

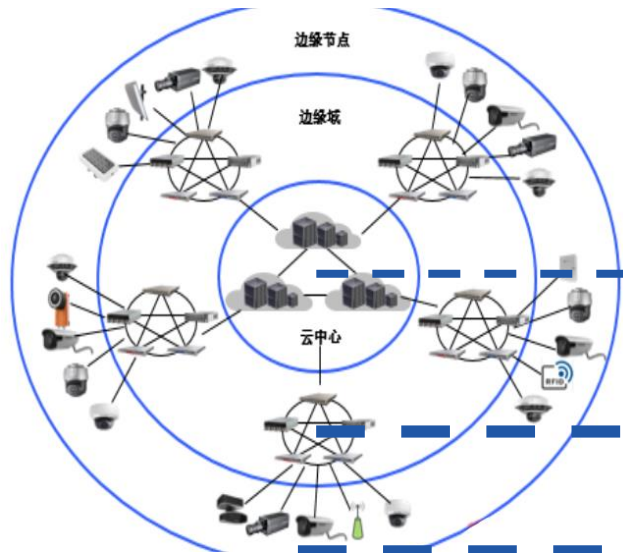
寒武纪MLU220解决方案



速度 · 创新 · 热情 · 客户 · 质量

Cambricon
寒 武 纪

寒武纪端云一体解决方案



• 云中心

寒武纪云端加速卡

- 大数据集中融合;
- 云数据深度分析;

• 边缘侧

寒武纪低功耗小尺寸边缘侧模组

- 数据就近清洗, 数据按需汇聚和存储;
- 实时快速智能分析

• 终端

寒武纪IP SOC芯片

- 前端数据智能采集;
- 边缘感知

云中心

- MLU290 : PCIe X16 云端训练卡, 512TOPS , 非同一般的计算能力造就出众的训练性能, 可让现代化的数据中心更快速地处理涉及人工智能和高性能计算的工作负载, 提供企业级的人工智能基础设施;
- MLU270 : PCIe X16 云端推理卡 : 128TOPS, 世界领先的推理加速器, 提供革命性的多精度推理性能, 以加速现代人工智能的各种应用

边缘侧

- MLU220: 高性能低功耗加速计算模组, 16T/10W, 实现在板载设备上快速且精准的复杂推理, 将AI应用于边缘设备, 发挥设备在网络受限环境下的巨大潜力;

终端IP:

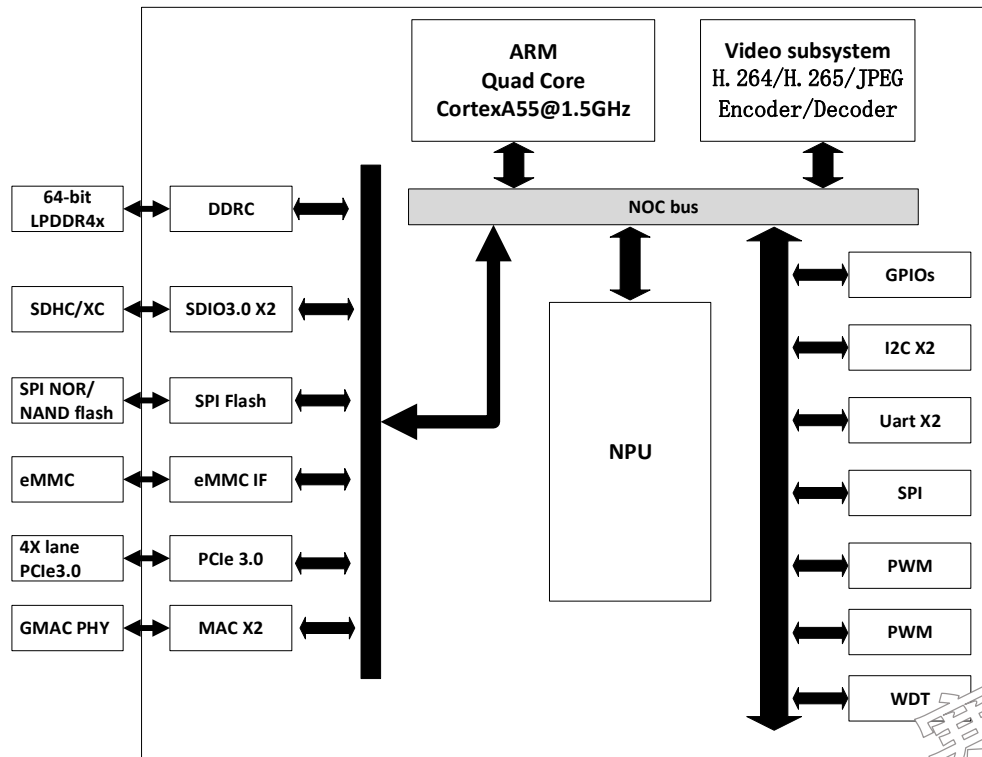
- 1M、1V 神经计算IP: 高性能低功耗的IP产品, 针对不同前端应用定制化的IP产品, 将智能延伸到前端移动设备;

