Cambricon
寒武纪科技

寒武纪：人工智能芯片的创新力量

寒武纪科技

钱诚

副总裁

Cambricon
寒 武 纪 科 技

我们从这里来

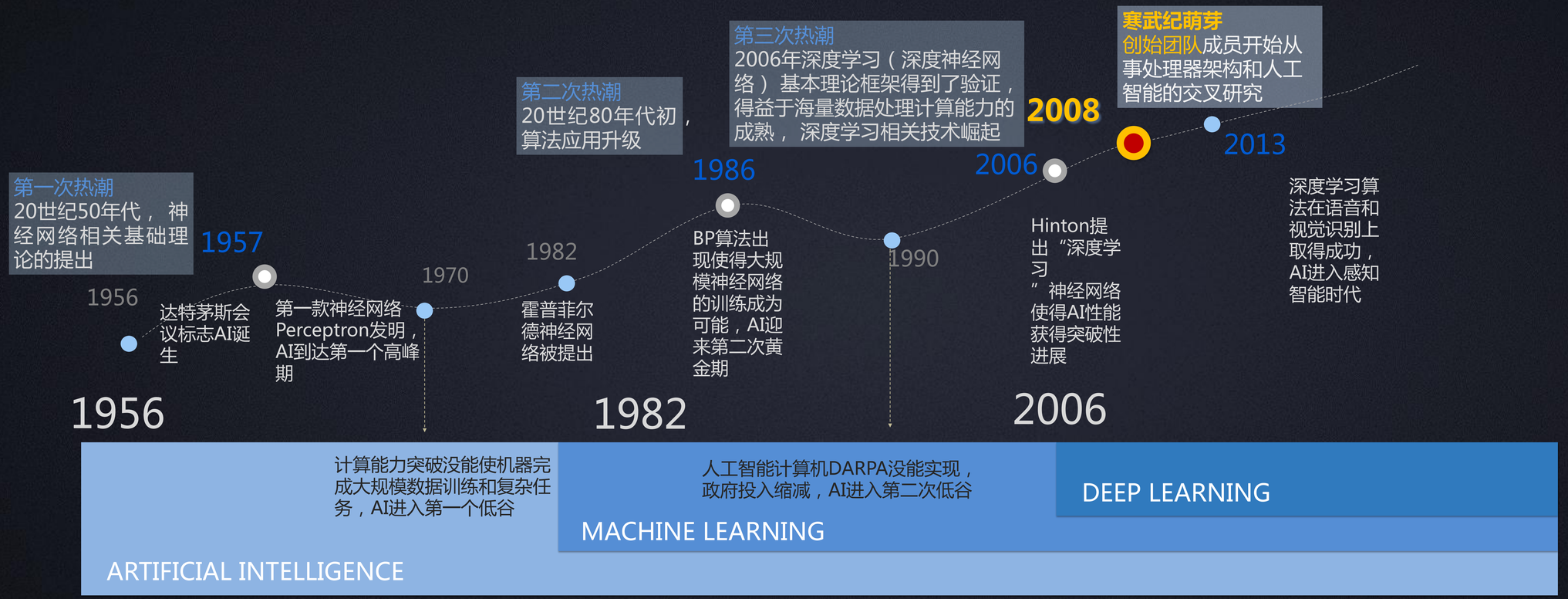
在5~6亿年前的寒武纪，大量较高等物种出现，物种多样性大幅提升。这个现象被称为寒武纪物种大爆发。

