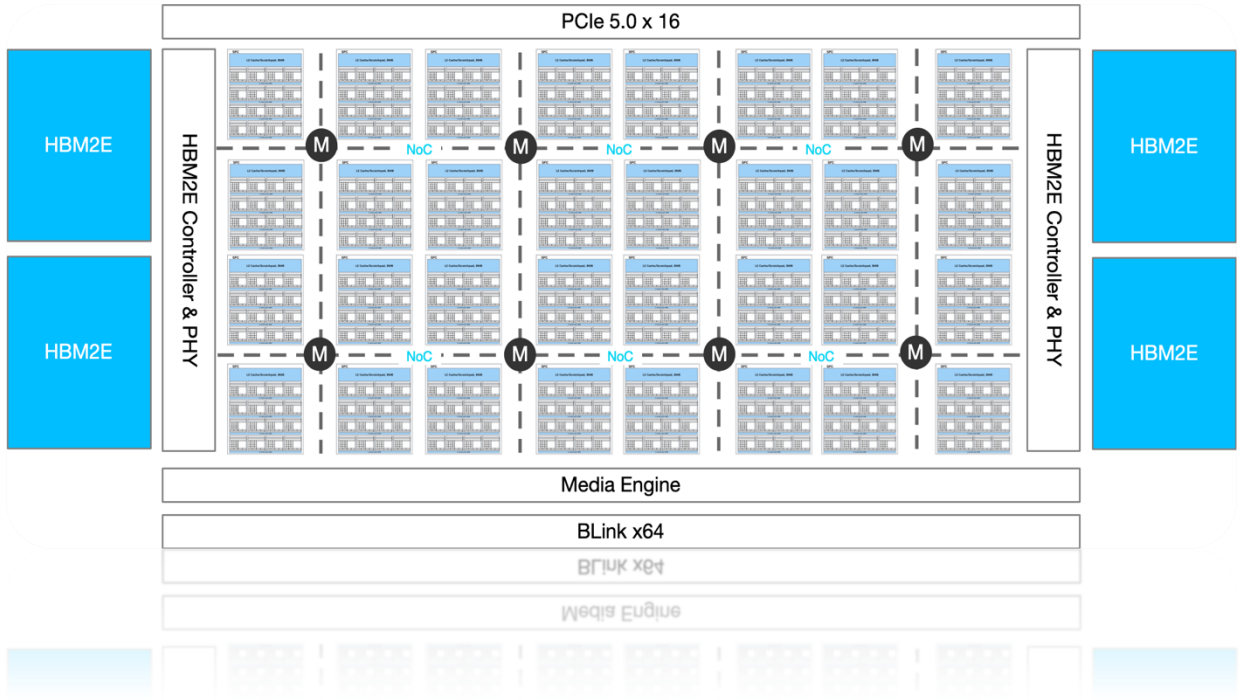


类 GPU 架构可提供相类似的 SIMT 编程方式

A100 GPU 芯片架构



BR100 GPU 芯片架构



BIRENSUPA 编程模型SDK与加速库 —— 编程模式与生态库

核心SDK

提供基于C/C++的GPU编程支持

- SUPA 编程模型
- BRCC 编译器

加速库

高性能实现，快速支持应用构建

- 深度学习算子库
- 多卡通讯库 SCCL

工具集

完整支持GPU管理与应用开发

- GPU 卡管理接口
- 性能采集与分析

Question

- **软件生态：**CUDA 让你兼容了吗？为什么 AMD、APPLE、TPU 都不去做这个事情？是因为 Google、AMD 等工程师没意识到生态兼容的重要性？还是 NVIDIA 专利墙厚？
- **产品化：**目前大部分国产AI芯片/GPU芯片主要停留在小规模供货/PPT官宣，壁仞科技的产品什么时候才能产品化，规模化出货？ TSMC 在两片 die 封装的良品率是多少？美国禁止代工问题如何解决？

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