MLU220 Architecture TopView

PART ONE

寒武纪授权

MLU220 System Architecture



寒武纪授权

MLU220 SKU

PART Two

寒武纪授权

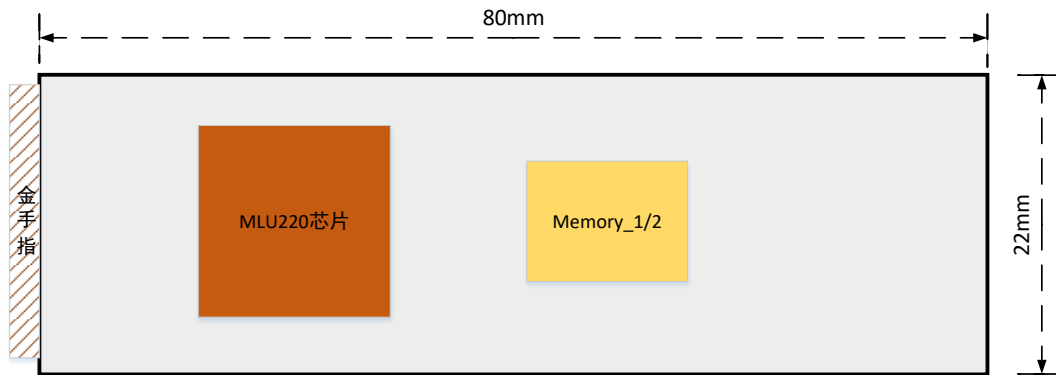
SKU—M.2卡



参数	规格
型号	MLU220-M.2
内存	4GB , LPDDR4x 64bit, 3200Mhz
AI算力	8TOPS (INT8)
编解码能力	- 支持H.264 , HEVC (H.265) , VP8, VP9 - 解码性能: 可以支持16X 1080P @30fps - 编码性能: 可以支持8X 1080P@30fps
图片解码	- JPEG, 最大图片分辨率8192 x 8192 - 解码性能: 410 fps@FHD, 编码性能: 400 fps@FHD
接口规格	M.2 2280, B+M key (PCIe3.0 X2)
功耗	8.25W (3.3V, 2.5A)
结构尺寸	长80mm, 宽22mm, 高7.3mm (无散热) 21.3mm (带散热)
散热	被动散热
表面温度	-20~80 度

寒武纪授权

SKU—M.2卡



(内存颗粒位置为板卡正反面分别有一颗内存颗粒)