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。

Cambricon = Cambrian + Silicon

Cambricon
寒 武 纪 科 技

深度学习的崛起与AI的第三次热潮

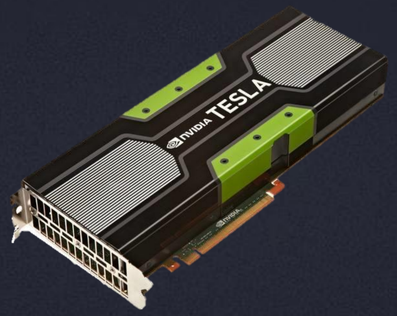


从1950年至今，人工智能历经三次发展热潮，从诞生到机器学习再到深度学习

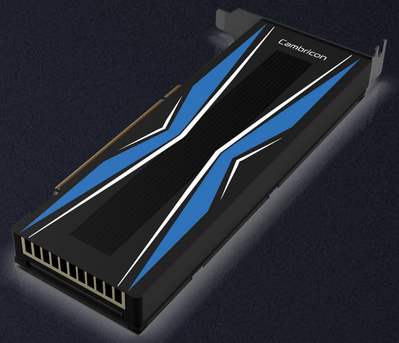
Cambricon
寒 武 纪 科 技

人工智能为什么需要专门的处理器？

CPU-通用计算



GPU-图形计算



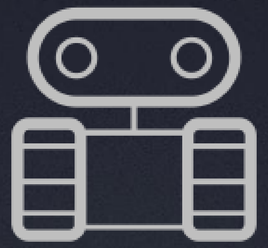
MLU-智能计算

Cambricon
寒 武 纪 科 技

智能创新的物质载体：智能芯片

智能终端，VR、AR设备

虚拟场景理解
人机交互
…



机器人、无人驾驶

感知智能
增强学习
…

互联网、数据中心

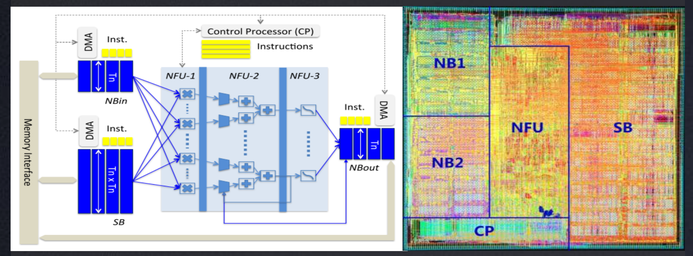
认知智能
决策判断
…

寒武纪是AI专用芯片的先行者

Cambricon
寒 武 纪 科 技

寒武纪：智能芯片的先行者

寒武纪的学术工作



DianNao, 2014

PRIME

Fused CNN

Eyeriss

EIE

Stripes

RedEye

Cnvlutin

Bit-Pragmatic

Pipelayer

FlexFlow

ScaleDeep

SCNN

TPU

DianNao

DaDianNao

ShiDianNao

PuDianNao

Cambricon

Cambricon-X

2014

2015

2016

2017

*数据from ISCA，HPCA，ASPLOS，MICRO，2010~2017

DianNao系列学术研究由来自法国Inria等机构的国际学术合作者共同完成

Cambricon
寒 武 纪 科 技

智能芯片如何做到通用和好用？

