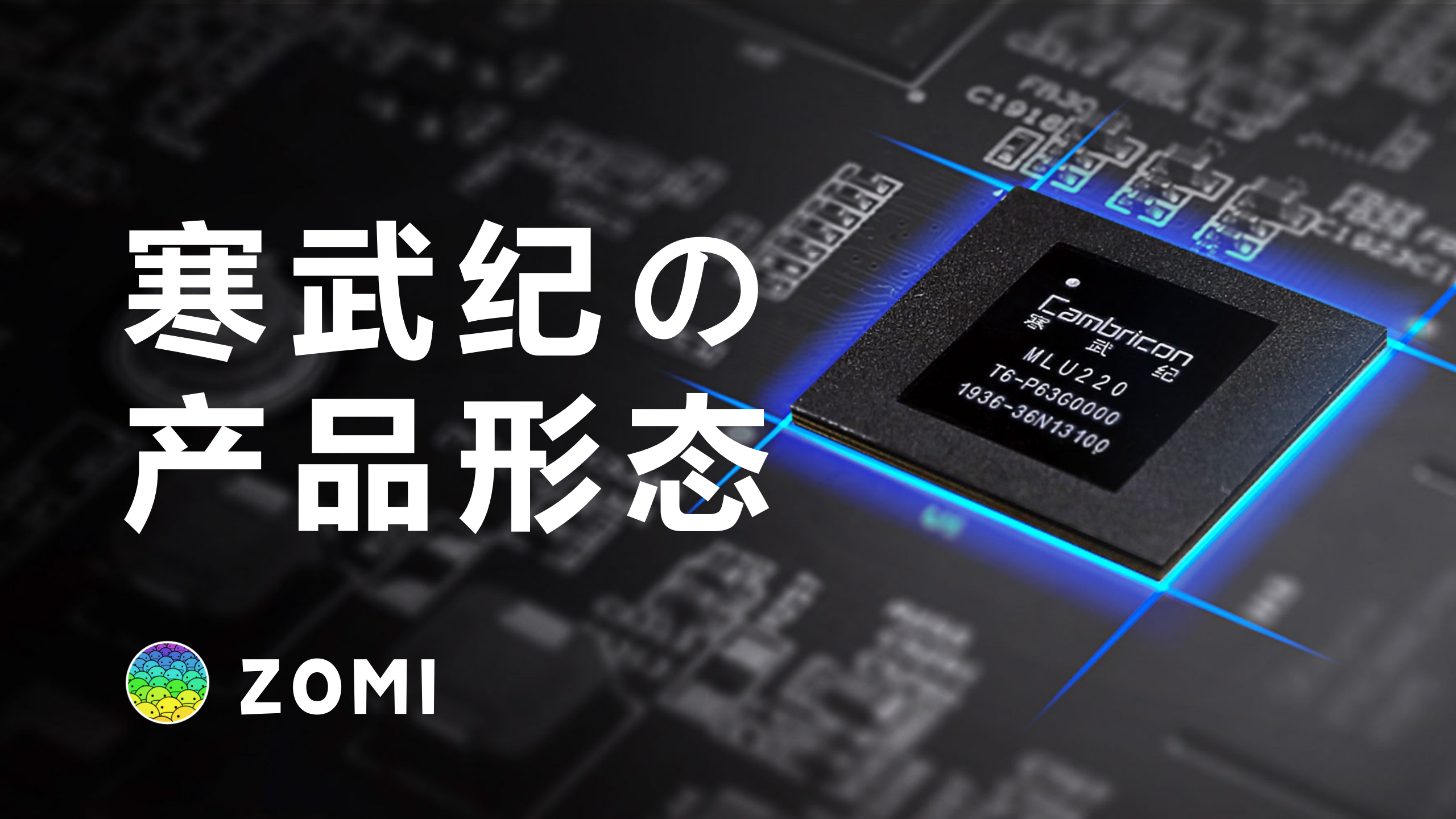
寒武纪の产品形态



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

1. 寒武纪是什么

寒武纪 Cambricon

Cambricon =
Cambrian + Silicon

- 在5~6亿年前的寒武纪，大量较高等物种出现，物种多样性答复提升。这个现象被称为寒武纪物种大爆发。
- 先进的智能技术已呈现大爆发之势，我们希望为智能技术的大爆发提供核心物质载体。