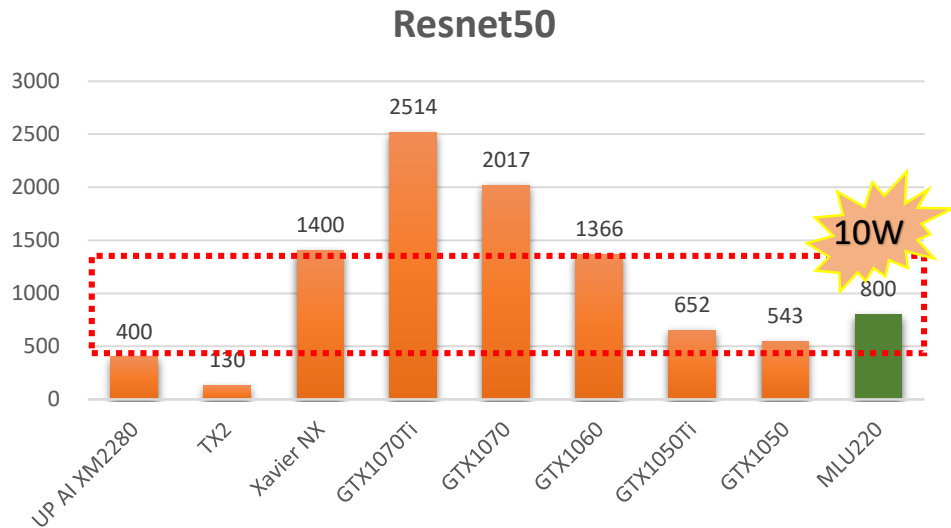
- 1.体积小，性能强：只有半张信用卡大小，内置的NPU算力最高可达8.0TOPs；
- 2.功耗更低：系统最大功耗只有8.25W；
- 3.扩展性强：在同一平台上，支持多个设备叠加使用，以扩展主机性能。

竞品分析—边缘端解决方案

边缘模块分析	名称	MLU220	Atlas200	Jetson TX2	UP AI Core XM2280
	Vendor	寒武纪	华为	NV	Intel
	INT8 (TOPS)	16	16	4.4	2
	功耗(W)	10W	13.8	7.5/15	5
	制程	16nm	12nm	16nm	16nm
	内存	LPDDR4X 8GB	LPDDR4X	LPDDR4	LPDDR4 4Gb/chip 1GB
	系统接口	PCIe3.0 2X2	PCIe3.0 1X4	PCIe 2.0 X1, X2, X4	M.2
	网络接口	2X GE	1X GE	1X GE	-
	SATA	-	-	SATA2.0	-
	USB	-	X1 2.0/3.0	USB3.0	-
	eMMC/SDIO	eMMC5.1, SDIO3.0 SDIO X2:104MB/s eMMC:400MB/s	COMBO SDR52 50MB/s	eMMC5.1, SDIO4.0	-
	其他接口	SPI, PWM, I2C, UART, GPIO	Misc I/O	Misc I/O	-
	视频编解码	16X Video decoder 1080P 8X Video encoder 1080P	16X Video decoder 1080P 1X Video encoder 1080P	8X Video decoder 1080P 8X Video encoder 1080P	4X Video 1080P decoder

竞品分析—更高的算力功耗比

Product	TOPS	Benchmark	
UP AI XM2280	8	400	INT8
TX2	1.3TFLOPS	130	FP16
Xavier NX	21	1400	INT8
GTX1070Ti	32	2514	INT8
GTX1070	26	2017	INT8
GTX1060	18	1366	INT8
GTX1050Ti	8	652	INT8
GTX1050	7	543	INT8
MLU220	16	800	INT8