分析和抽取应用负载特征

设计灵活的指令集

设计可扩展性强、高效的架构

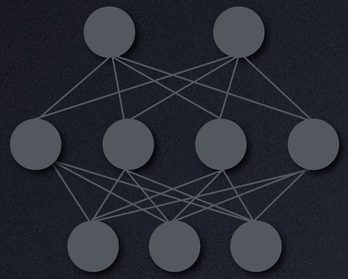
提供灵活的运算器方案

支持主流编程框架

在大规模商用中得到反馈和修正

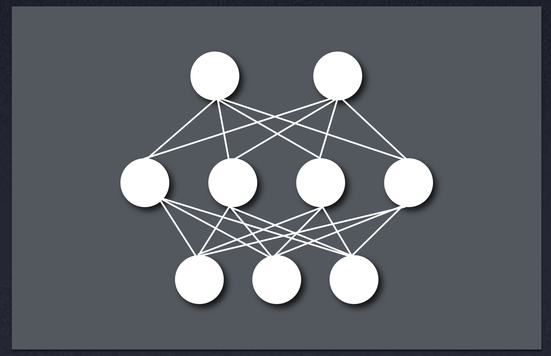
Cambricon
寒 武 纪 科 技

传统思路

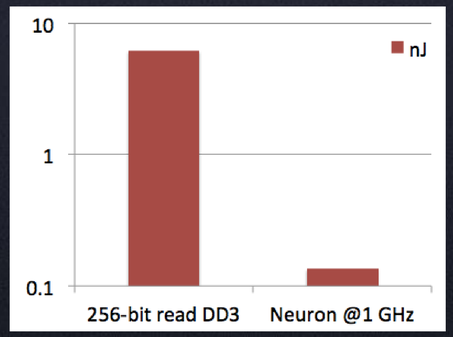


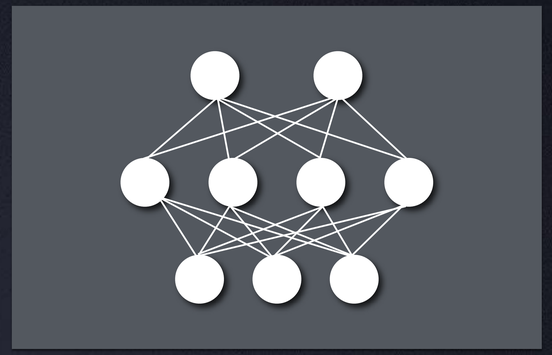
算法

把数据从内存搬运到硬件运算单元，甚至比运算本身更耗能量



硬件运算单元和
算法神经元一一对应



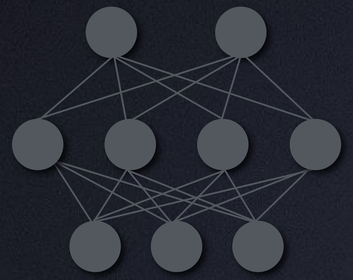


片外内存

硬件运算单元数量稍微一多，
访存带宽就供应不上数据

Cambricon
寒 武 纪 科 技

寒武纪思路



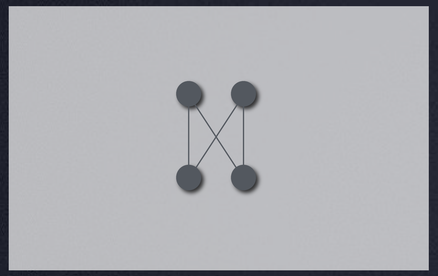
算法

对硬件运算单元分时复用

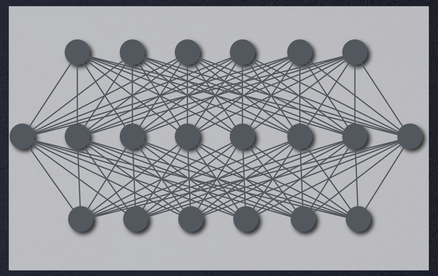
- 小尺度但支持大规模神经网络
- 速度：把访存带宽用起来，尽可能提高性能
- 能耗：通过优化片上存储层次**尽量减少访存次数**

神经元虚拟化

有限规模的硬件 vs 任意规模的算法



单芯片只能集成
数千硬件神经元



通过对硬件神经元时分复用，可虚拟化出**千亿**级别超大规模神经网络

Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam, "DianNao: A Small-Footprint High-Throughput Accelerator for Ubiquitous Machine-Learning," In Proceedings of 19th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'14), 2014. (Best Paper Award)