寒武纪 Cambricon

- 成立时间：
  - 2016年3月15日
- 企业使命：
  - 为客户创造价值，成为持续创新的智能时代领导者。
- 企业愿景：
  - 寒武纪聚焦端云一体、端云融合的智能新生态，致力打造各类智能云服务器、智能终端以及智能机器人的核心处理器芯片，让机器更好地理解和服务人类。

发展历程


2016
Cambricon
寒武纪成立
2016.3.15
1A处理器
0.5 TOPS(INT8)
2017
1H处理器
1 TOPS(INT8)
2018
1M处理器
8 TOPS(INT8)
思元100
中国首颗云端智能芯片
MLU100
32 TOPS(INT8)
2019
思元270
第二代云端智能芯片
MLU270
128 TOPS(INT8)
思元220
边缘端智能SoC芯片
MLU220-M.2
8 TOPS(INT8)
2020
寒武纪
登陆科创板
688256.SH
2020.7.20
MLU220-SOM
16 TOPS(INT8)
2021
思元290
寒武纪首颗7nm训练芯片
MLU290-M5
512 TOPS(INT8)
MLU-X1000
2.05 POPS(INT8)
思元370
寒武纪首颗采用chiplet芯粒技术的AI芯片
MLU370
256TOPS(INT8)

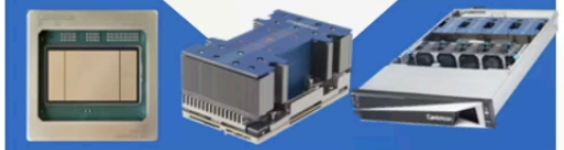
产品一览

终端智能处理器IP

面向嵌入式终端
提供人工智能芯片IP授权

- 手机等各类智能终端设备的SoC芯片
- 超强的片上推理能力

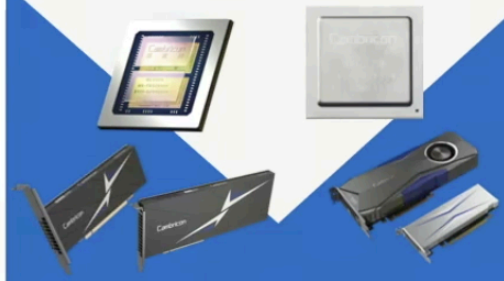
云端训练芯片



面向数据中心训练任务

- 面向云端人工智能训练任务的云端智能芯片
- 为云端训练任务提供强大的算力支撑

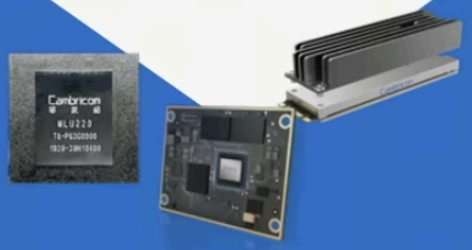
云端推理芯片



面向数据中心推理任务

- 面向云端人工智能推理任务的云端智能芯片
- 为云端推理任务提供强大的算力支撑

边缘端智能芯片



面向超紧凑/低功耗的边缘端
AI推理部署方案

- 智慧交通、智能制造、智能电网、智慧金融
- 低功耗，小尺寸边缘侧模组
- 数据边缘清洗，智能分析

产品一览

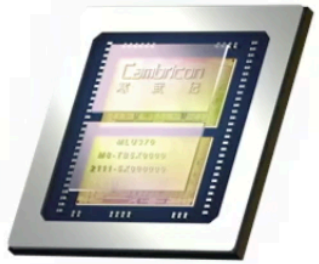

IP终端
边缘端
云端推理
云端训练
0.5-8TOPS
8-16TOPS
128-256TOPS
256-2048 TOPS
1H16
1H8
1M
思元220
MLU220-SOM
MLU220-M.2
思元370
MLU370-S4
MLU370-X4
思元290
ML290-M5
MLU-X1000
思元270
MLU270-S4
MLU270-F4
思元370
MLU370-X8

产品一览

场景	类型	产品	推出时间	收费方式
云端	推理	思元100	2018	买卡
	推理	思元270	2019	买卡
	训练	思元290	2020	买卡
	训练	思元370	2021	买卡
	训练	思元590	2023	买卡
	训练	玄思1000	2022	整机
边缘	推理	思元220	2019	买卡
终端	AI处理器IP	1A	2016	IP授权
		1H	2017	IP授权
		1M	2018	IP授权
系统软件	基础系统软件	Neuware	持续升级	部分免费