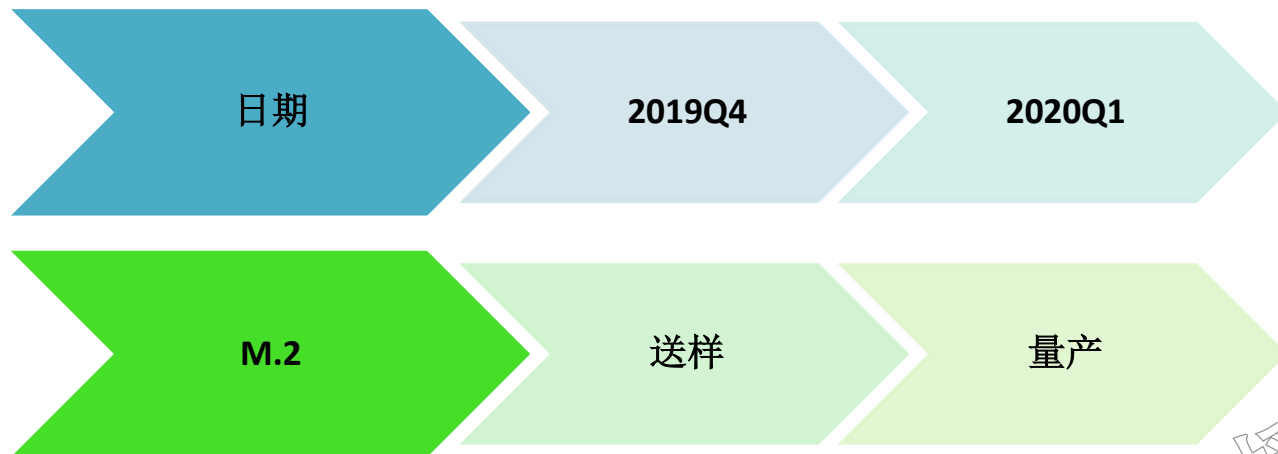


MLU220 Comparison & Plan

PART Three

寒武纪授权

RoadMap

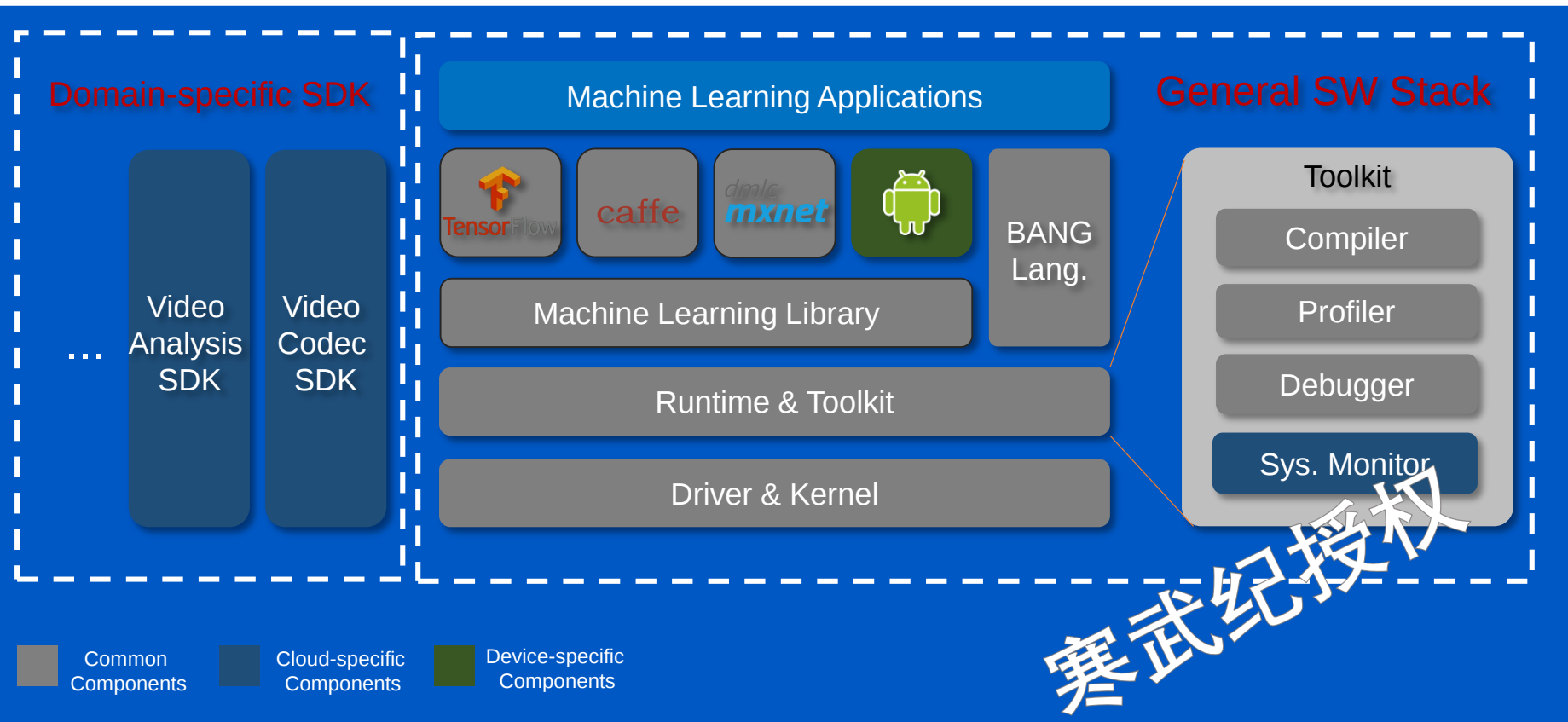


寒武纪授权

MLU220 SDK & Application

PART Four

寒武纪授权



