寒武纪 の MLU03芯片架构

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

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

- 寒武纪的产品形态
- 寒武纪MLU03芯片架构
- 寒武纪软件栈和通信

Talk Overview

- 不是为了吐槽友商和国产 AI 芯片
- 通过实际产品，深入理解 AI 芯片相关技术
- 洞察不同厂商芯片架构，从而思考 AI 芯片的未来发展方向

1. 架构发展

历经5代AI芯片架构

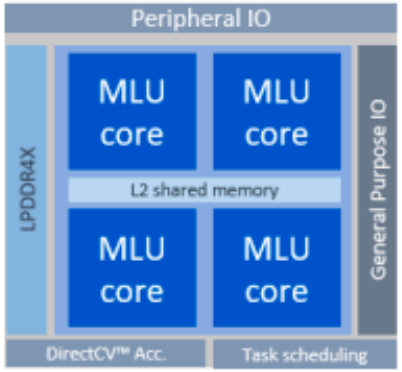
- 麒麟980
- 麒麟970

- 思元100

- 思元220
- 思元270
- 思元290

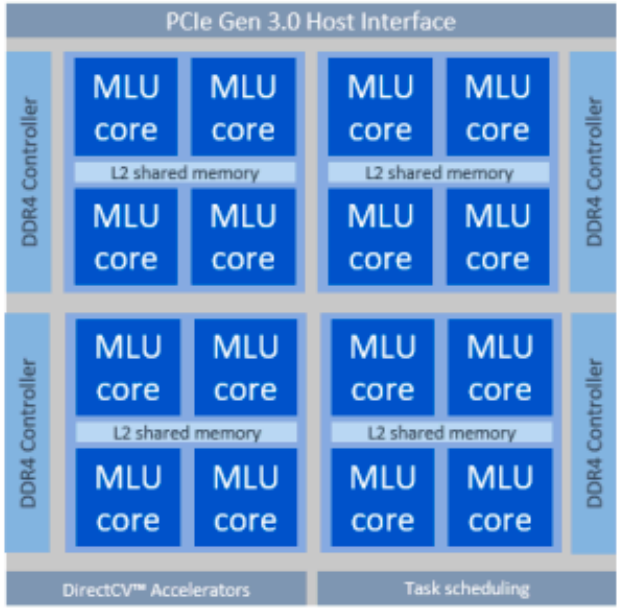
- 思元370

整体架构



MLU220

6MB SRAM
LPDDR4x



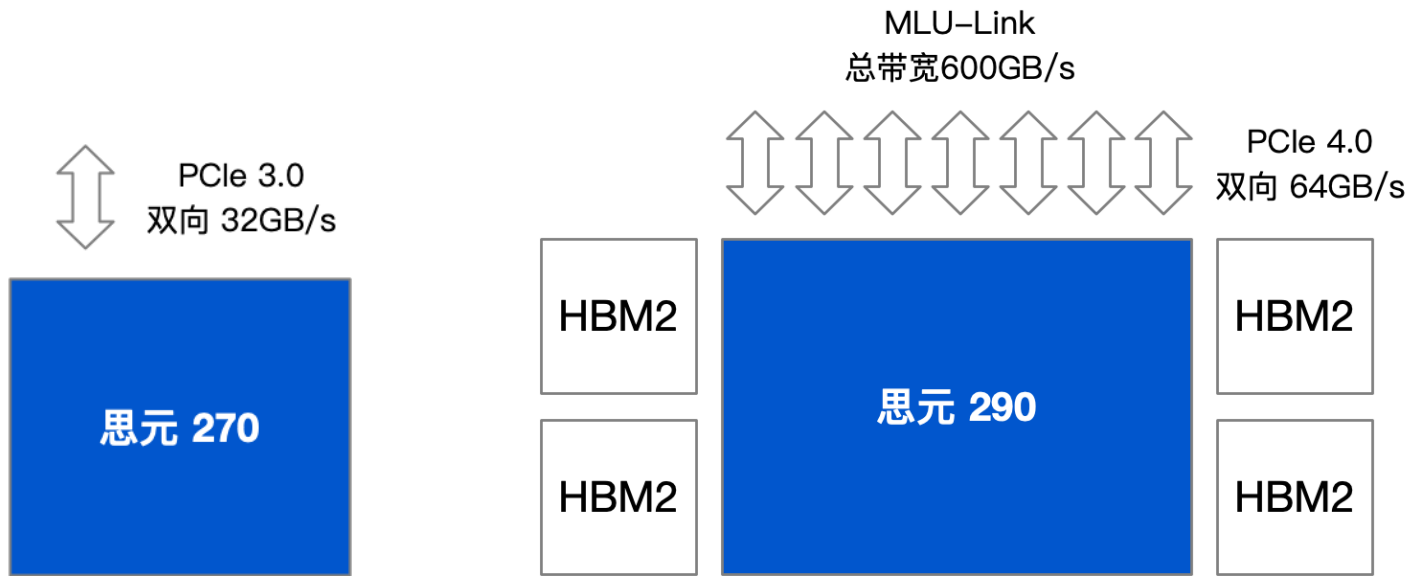
MLU270

32MB SRAM
DDR4

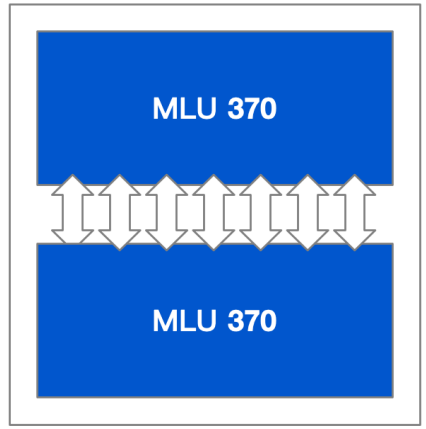
MLU290

96MB SRAM
HBM2