通用智能指令集

结构固定的硬件 vs 千变万化的算法

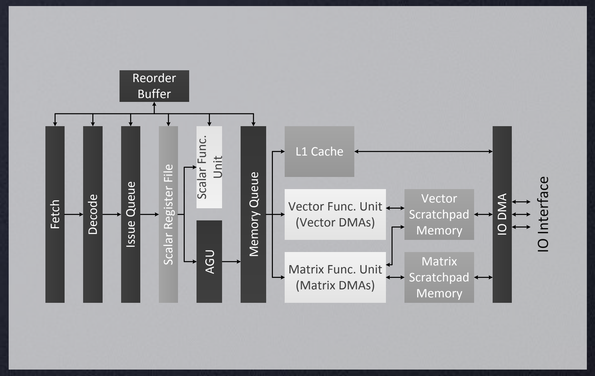
抽取和提炼
各种智能算法共性基本运算

+

设计通用智能指令集
来高效处理千变万化的智能算法

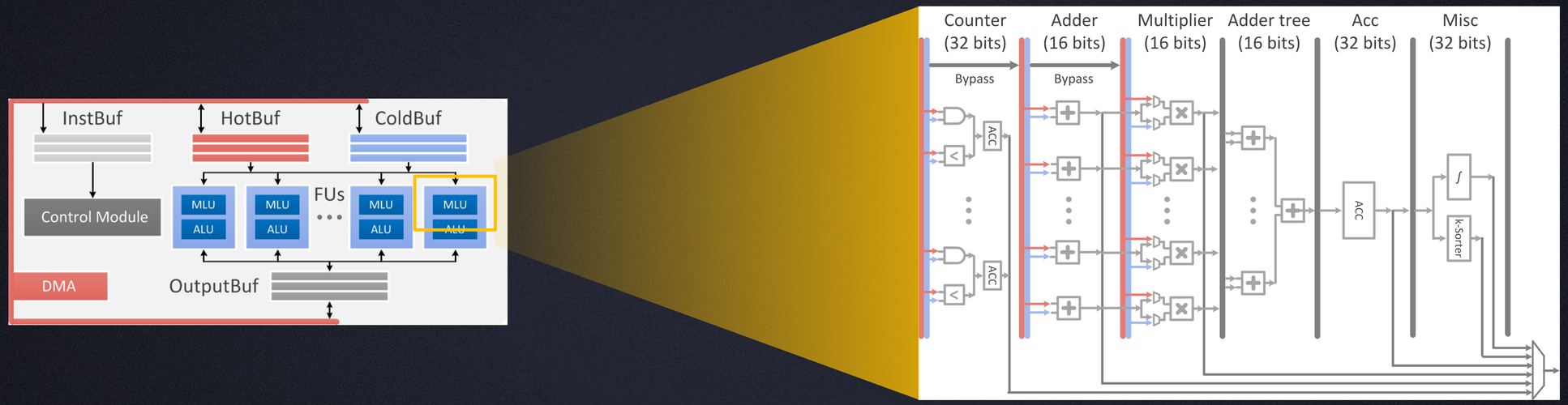
Table I. An overview to Cambricon instructions.

Instruction Type		Examples	Operands
Control		jump, conditional branch	register (scalar value), immediate
Data Transfer	Matrix	matrix load/store/move	register (matrix address/size, scalar value), immediate
	Vector	vector load/store/move	register (vector address/size, scalar value), immediate
	Scalar	scalar load/store/move	register (scalar value), immediate
Computational	Matrix	matrix multiply vector, vector multiply matrix, matrix multiply scalar, outer product, matrix add matrix, matrix subtract matrix	register (matrix/vector address/size, scalar value)
	Vector	vector elementary arithmetics (add, subtract, multiply, divide), vector transcendental functions (exponential, logarithmic), dot product, random vector generator, maximum/minimum of a vector	register (vector address/size, scalar value)
	Scalar	scalar elementary arithmetics, scalar transcendental functions	register (scalar value), immediate
Logical	Vector	vector compare (greater than, equal), vector logical operations (and, or, inverter), vector greater than merge	register (vector address/size, scalar)
	Scalar	scalar compare, scalar logical operations	register (scalar), immediate



Shaoli LiZidong Du, Jinhua Tao, Dong Han, Tao Luo, Yuan Xie, Yunji Chen, and Tianshi Chen, "Cambricon: An Instruction Set Architecture u, for Neural Networks," In Proceedings of the 43rd ACM/IEEE International Symposium on Computer Architecture (ISCA'16), 2016. (Highest Score in Peer Review)