Framework Support


TensorFlow

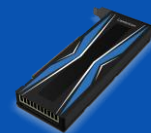
caffe

dmlc
mxnet

 ONNX



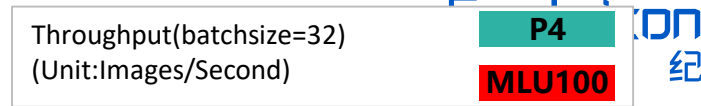
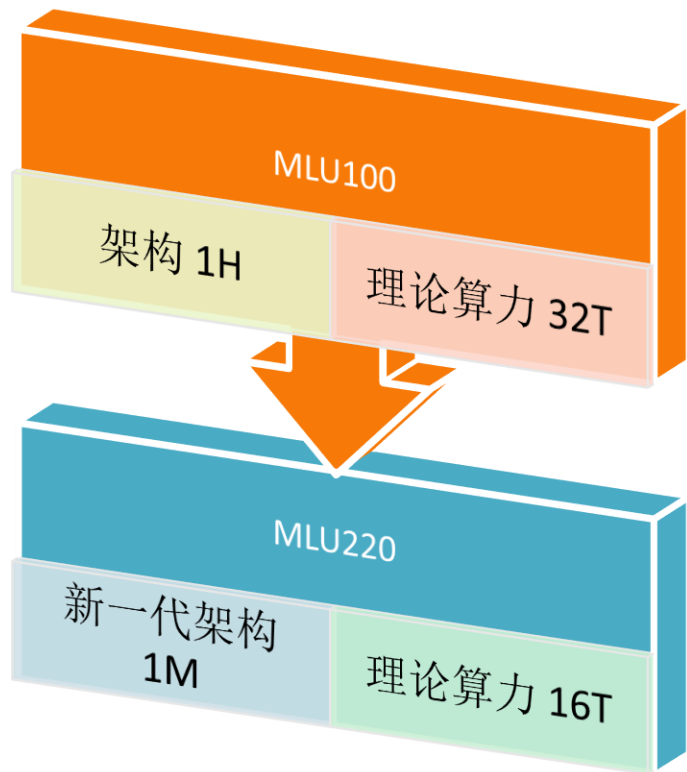
Inference