2. 训练产品

云端产品



思元370芯片

MLU370 加速卡

MLU290 OAM

MLUX1000 加速器

- 面向数据中心业务
- 涵盖训练和推理业务场景
- PCIE，OAM多种产品形态
- 训练整机产品方案

- 高性能AI运算能力
- 高速卡间互联网技术
- 提供浮点，定点多种计算精度，满足云端不同业务需求

训练芯片 思元 vs NVIDIA

	思元290	思元370-X8	思元590	NV A100
制造工艺	TSMC 7nm @ 102mm × 165mm	TSMC 7nm	TSMC 7nm	TSMC N7 @ 54.2B 828mm^2
算力	512 TOPS (INT8) 256 TOPS (INT16) 64 TOPS (CINT32) ?	256 TOPS (INT8) 128 TOPS (INT16) 96 TFLOPS (FP16) 96 TFLOPS (BF16) 24 TFLOPS (FP32)	期待中	624 TOPS @ INT8 312 TFLOPS @ BF16 312 TFLOPS @ FP16 156 TFLOPS @ TF32 支持 FP32、FP64 等类型
AI计算核心数	64个MLU Core	96个MLU Core	期待中	432 Tensor Core
多实例GPU	不支持	不支持	不支持	MIG(7个，每个10G)
架构	MLU02	MLU03	**MLU05**	Ampere
内存	32GB HBM2	48GB LPDDR5	期待中	80GB HBM
互联	聚合带宽600GB/s Bi-direction	聚合带宽200GB/s Bi-direction	期待中	NVLink 600GB/s
功耗	350W	250W	期待中	400W
接口	x16 PCIe Gen4	x16 PCIe Gen4	期待中	SXM5
发布(量产)	2020	2021	2023(?)	2020(2021)