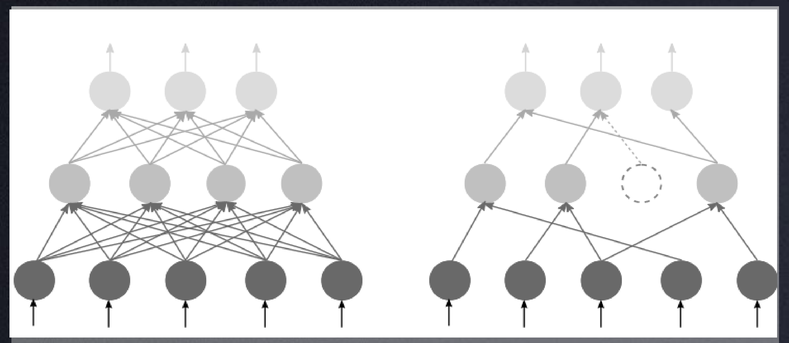
硬件架构举例



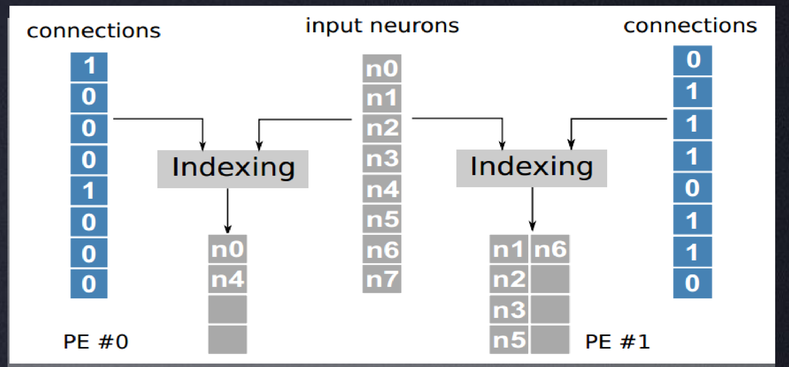
Daofu Liu, Tianshi Chen, Shaoli Liu, Jinhong Zhou, Shengyuan Zhou, Olivier Temam, Xiaobing Feng, Xuehai Zhou, and Yunji Chen, "PuDianNao: A Polyvalent Machine Learning Accelerator," In Proceedings of 20th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'15), 2015.

稀疏化模型处理

能耗受限的硬件 vs 计算量大的算法



稀疏化包括权值稀疏化和神经元稀疏化，只要二者之一为0，相应乘加运算即可被跳过以达到加速计算的目的。



权值只存储非0元素及其index信息，可以节省存储容量和访存带宽，index还可用作判断是否跳过计算的辅助信息。

Shijin Zhang, Zidong Du, Lei Zhang, Huiying Lan, Shaoli Liu, Ling Li, Qi Guo, Tianshi Chen, and Yunji Chen, "Cambricon-X: An Accelerator for Sparse Neural Networks," In Proceedings of 49th IEEE/ACM International Symposium on Microarchitecture (MICRO'16), 2016.