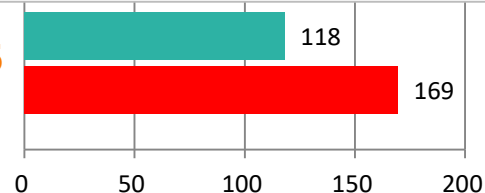
MLU

寒武纪授权

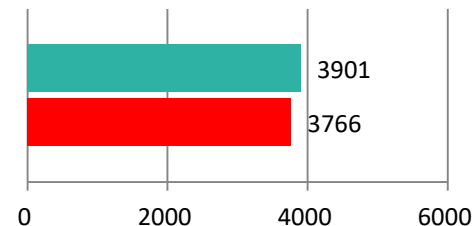
Capability Evaluation



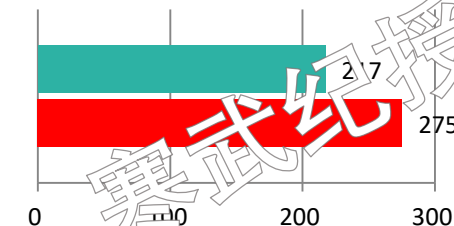
YOLOV3 -416



Resnet18



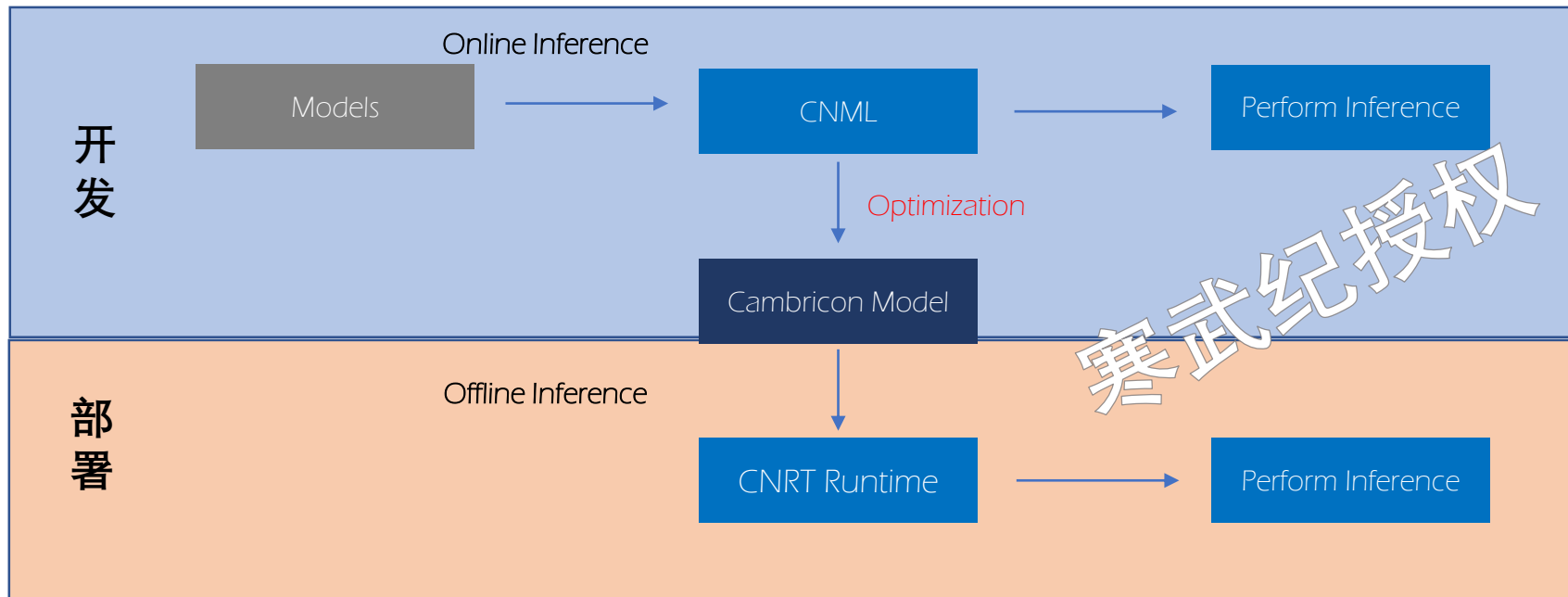
SSD-VGG16



MLU220 Software Development flow

开发调试 --- 在线模式, JIT编译生成和在线更新网络模型, 通用性更高;

业务部署 -- 离线模式, 离线编译生成和更新网络模型, 适合于网络模型相对固定场景, 性能更优;



Typical Application Scenario

- 可大量应用于对数据处理有低时延，高安全需求的边缘应用场景。
- 可实现数据的就近清洗，本地处理；



□通过在生活娱乐中引入人工智能技术，使生活娱乐与人工智能高度融合，对生活方式产生前所未有的深刻影响。

- **内容审核：**利用深度学习和图像识别的综合性技术，可以快速精准识别图片和视频中的涉黄、涉暴涉恐、政治敏感、微商广告及令用户不适的内容和事件场景等。
- **智能点播：**通过运用智能语音识别及交互技术打造的新型语音交互类业务，将传统遥控器的按键操作、拼音搜索全部转换为语音操控，进行云端IPTV视频源的智能点播。
- **广告推荐：**利用大数据技术使用户画像变成更简单，通过对特定人群的分析来准确知晓用户的购买需求，从而实现广告的精准投放。