MLU370-X8双芯思元370，双槽位250w，提供24TFLPOS(FP32)训练算力和256TOPS (INT8)推理算力。

同时MLU370-X8搭载MLU-Link多芯互联技术，每张加速卡可获得200GB/s的通讯吞吐性能，是PCIe 4.0带宽的3.1倍。

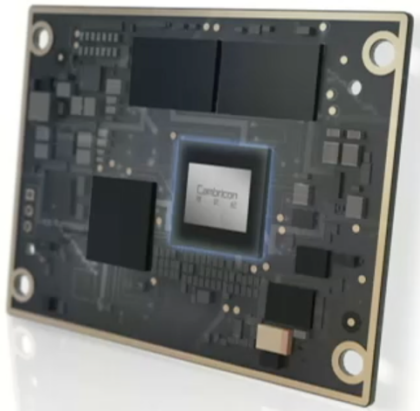
训练芯片 - 思考

- 产品手册相对国内其他厂商比较详细，软硬件架构可以进一步打开，让更多的开发者了解寒武纪的细节，参与到寒武纪在AI界的对接。
- 590！590！！590！！！

3. 推理产品

边缘产品

思元220芯片



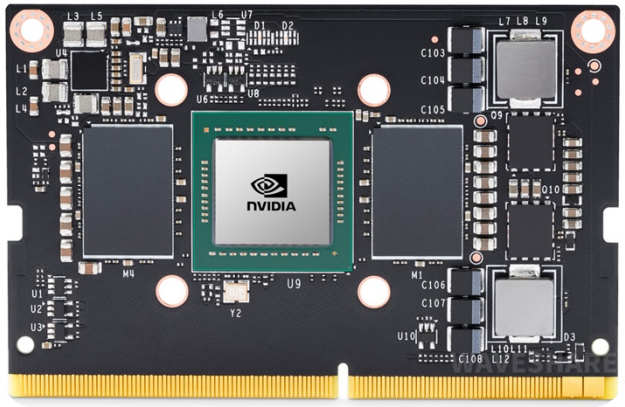
MLU220-SOM 模组

MLU220-M2 加速卡

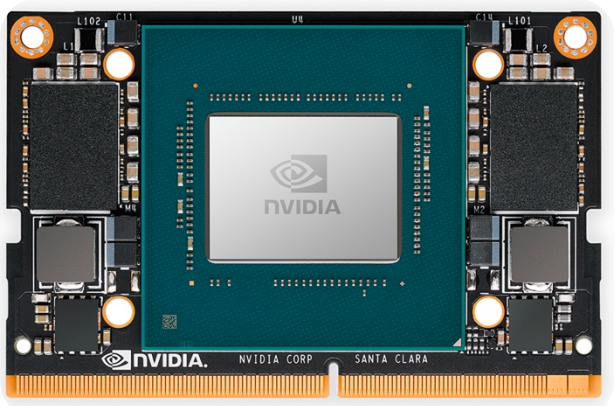
- 极致能效，灵活部署
- 模组，板卡，SOC多种产品形态
- 可提供最多十几个TOPS的算力
- 满足智能边缘计算各种使用场景
- 方便用户打造丰富多样的智能边缘软硬件解决方案

边缘芯片 思元 vs NVIDIA

	思元220	英伟达 Jetson	英伟达 Xavier
制造工艺	16nm@94.8mm2	87 mm x 50 mm	70 mm x 45 mm
产品形态	SOM/M2	TX2 NX/TX 4GB/TX2/TX2i	Xavier NX 16GB/Xavier NX
算力	INT8@16TOPS	FP16@1.33 TFLOPS not support INT8	INT8@21 TOPS
架构	MLU02	Pascal	Volta
AI计算核心数	/	256 CUDA core	48 Tensor Core 384 CUDA Core
内存	8GB 64bit LPDDR4x	8GB 128-bit LPDDR4	16 GB eMMC 5.1
互联	/	59.7 GB/s	59.7GB/s
功耗	15W	7.5W / 15W	10 W | 15 W | 20 W
接口	PCIe3.0 2×2 (RC)； SDIO3.0×2	1 x1 + 1 x4 OR 1 x1 + 1 x1 + 1 x2 (PCIe Gen2)	1 x1 (PCIe 第 3 代) + 1 x4 (PCIe 第 4 代)，共每秒 144 GT*
发布(量产)	2019	2017	2020



英伟达 Jetson TX2



英伟达 Jetson Xavier

边缘芯片 – 思考

- 1）边缘的技术护城河比较短，2）竞争激烈，国内AI边缘芯片和对应的产品特别多（沐曦、壁仞、燧原、天数智芯），3）利润较低，NVIDIA Xavier 和 Jetson 系列价格感人。因此，寒武纪的聚焦重点是否在于训练而不是推理？

4. 端侧产品

终端处理器 IP

寒武纪
终端智能处理器IP

寒武纪1H16处理器

更高性能、更完备的终端智能处理器IP

寒武纪1H8处理器

面向计算机视觉领域的专用处理器IP

寒武纪1M处理器

面向智能驾驶的处理器IP

终端智能处理器 IP

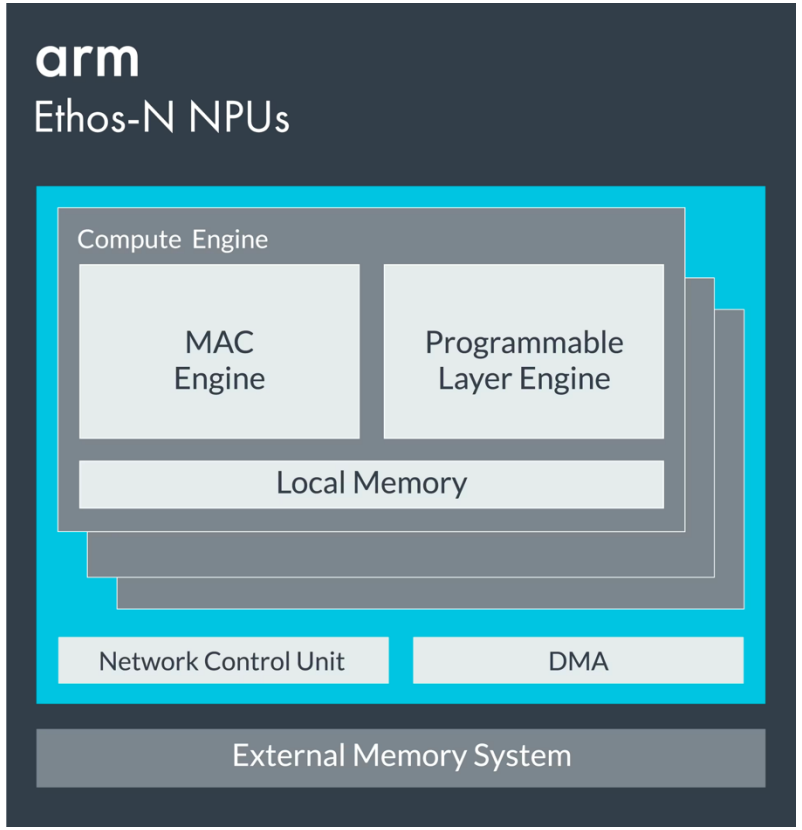
	寒武纪1A	寒武纪1H	寒武纪1M
峰值算力	FP16@0.5TOPS	INT8@1TOPS FP16@0.5TOPS	INT8@8TOPS INT4@16TOPS
制造工艺	IP Only	IP Only	IP Only
功耗	2TOPS/W @7nm	4TOPS/W @7nm	5TOPS/W @7nm
场景	端侧推理	端侧推理	端侧推理
产品	华为麒麟970 1种型号	华为麒麟980 3种子型号	华为麒麟990 3种子型号
软件栈	None	None	None
发布时间	2016	2017	2018