Cambricon
寒 武 纪 科 技

端云结合，端云一体

/ 人工智能 / & / 大数据 / & / 云计算 /

智能终端处理器IP

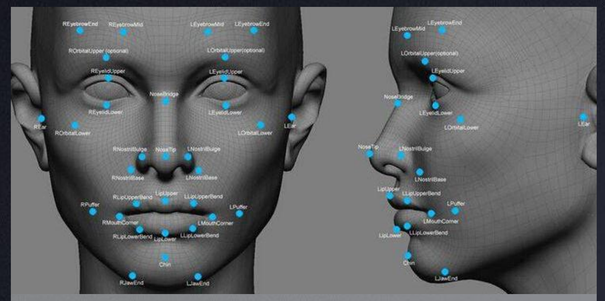
手机、安防产品的SOC芯片
相机传感器芯片组
超强的片上推理能力

智能云服务器芯片

面向深度学习/机器学习的专用处理器
云端推理+训练

Cambricon
寒 武 纪 科 技

应用领域



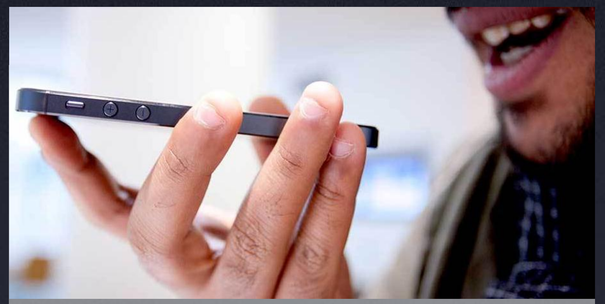
机器视觉

典型技术

人脸/行人/车辆的检测、追踪、识别和属性分析，文字/物体的检测和识别

典型应用

安防监控，人脸身份认证，智能交通，机器人视觉（如无人机），图像搜索引擎，图像/视频的理解与美化



语音识别

典型技术

语音识别、声纹识别、多麦克风阵列系统

典型应用

语音输入，语音控制，智能助手，机器翻译、机器人听觉

自然语言

典型技术

词句嵌入、语义建模

典型应用

聊天机器人，智能助手，智能客服，视频理解，机器翻译

Cambricon
寒 武 纪 科 技

智能终端处理器IP

寒武纪深度学习处理器IP

授权华为海思使用寒武纪1A处理器
为华为Kirin970手机芯片和Mate10手机插上智慧之翼

Kirin + Cambricon 寒 武 纪 科 技 **=全球首款AI手机芯片**

寒武纪1H16处理器

更高性能、更完备的深度学习处理器IP（2017年Q1上市）

寒武纪1H8处理器

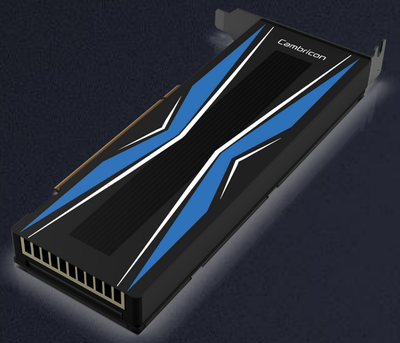
面向计算机视觉领域的专用处理器IP（2017年Q3上市）

寒武纪1M处理器

面向智能驾驶的处理器IP
（ 2018年Q2上市）

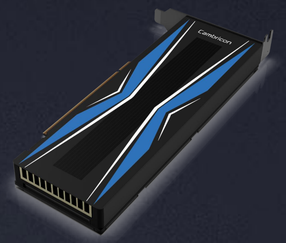
Cambricon
寒 武 纪 科 技

智能云服务器芯片



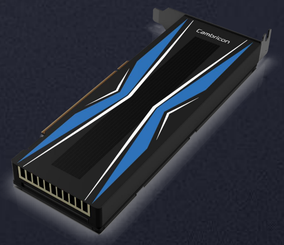
寒武纪智能处理卡MLU100

2018年5月3日在中国上海正式发布，标志着寒武纪已成为中国第一家（也是世界上少数几家）同时拥有终端和云端智能处理器产品的商业公司



寒武纪MLU100智能处理卡

寒武纪推出的第一款通用智能处理卡，侧重推理（2018年Q2上市）



寒武纪MLU200智能处理卡

支持训练和推理，侧重训练
（预计2019年上市）

Cambricon
寒 武 纪 科 技

寒武纪MLU100详细参数

核心架构	**Cambricon MLUv01**
核心频率	1Ghz
半精度浮点运算速度（FP16）	16 TFLOPS（关闭稀疏模式时峰值） 64 TFLOPS（打开稀疏模式时峰值）
定点运算速度 (INT8)	32 TOPS（关闭稀疏模式时峰值） 128 TOPS（打开稀疏模式时峰值）
内存容量	8GB/16GB/32GB
内存位宽	256-bit
内存接口	102.4 GB/s
系统接口	PCI Express 3.0 x16
ECC保护	是

Cambricon
寒 武 纪 科 技

软件开发套件：Cambricon NeuWare

开发

软件开发工具包

TensorFlow
Caffe
MXNet
CN-Lib（高性能函数库）
CN-CC（编译器）
CN-Gen（模型生成工具）

调试

功能调试工具包

CN-GDB（调试工具）
CN-Dmp（转储工具）
CN-Cmp（比较工具）

调优

性能调优工具包

CN-Prof（性能剖析）
CN-Adv（优化建议）
CN-Est（性能评估）

Cambricon
寒 武 纪 科 技

寒武纪产品路线图

Cambricon
寒 武 纪 科 技

寒武纪的三年目标

3年10亿台

集成寒武纪处理器的智能终端

3年30%

中国高性能智能芯片市场

边缘计算
智能终端

高性能计算
云数据中心

让机器更好地理解和帮助人类