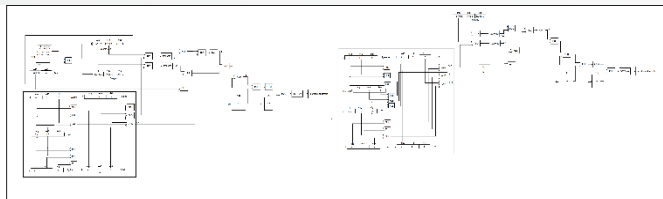


Latency
 $\frac{1}{4} \times$

8X
Throughput

语音合成

基于业界领先的深度神经网络技术，提供高度拟人、流畅自然的语音合成服务，让您的应用、设备开口说话，更具个性



提供多种发音人

提供基础音库和精品音库共9种发音人供您选择，适用于泛阅读、订单播报、智能硬件等应用场景，即将推出更多特色发音人



支持多音字标注

中文多音字可通过标注拼音、音调自行定义发音，例如“轻舟已过万重(chong2)山”、“脑筋急转(zhuan3)弯”



语速、音调可调节

支持多种参数配置，可根据场景需求对发音人的语速、音调、音量进行灵活设置，满足个性化需求



多种调用方式，满足多场景需求

提供REST API接口、离在线融合SDK、纯离线SDK，满足不同网络环境下的语音合成需求，提供流畅自然的合成体验

应用案例- 智慧教育

Cambricon
寒 武 纪



01

AI Teacher

Learning data is applied to the classroom through live and AI technologies to make learning more effective.



02

Emotion Recognition

Real-time identification and classification of students' expressions can help track students' focuses and provide teachers with more insights.

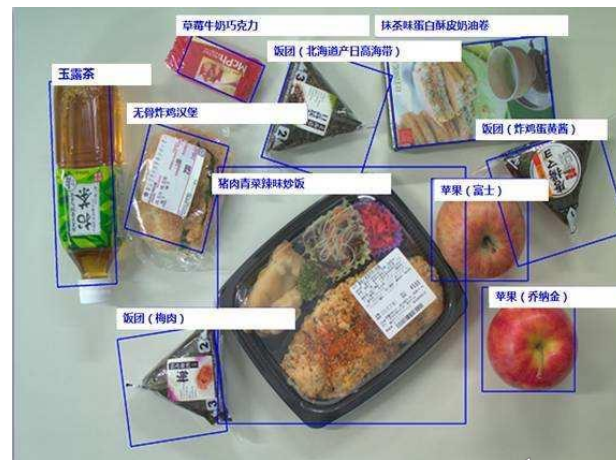
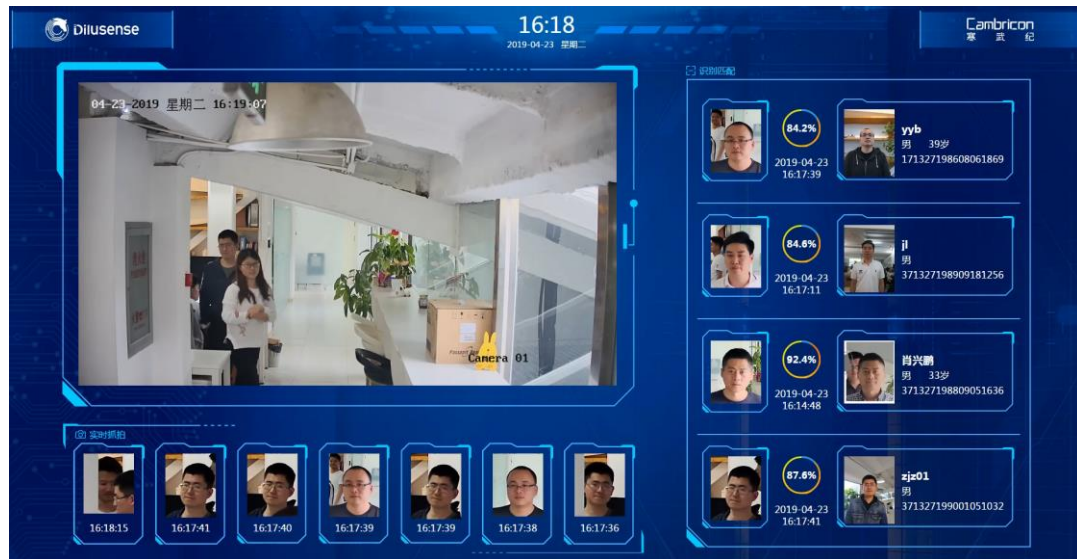
应用案例- 智慧农牧