麒麟990采用自研华为达芬奇架构NPU，采用2个超大核+2个大核+4个小核的三档能效架构

终端智能处理器 IP - ARM

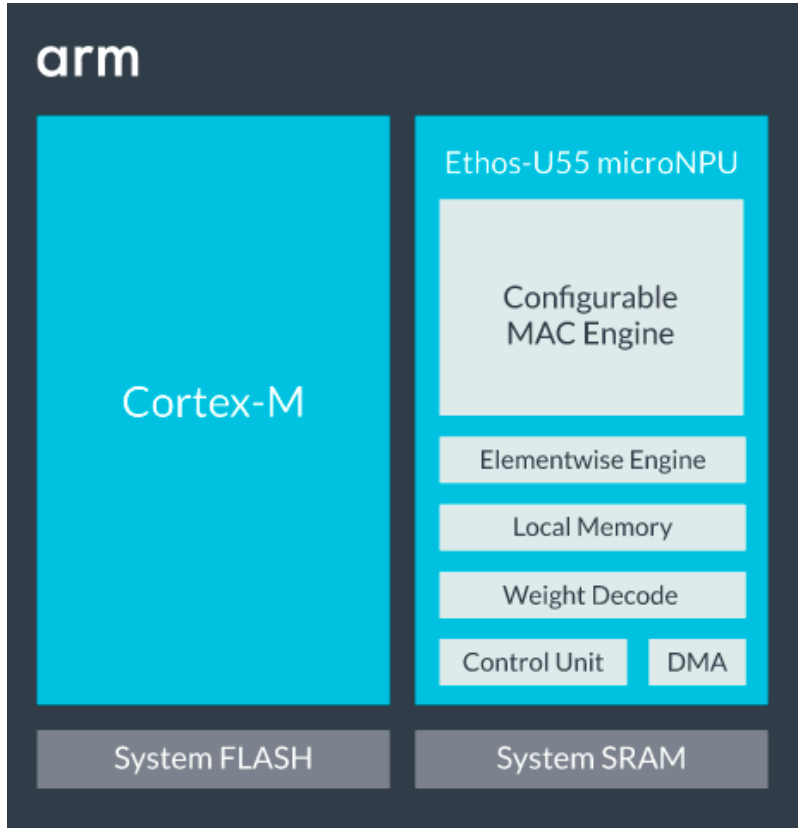
ARM Ethos-N57

Ethos-N 可以作为独立IP集成在SoC中



Ethos-U55

与Cortex-M类NPU耦合的首款microNPU



终端智能处理器 IP – 思考

- **销售策略：**同期 ARM、Qualcomm、MediaTek 等终端IP和芯片厂商，都推出自己的 NPU 和对应的 AI 软件栈，终端 IP 捆绑销售比较常见，失去了华为麒麟系列，技术和商业上应该如何突围？
- **技术竞争力：**终端 NPU 目前技术上集中在推理，但是终端的推理引擎（MNN、NCNN、XXX Lite）目前为止，因为软件的易用性和开发效率问题，严重依赖于 CPU 和 GPU 能力，软件栈如何构建核心竞争优势？

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