MLU-Link™

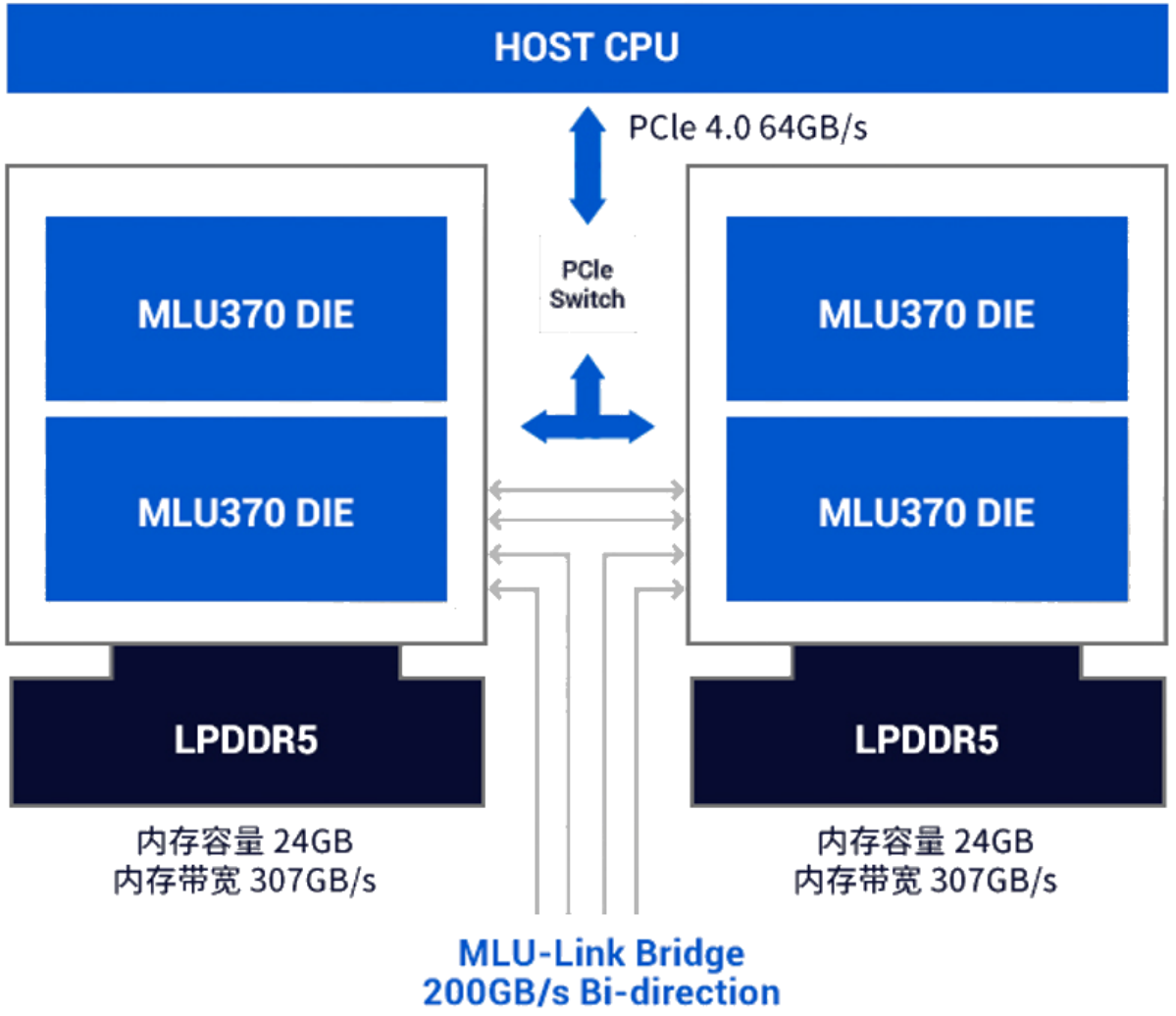
芯颗封装



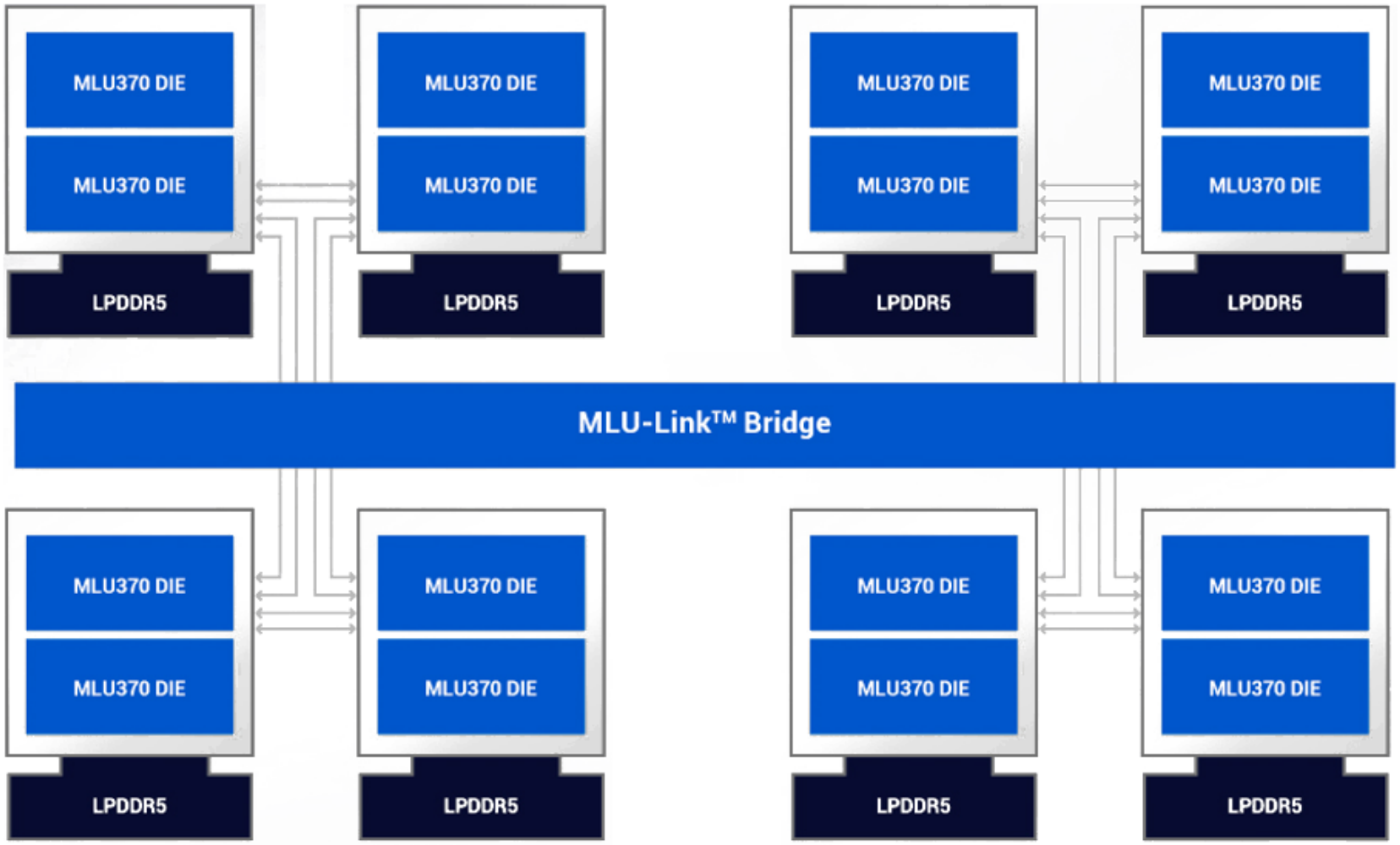
- 2颗DIE封装
- 全局UMA内存访问
- 低功耗、低时延、高带宽



产品形态 I


HOST CPU
PCIe 4.0 64GB/s
PCIe Switch
MLU370 DIE
MLU370 DIE
MLU370 DIE
MLU370 DIE
LPDDR5
LPDDR5
内存容量 24GB
内存带宽 307GB/s
内存容量 24GB
内存带宽 307GB/s
MLU-Link Bridge
200GB/s Bi-direction

产品形态 II


MLU370 DIE
MLU370 DIE
LPDDR5
MLU370 DIE
MLU370 DIE
LPDDR5
MLU370 DIE
MLU370 DIE
LPDDR5
MLU370 DIE
MLU370 DIE
LPDDR5
MLU-Link™ Bridge
MLU370 DIE
MLU370 DIE
LPDDR5
MLU370 DIE
MLU370 DIE
LPDDR5
MLU370 DIE
MLU370 DIE
LPDDR5
MLU370 DIE
MLU370 DIE
LPDDR5