Cambricon
寒 武 纪

寒武纪授权

应用案例- 智慧商超

Cambricon
寒 武 纪



寒武纪授权

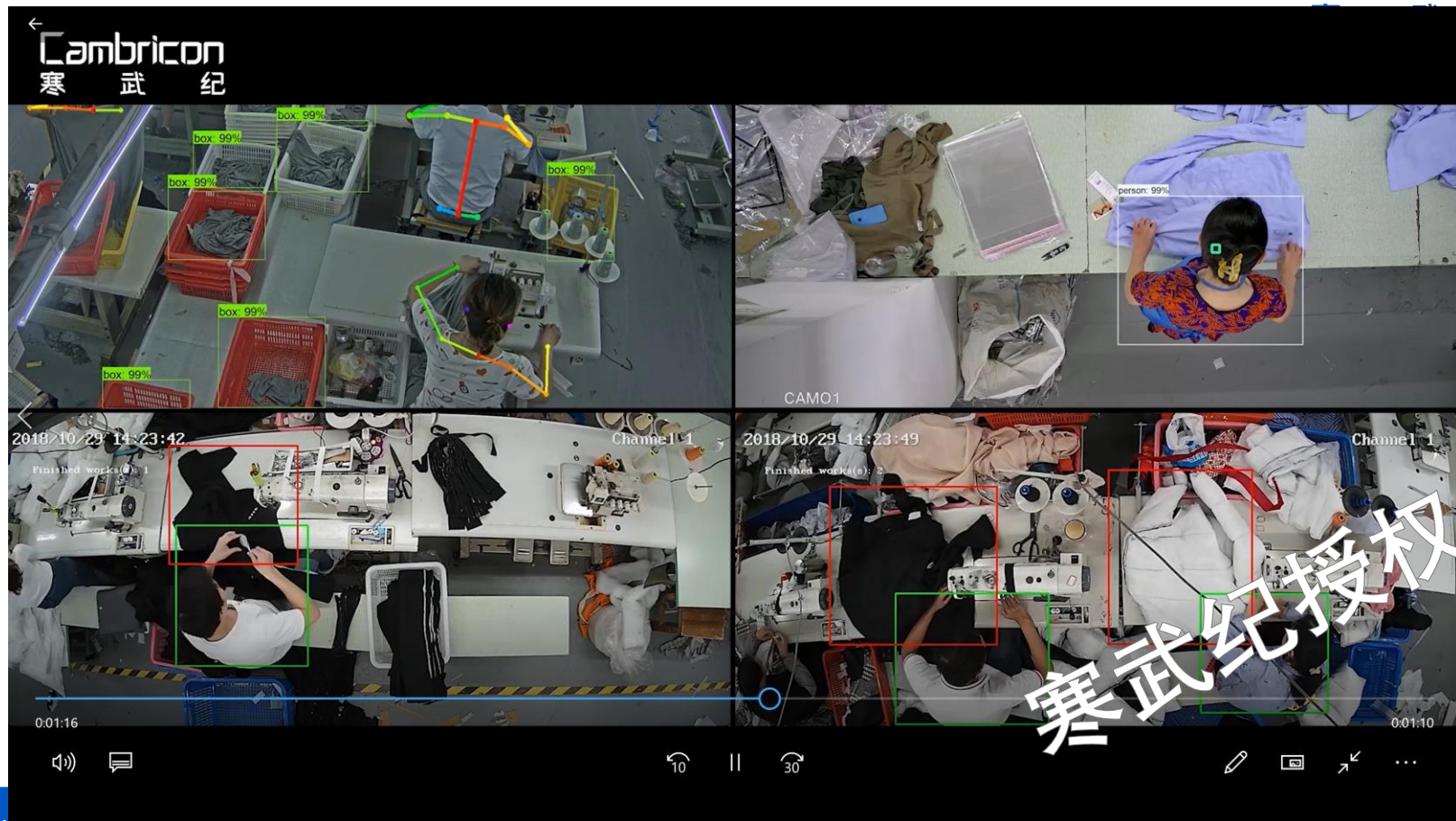
应用案例-智慧城市

Cambricon
寒武纪



应用案例- 智慧工厂

Cambricon
寒武纪



THANKS

让机器更好地理解和服务人类

Cambricon
寒 武 纪