2. MLU03

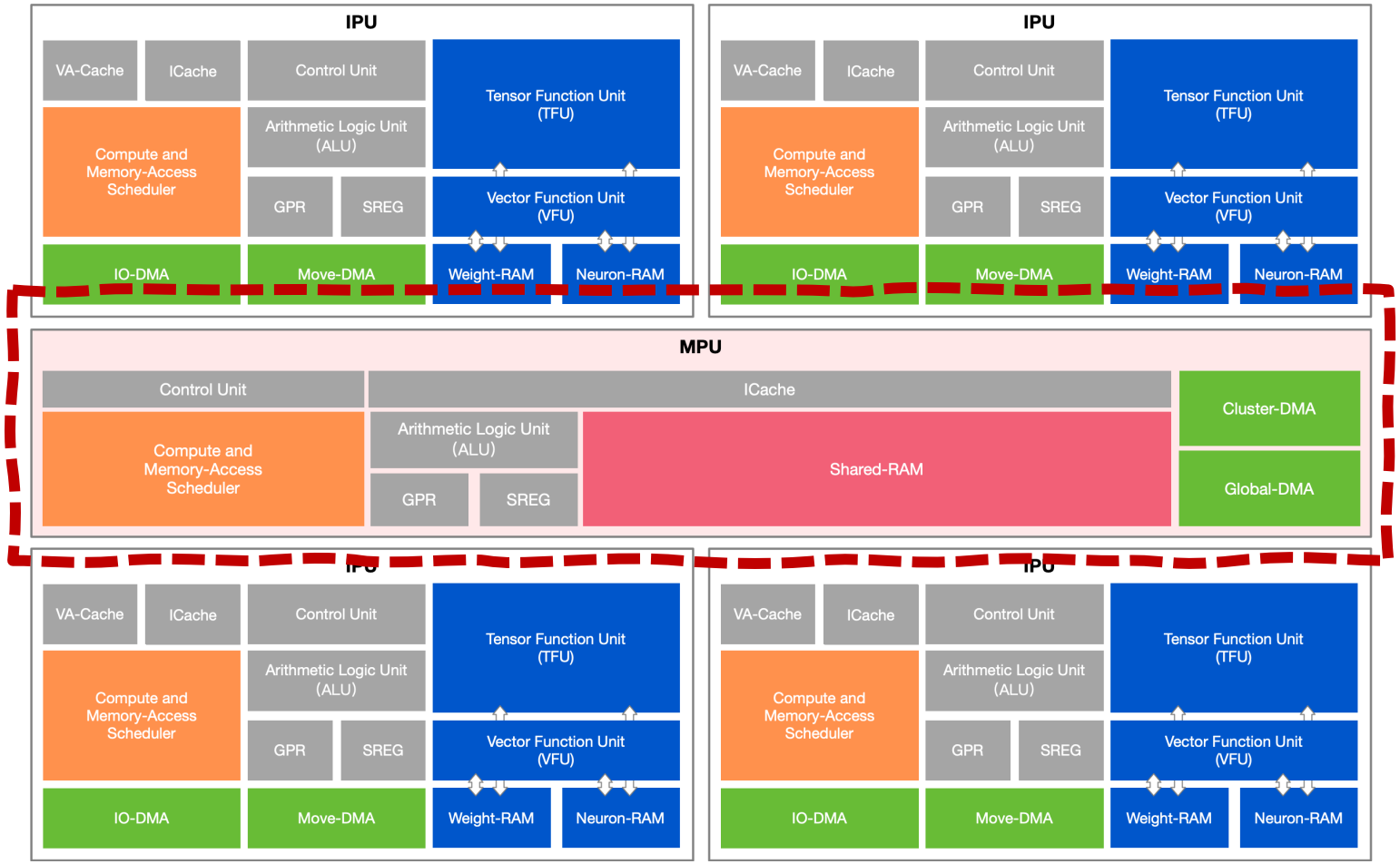
整体架构

MLUv03 - Cluster

- IPU：Intelligence，作为 MLU 核心模块，负责 AI 应用的计算、访存和控制指令执行。
- MPU：Memory，MLU中负责片上Shared-RAM相关协处理器核心，对数据进行处理。
- Cluster：NLUv03架构，4个IPU和1个MPU作为核心构成1个Cluster。多个Cluster组合不同产品形态。

MLUv03 - Cluster

- IPU：Intelligence，作为 MLU 核心模块，负责 AI 应用的计算、访存和控制指令执行。
- MPU：Memory，MLU中负责片上Shared-RAM相关协处理器核心，对数据进行处理。
- Cluster：NLUv03架构，4个IPU和1个MPU作为核心构成1个Cluster。多个Cluster组合不同产品形态。



MLUv03 - Cluster

- IPU：Intelligence，作为 MLU 核心模块，负责 AI 应用的计算、访存和控制指令执行。
- MPU：Memory，MLU中负责片上Shared-RAM相关协处理器核心，对数据进行处理。
- Cluster：NLUv03架构，4个IPU和1个MPU作为核心构成1个Cluster。多个Cluster组合不同产品形态。

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem