

# 算力扩散，边缘场景和投资价值

---

## 通信行业专题报告

分析师：李宏涛 S0910523030003

- ◆ 算力向推理和边缘扩散；扩散呈现三个趋势：1、低延时低功耗需求增强、2、AI芯片向端侧推理演进、3、缩减数据处理成本；边缘算力三种体现：模组、终端、边缘计算中心。
- ◆ 产业链重构带来价值：芯片侧：RISC-V更适合边缘架构，高通在物联网侧一骑绝尘；系统侧：“平台+操作系统”边缘算力中枢；设备侧：边缘计算载体；设施侧：支撑边缘计算高效响应运行。
- ◆ 车和垂直行业弹性爆发：车辆边缘算力的iot模组产业链拆解；扫地机器人：“大脑+小脑”扫地机器人三大处理器；微模块idc边缘算力拆解
- ◆ 投资逻辑：投资价值一：边缘预处理最先落地；投资价值二：大模型百花齐放参数不一、大模型转小模型是必须之路、模型嵌入所需硬件支撑—算力/带宽/内存
- ◆ 建议关注标的：美格智能、移远通信、映翰通、东土科技、移为通信、三旺通信、广和通
- ◆ 风险提示：边缘模型算法开发进展不及预期；边缘应用场景需求不及预期；芯片成本过高影响落地进度。

- 01 算力向推理和边缘扩散
- 02 产业链重构带来价值
- 03 车和垂直行业弹性爆发
- 04 投资逻辑和重点标的
- 05 投资建议与风险提示

# 趋势一：低延时低功耗

- ◆ **边缘算力低延迟、占用带宽资源少。**云端受限于延时性和安全性，不能满足部分对数据安全性和系统及时性要求较高的用户需求。边缘计算是5G网络架构中的核心环节，能够解决5G网络对于低时延、高带宽、海量物联的部分要求，大幅提升生产效率。
- ◆ **下游应用场景爆发，边缘AI芯片需求旺盛。**ABI Research预计，边缘AI芯片市场规模将从2019年的26亿美元增长到2024年的76亿美元，边缘AI芯片市场将超过云AI芯片市场。根据前瞻产业研究院的预测数据显示，中国人工智能芯片市场规模将保持40%-50%的增长速度，到2024年市场规模将达到785亿元。我国人工智能芯片行业的下游应用场景主要聚集在云计算与数据中心、边缘计算、消费类电子、智能制造、智能驾驶、智慧金融、智能教育等领域。

图表1 云端AI弊端

主要问题	问题描述
功耗过高	与云端进行大量的数据传输将产生极大的功耗，限制了终端设备的应用。以比特币“挖矿”为例，矿工们一般都是采用英伟达的GPU，支付高昂的电费。
实时性不强	本地数据通过网络传输到云端，云端再将计算结果返回至终端，这一过程存在数秒乃至数十秒的延迟。自动驾驶、工业现场领域，毫秒级的延迟实时性不强。
带宽不足	传感器的大范围普及和低功耗广域网等连接技术的飞速发展，设备数呈指数型飞速增长，无法满足互联的带宽
安全问题	网络传输过程中存在数据被劫持的风险，隐私与安全性缺乏。

图表2 2019-2024 年中国边缘智能芯片市场规模情况



# 趋势二：AI芯片向端侧推理演进

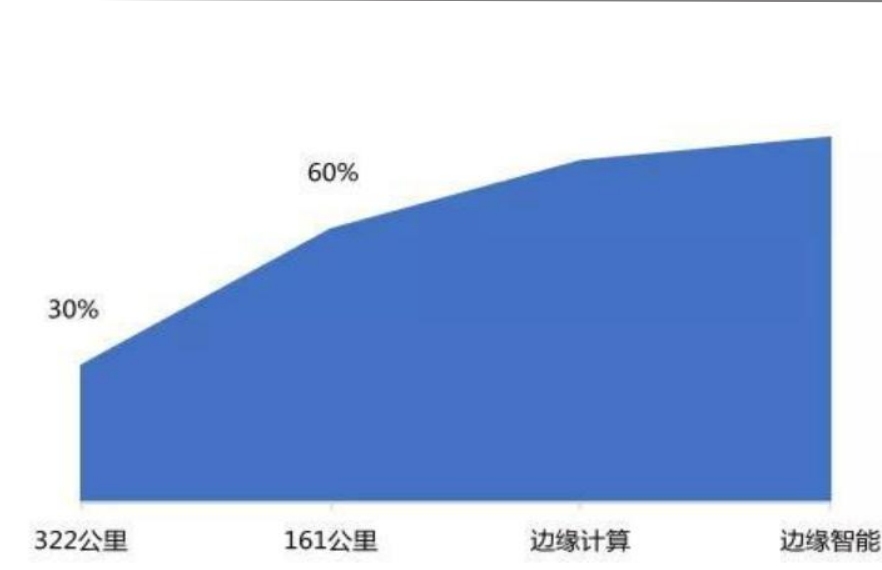
◆ AI芯片以高并行架构为主流模式。针对各类深度学习算法基础共性需求优化的指令集和高并行计算架构、高能效的内存存取架构和高速易拓展的互联接口。面向不同应用场景的需求，不同芯片产品间差异较大。

	终端	云端	边缘端
芯片需求	<div><ul style="list-style-type: none"><li>· 低功耗、高能效</li><li>· 推理任务为主、成本敏感</li><li>· 硬件产品形态众多</li></ul></div>	<div><ul style="list-style-type: none"><li>· 高性能、高计算密度</li><li>· 兼有推理和训练任务、单价高</li><li>· 硬件产品形态少</li></ul></div>	<div><ul style="list-style-type: none"><li>· 对功耗、性能、尺寸的要求常介于终端与云端之间</li><li>· 推理任务为主、多用于插电设备</li><li>· 硬件产品形态相对较少</li></ul></div>
典型计算能力	<8TOPS	>30TOPS	5TOPS-30TOPS
典型功耗	<5瓦	>50瓦	4瓦-15瓦
典型应用领域	各类消费类电子、物联网产品等	云计算数据中心、企业私有云等	智能制造、智能家居、智能零售、智慧交通、智慧金融、智慧医疗、智能驾驶等众多应用领域

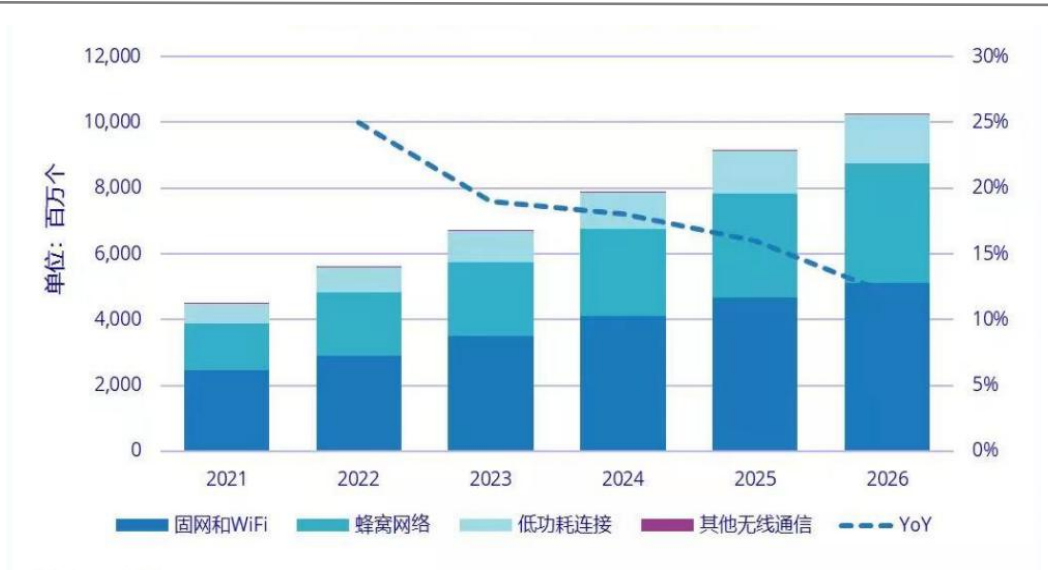
# 趋势三： 缩减数据处理成本

- ◆ 物联网连接设备规模猛增，下沉各行业。IDC预测，2022年中国物联网连接规模达56亿个，到2026年将增至102.5亿个，复合增长率约为18%，其中，消费者行业是最大的物联网连接组成，智能家居、可穿戴依然是重要增长点，连接数量到2026年将近59.8亿个。
- ◆ 边缘计算降低数据处理成本。当边缘与云端距离越短，数据处理成本越低。边云距离减少到 322 公里的时候，成本将缩减 30%，当距离为 161 公里的时候， 成本将缩减 60%。

图表3 数据处理成本缩减（%）结果



图表4 中国物联网连接规模预测，2022-2026

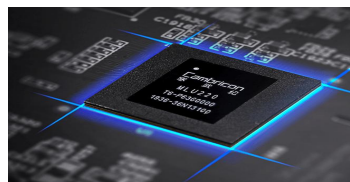




# 边缘算力三种体现：模组、终端、边缘计算中心

- ◆ 第一类，通过边缘算力芯片提供，形式为模组，通过定制 PCB 板输出，或者通过模组形式输出。

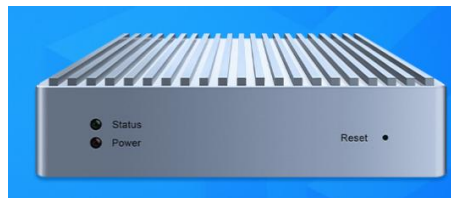
图表5 寒武纪思元220边缘计算模组



- ◆ MLU220是一款专门用于边缘计算应用场景的AI加速产品（边缘人工智能加速卡），可应用于智能机器人、智慧零售、智慧金融、智慧工厂等场景。

- ◆ 第二类，通过边缘算力芯片提供，形式为服务器或者边缘盒子。

图表6 瑞驰AI边缘终端



- ◆ AI边缘智能终端是瑞驰自主研发、基于嵌入式架构的软硬一体机设备，包含智能终端和智能一体机两种形态，具备高性能、低功耗、国产化等优势特色。

- ◆ 第三类，类似于传统数据中心，通过将机柜布置在离用户较近的机房中，便捷的活动本地算力。

图表7 闪讯边缘云EdgeON智能边缘机柜



- ◆ EdgeON智能边缘机柜采用42U服务器机柜，可以部署在通讯机房以及冷却设备的独立安全计算环境，可根据业务需要安装部署标准服务器和边缘云计算环境，满足客户对边缘云计算的需求。

边缘计算将构筑“连接+计算+智能”ICT一体化信息技术底座，使能新型基础设施

网络能力

计算能力

存储能力

感知能力

.....

协同联动

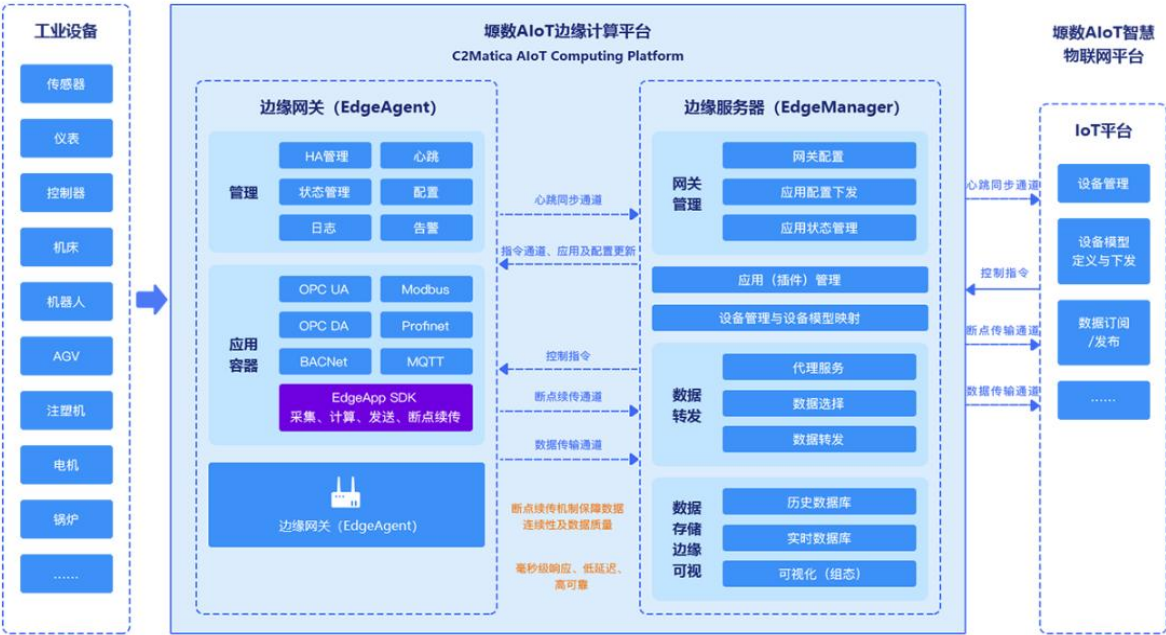
**工业现场边缘计算：**可依靠算力支持百万点高并发及高频工业设备接入，涵盖PLC、DCS、DTU、RTU、数控机床等设备。可以实现AI服务的调用与管理，工业设备灵活部署，提高现场总线能力，提高分析分析和操作效率。

**电力监测场景：**国家电网接入终端包括无人机、巡检机器人、长距离智能检测、智能传感器，在线监测中压配电网线路电流和对地电场；实现接地故障准确定位、复杂故障回溯反演、线路异常提前预警。

图表8 智能化配电网线路状态监测系统



图表9 工业场景中边缘案例



资料来源：映翰通官网，源数科技官网，华金证券研究所

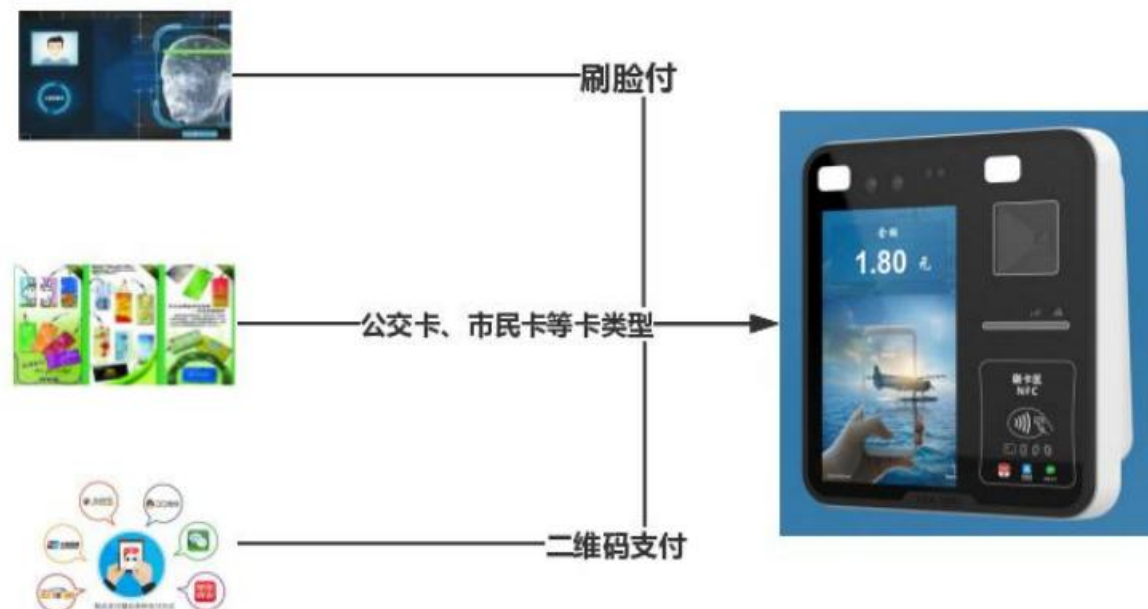
请仔细阅读在本报告尾部的重要法律声明



# 消费级行业:自动售货、POS机人脸识别

- ◆ **POS机人脸识别**：采用云计算，将人脸识别部署在服务器，发挥服务器强大的运算能力和庞大的人脸库数据优势;商米科技SUNMI P2 Xpro采用8核2.0GHz处理器，运用3D结构光识别技术可将20000多个投射到人脸的散斑点进行360度全方位轮廓追踪。
- ◆ **AI自动售货柜**：采用端计算，将识别模型部署在本地,从而消除网络传输的延迟; SandStar视达福星AI智能售货机采用6NM制程芯片。

图表10 POS机人脸识别支付场景



图表11 自动售货柜场景



01 算力向推理和边缘扩散

02 产业链重构带来价值

03 车和垂直行业弹性爆发

04 投资逻辑和重点标的

05 投资建议与风险提示

# 芯片侧：RISC-V更适合边缘架构

- ◆ 三大主流架构：X86架构多用于高性能计算领域，ARM架构多用于移动互联网领域，RISC-V具有架构永久开源、指令集精简且高效、CPU微架构模块化、架构扩展性强等若干特征，完美契合物联网领域设备多元化、碎片化的场景；
- ◆ 三星、英特尔、英伟达、高通、联发科、谷歌发起全球 RISC-V 软件生态计划“RISE”，推动 RISC-V 处理器在移动通信、数据中心、边缘计算及自动驾驶等领域的市场化落地。

图表12 不同芯片性能示意图

		X86	ARM	RISC-V
架构类别		CISC(复杂指令计算机)	RISC(精简指令集计算机)	RISC(精简指令集计算机)
流水线及硬件复杂度		流水线指令复杂、硬件实现难度大	流水线指令精简、硬件实现相对容易	流水线指令精简、硬件实现相对容易
指令集整体性能	模块化	不支持	不支持	支持模块化指令集
	可扩展性	不可扩展	不可扩展	支持第三方扩展定制指令
	能耗	高	低功耗	低功耗
应用实例		高通600E和410E、华为Boudica 120和150以及三星Artik1、5、10等物联网芯片	英特尔Edison、Curie	目前仍是微处理器，特斯拉已加入RISC-V基金会；
适用领域		桌面、HPC	移动互联网	物联网等新兴领域

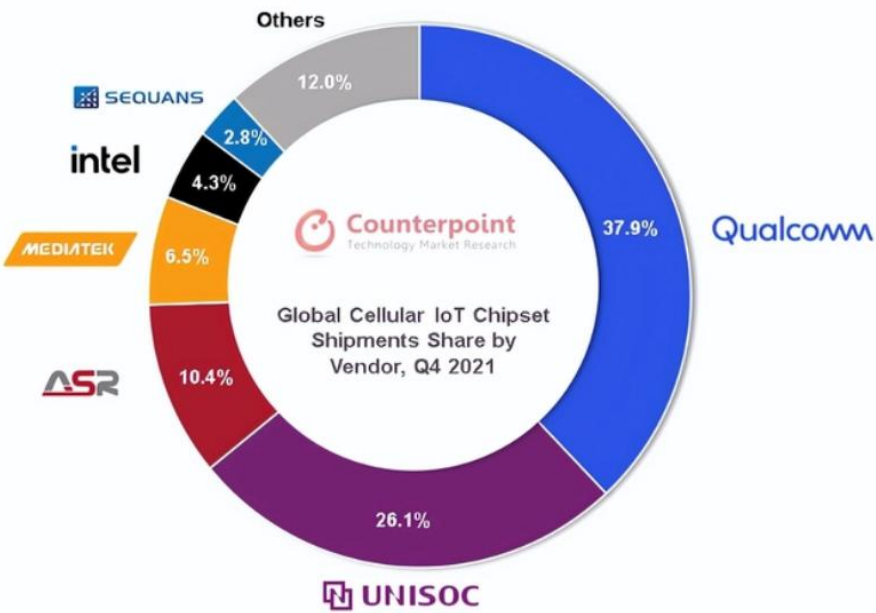
# 芯片侧：高通在物联网侧一骑绝尘

苹果M1、M2计算芯片：基于ARM架构，主要用于其生态体系内的如Iphone手机、电脑、Ipad 等产品内；

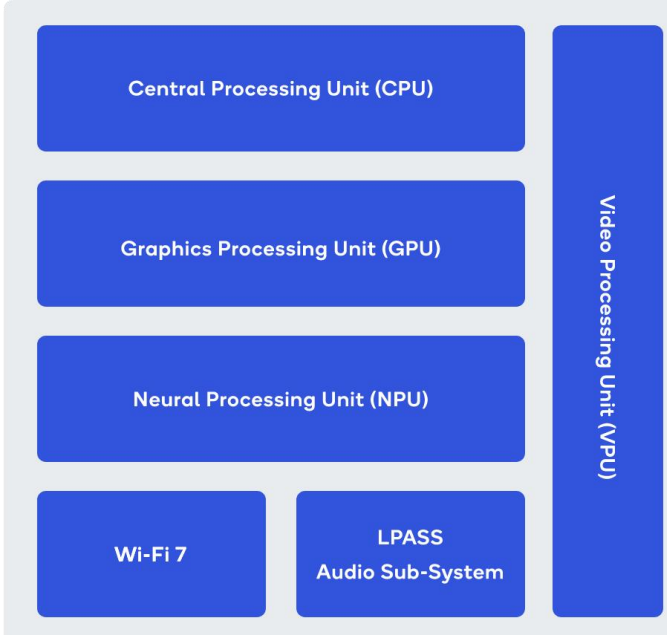
英伟达JESTON AGX Orin：边缘的AI计算平台级芯片，同时继承了安培架构的GPU和 ARM Cortex-A78，既可以做推理也可以做训练。作为一个边缘智能芯片产品，其有200Tops的处理性能（INT8）边缘产品主要是车侧的自动驾驶芯片如 Orin；

高通：基于骁龙系列手机芯片推出了一系列专为边缘侧设计的模组芯片，当下主流的物联网算力场景，如智能车机，智能零售等，普遍采用高通芯片来提供算力和搭载系统。

图表13 Cellular IoT芯片供应商份额



图表14 高通芯片



- ◆ 全新高通QCS8550和高通QCM8550处理器整合强大的算力和边缘侧AI处理、Wi-Fi 7 连接以及栩栩如生的图形和视频功能；
- ◆ 配备5G和Wi-Fi 6E连接，可实现数千兆比特传输速率、更广泛的覆盖范围、低时延和高效的处理能力

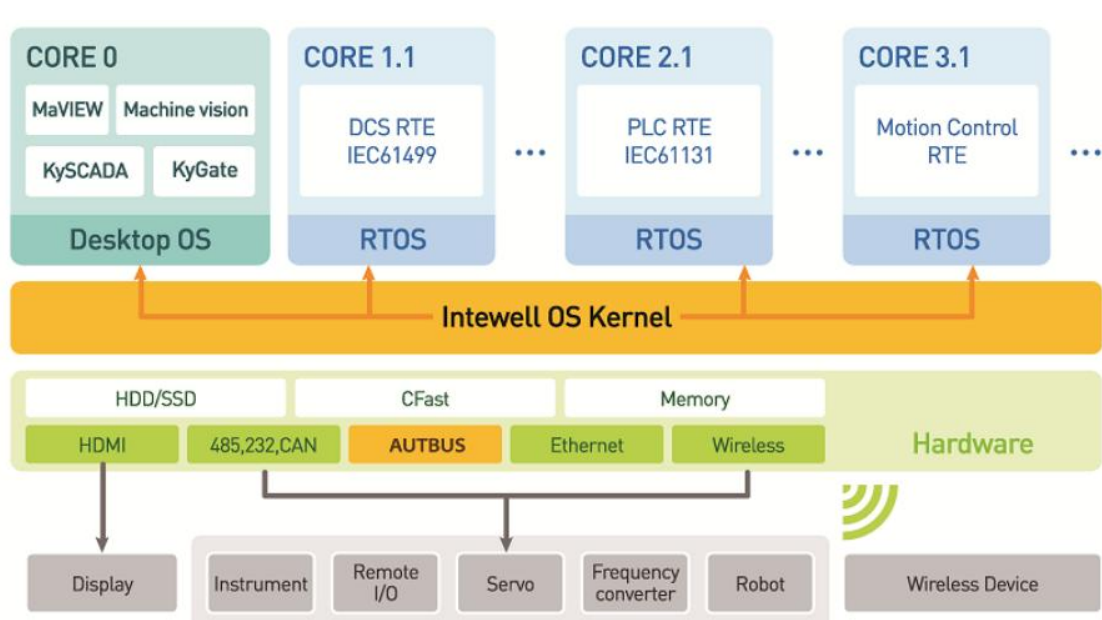
# 系统侧：“平台+操作系统” 边缘算力中枢

- ◆ **边缘物联网平台。**基于边缘侧向上延伸的物联网平台，能够快速部署或快速扩展，无需现场调试。移远通信基于移远云构建了全新的物联网开放平台，提供SaaS、APP、数据大屏等应用，一站式为全球客户提供创新有效的解决方案；
- ◆ **边缘侧操作系统。**给边缘侧设备提供支撑运行智能化应用以及高实时控制应用。东土科技Intewell操作系统，嵌入式控制/仪器仪表/传感器，支持工业总线、工业无线互联互通，支持X86/ARM/飞腾等体系架构。

图表15 移远通信物联网云服务平台QuecCloud



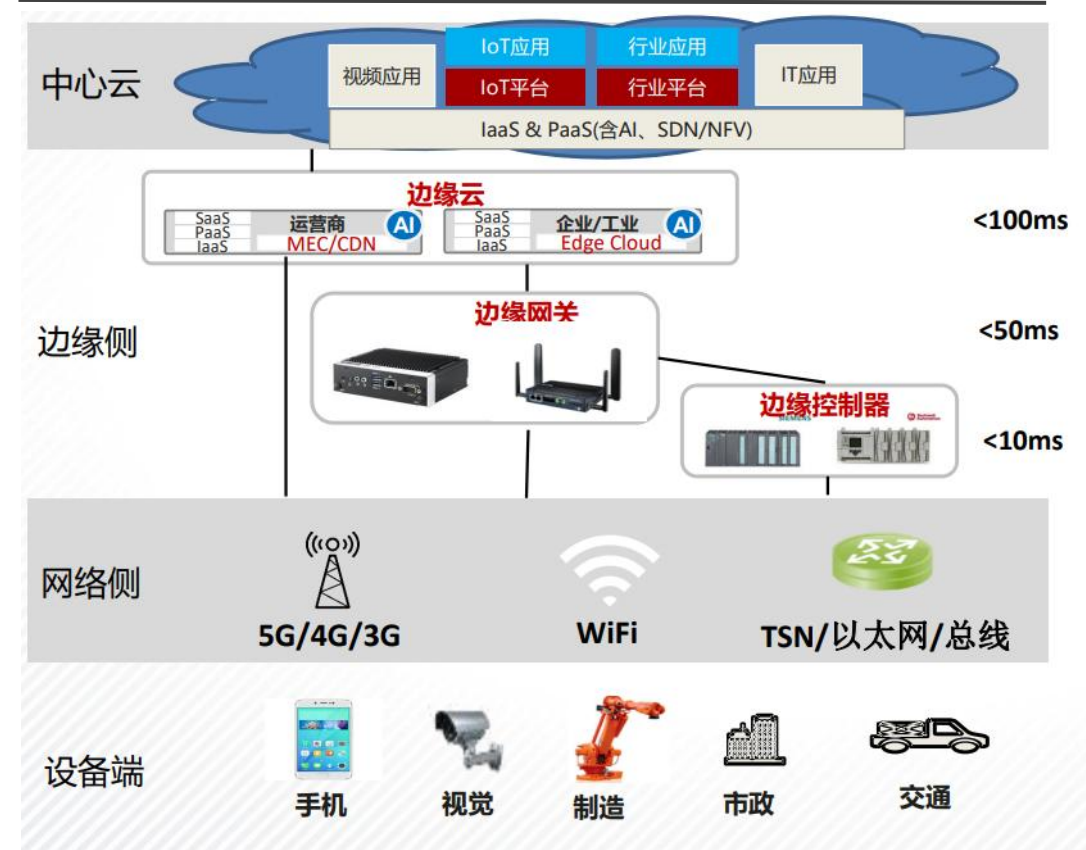
图表16 东土科技边缘操作系统





设备侧主要有工业交换机、边缘服务器、边缘网关，三大设备支撑构建边缘侧解决方案；边缘网关：对接不同网络协议，保障不同网间数据传输；工业交换机：汇聚不同节点数据，保障数据的安全传输及分发；边缘服务器：边缘算力的载体，实现边缘端数据的处理及控制。

图表17 边缘计算载体



图表18 工业交换机



- ◆ 应用于工业控制领域的以太网交换机设备。其在边缘侧的作用在于拥有零治愈环网技术性，可以保证数据的可靠性及一致性。

图表19 边缘服务器



- ◆ 将传统服务器的计算能力从中心点下沉到靠近用户端的小型服务器。其在边缘侧的作用主要有数据延迟低、可适应恶劣场景、可降低带宽需求和成本。

图表20 边缘网关



- ◆ 连接终端设备和云端服务器之间的设备，可以实现对设备和物联网环境的实时监测和控制。其在边缘侧作用主要有降低整体能耗、简化通信层架构、灵活部署。

- ◆ IDC建设从开始的传统机房，逐渐演变成模块化、微模块化、全微模块化；模块IDC采用密闭通道，分为IT机房室、空调室、UPS&电池室。机柜密度一般为4-7KW/Rack，对建筑物要求低，施工周期短，一站式采购，移机能力强。
- ◆ 微小散热模块主要包含导热界面材料、热管、均热板和 3D Vapor Chamber等。其中均热板与热管是芯片级散热的重要元件，并且均热板扁平型的结构使得其转移热量更大。

图表21 模块IDC示意图



图表22 微小散热模块方案示意图



01 算力向推理和边缘扩散

02 产业链重构带来价值

03 车和垂直行业弹性爆发







04 投资逻辑和重点标的

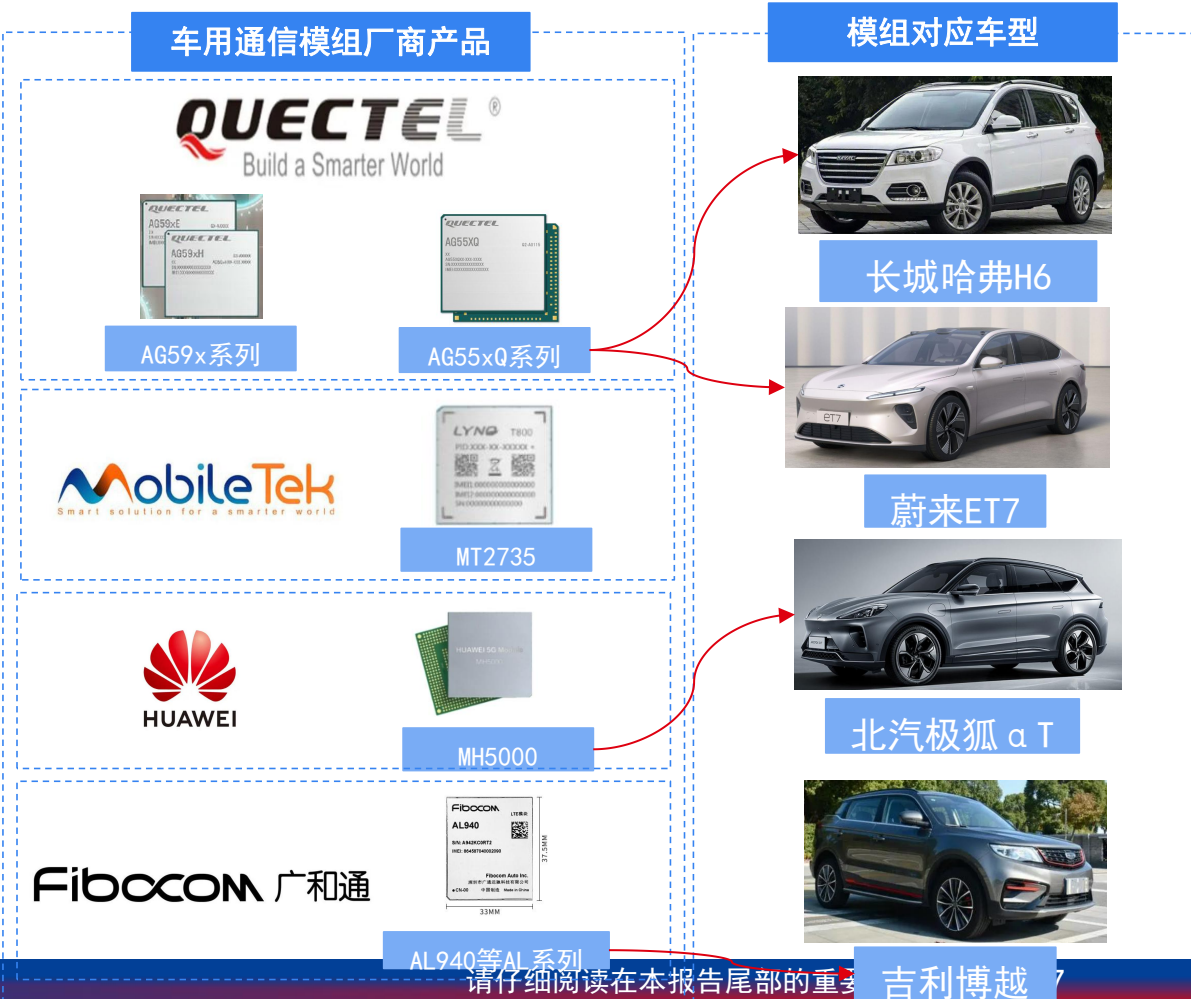
05 投资建议与风险提示

# 1：车辆边缘算力的iot模组产业链拆解

- ◆ 高通已经开发出多款专门面向IoT领域的芯片，涵盖0.2TOPS到48TOPS不等，包括eMTC、NB-IoT等，可以满足不同应用场景的需求；
- ◆ 多款车用通信模组应用主流车型，包括华为MH5000 5G芯片应用于北汽极狐。

图表23 高通物联网解决方案及其目标应用

	芯片	发布时间	算力	目标应用
	QCS8550/ QCM8550	2023.4	48TOPS	自主移动机器人、工业无人机、沉浸式云游戏、视频协作等
	QCS4490/ QCM4490	2023.4	1~10TOPS	先进零售和POS终端、控制和自动化应用、安防面板等
	QCS8250	2021.6	15TOPS	联网医疗、数字标牌、零售及视频协作等
	QCS6490/ QCM6490	2021.6	14TOPS	联网医疗、物流管理、零售、交通运输及仓储等
	QCS4290/ QCM4290	2021.6	1TOPS	摄像头、工业手持设备和安防面板等
	QCS2290/ QCM2290	2021.6	0.2TOPS	摄像头应用、工业手持设备、零售和资产跟踪等










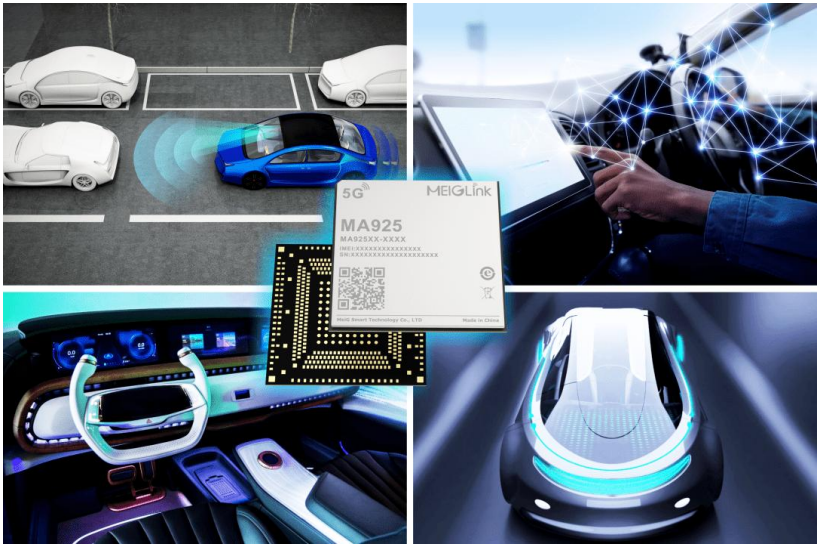
# 1：车辆边缘算力的iot模组产业链拆解

- ◆ 美格智能与比亚迪深度合作，依托高通平台推出面向智能座舱、网联化、辅助驾驶等不同场景的模组；基于比亚迪DiLink4.0和DiLink5.0平台，覆盖汉、唐、腾势、海豹等多种车型。
- ◆ 公司面向智能座舱的第一代5G智能模组产品已在比亚迪多个主流车型上大规模量产，公司推出的第二代5G智能模组已正式在比亚迪中高端车型上成功量产，智能座舱模组价值量达千元，远高于数传等传统模组。

图表24 美格智能模组汽车应用领域

	解决方案	代表模组	基于平台	目标应用
智能座舱	一芯多屏	SRM930 	高通QCM6490	可以实现“一芯多屏、多屏触控、多屏联动”，支持WIFI 6E，2x2 MU MIMO
	ADAS/DMS	SRM930 	高通QCM4290	支持LTE Cat. 6和2x20MHz载波聚合最大下行速率可达300Mbps，支持L1+L5双频GNSS
网联化	T-Box	MA525 	高通新一代5G车联网通讯平台	采用Soc架构芯片，支持3GPP R16标准，集成AP处理器，最大具备20K DMIPS算力
辅助驾驶	定位(车规级)	MA925 	第二代骁龙汽车5G调制解调器及射频平台	支持3GPP 5G R16标准，支持5G NR独立组网和非独立组网模式，支持选配C-V2X
	定位(工业通信)	SNM970 	高通QCS8550	具备SoC异构计算特性，支持WIFI 7，综合AI算力可达48Tops

图表25 美格智能与比亚迪合作

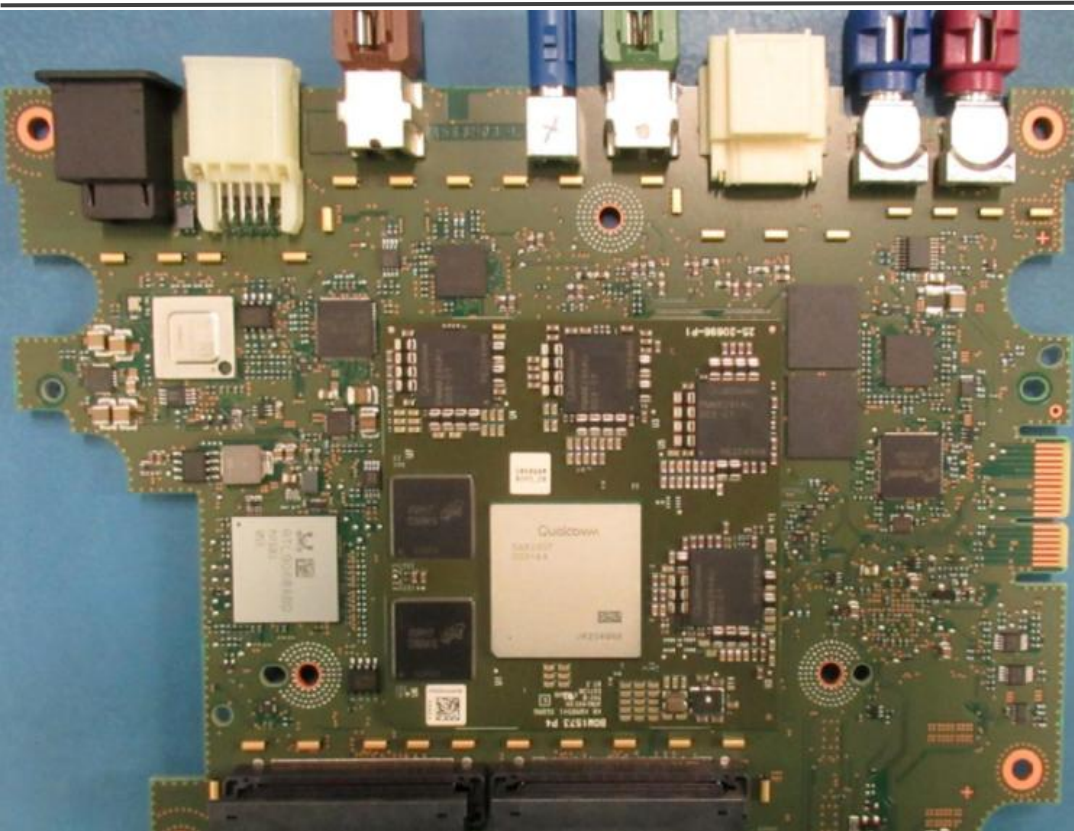




# 1：车辆边缘算力的iot模组产业链拆解

◆ 汽车座舱主要分为三个子系统，分别是收音的CSB、基础功能的BB和先进多媒体功能的MMB；其中MMB系统是智能化中枢，最新一代采用了高通第四代座舱旗舰SA8295P；该模块售价高达250美元，包括4片电源管理和2片LPDDR5、基础软件等，运行频率可能高达3.00GHz，算力达30TOPS，应用在奔驰E级车型。

图表26 奔驰座舱先进多媒体功能MMB板

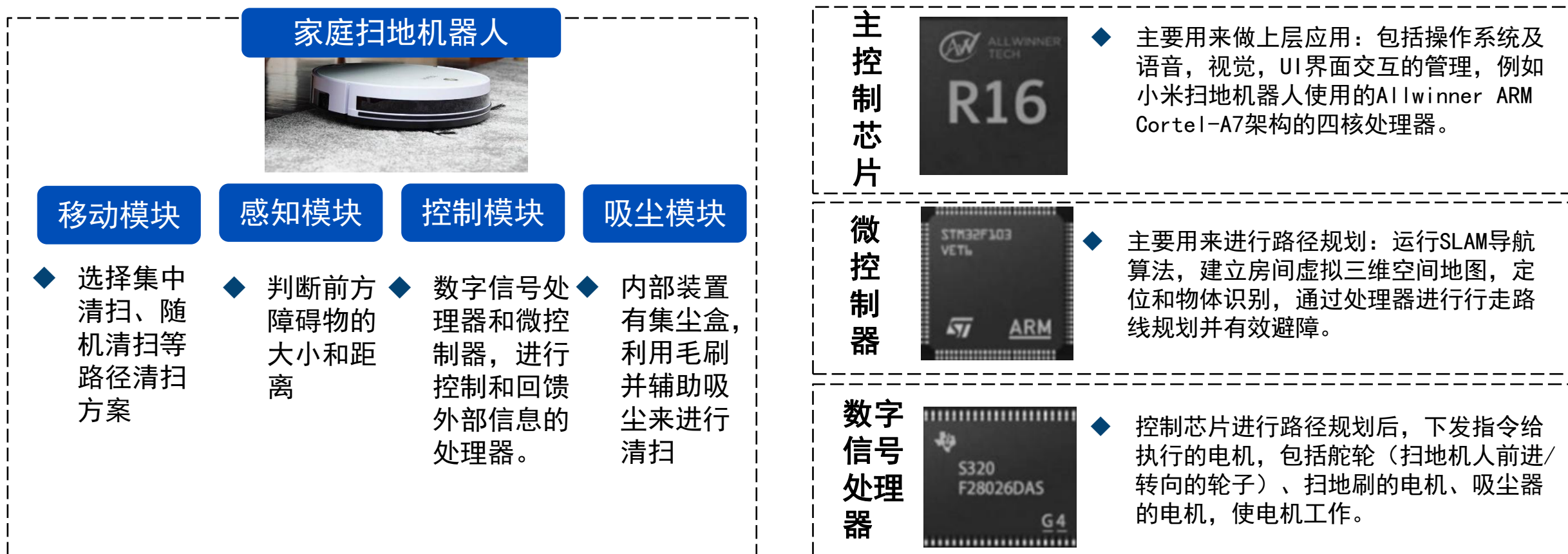


图表27 MMB板芯片拆解

芯片部位	芯片型号	芯片价格	芯片功能
主控芯片	高通 SA8295P	250美元	可支持MMB板卡中多个 ECU 和域的融合，包括仪表盘与座舱、AR-HUD、信息影音、后排显示屏、电子后视镜和车内监测等功能
以太网交换机芯片	MICROCHIP KSZ8895	4美元	为MMB板卡提供以太网连接，传输高质量、低延迟的音频和视频，支持AVB 2.0标准和IEEE 802.3at PoE+
接口芯片	HDBaseT V6000	100美元	使得MMB板卡可以传输高清视频和音频，符合HDCP 2.2标准可用于传输受保护的内容，支持4K 60Hz
串行芯片	美信 Max9288	20美元	解决MMB板卡的图形I/O功能，将HDMI转为GMSL同轴或STP输出，也可将360环视输入到串行芯片中处理。

## 2 扫地机器人：“大脑+小脑” 扫地机器人三大处理器

- ◆ 扫地机器人系统可分为移动模块、感知模块、控制模块、吸尘模块四个模块；需要主控制芯片、微控制器、数字信号处理器；
- ◆ 主控芯片是神经中枢，小米扫地机器人采用Allwinner ARM Cortel-A7架构的四核处理器；微控制器是移动控制中枢，用于规划路径、规避障碍；数字信号处理器是信号处理中枢，用于传递控制信号给各执行器。





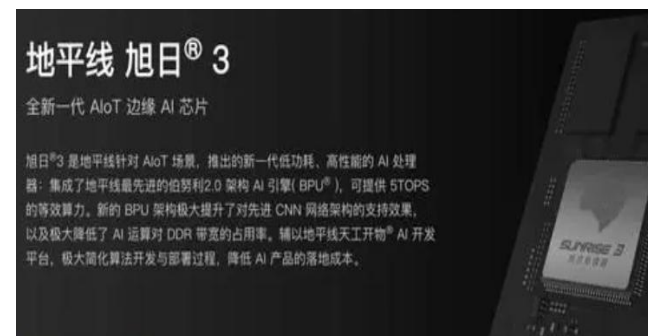
## 2 扫地机器人：边缘AI算力芯片

- ◆ 目前各大厂商针对扫地机器人场景推出边缘AI芯片，如地平线推出旭日芯片，达到5TOPS标准算力、可满足辅助驾驶AI计算需求性能；科沃斯搭载的地平线芯片，性能提升显著，感知速度提升4倍，最低延迟可降至60毫秒。石头科技搭载的高通骁龙芯片，嵌入卷积模型后，识别准确性提升。



- ◆ 石头扫地机器人T7 Pro运用AIoT开启智能家电；
- ◆ 采用了两颗500万像素，120° 广角摄像头，通过立体视觉识别技术和AI物体识别技术，用大视角双目摄像头获取环境深度信息

- ◆ 高通骁龙625处理器作为数据处理的后盾
- ◆ 基于对数万张图像训练而成的卷积神经网络，T7 Pro可以快速识别、准确处理、及时避障。



- ◆ 自动驾驶芯片公司地平线一款AIoT边缘AI芯片；
- ◆ 采用16nm工艺的应用SoC处理器；2.5W的功耗达到等效5TOPS的标准算力；可满足L2+辅助驾驶AI计算需求性能。

- ◆ 科沃斯搭载了旭日3，发布地宝X1
- ◆ 该产品可识别物体从原有的5种提升为15种，感知速度较之前提升4倍，最低延迟可降至60毫秒，并能够感知到物体的三维信息。



## 2 扫地机器人：关键部件核算

- ◆ 全规划扫地机器人单台成本大概是1000元，其中主控制板（PCB主板）（70-240元）和激光导航传感器（130-200元）；导航芯片（50-56元）；传感器（42-90元）；主芯片24元；电机35元；电池芯片70-100元。
- ◆ 地平线旭日3AI开发板，包含旭日X3M芯片，具有5TOPS端侧推理，价格（500-1000元）。

图表28 智能扫地机器人核心构成表

组成部分	功能
cpu	核心的运算芯片，控制及其的所有动作与反馈
mcu	微处理器，控制各种传感器，中断，实时性高
ddr	运用于机器人运作时数据存储
flash	存储产品软件
omic	电源管理集成电路
wifi	与智能设备连接，操控
acc	加速度计、陀螺仪

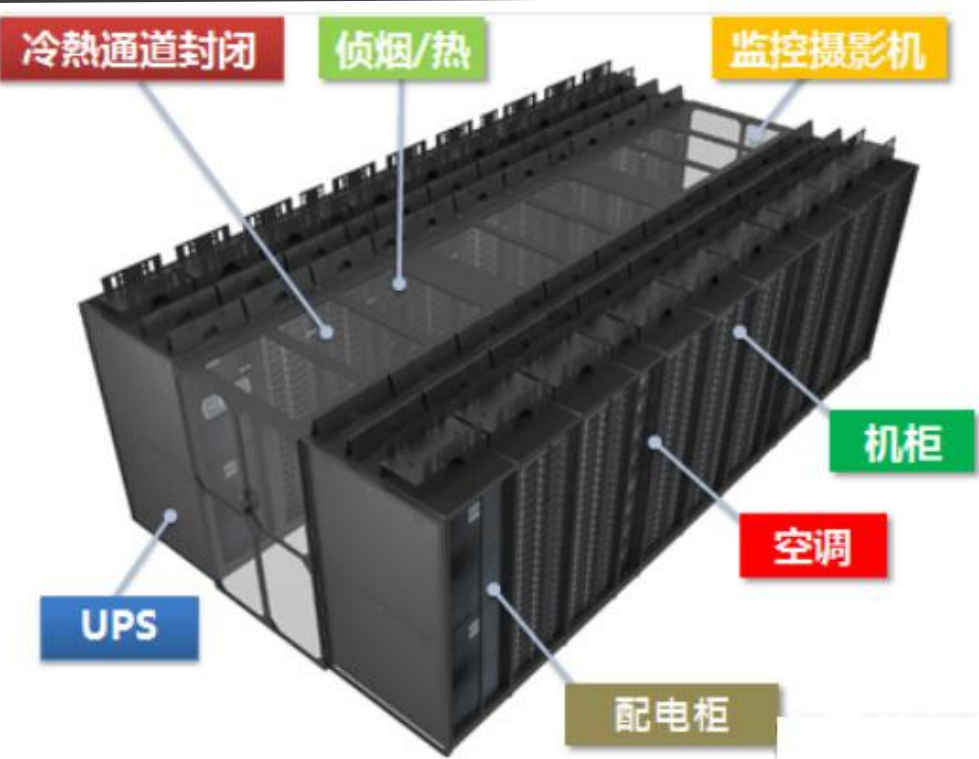
图表29 智能扫地机器人核心构成表

产品	公司
LDS测距模组（激光雷达）	Indemind、速感科技、诠视科技、远行时空、瑞盟、信泰光学、思岚科技、镭神智能、YDLIDAR等
芯片模组	国产厂商全志科技、瑞芯微
电池行走轮模组、风机、结构件	德赛电池、比亚迪、蓝微电子、力嘉塑料
WIFI模组	科沃斯
嵌入式MCU	高通、意法半导体、兆易创新
触发逆变器	米家
微处理器	inxni

### 3：微模块idc边缘算力拆解

- ◆ 边缘计算与微模块IDC相辅相成。边缘计算提供让数据处理和存储能力更接近数据源的技术或模型，微模块IDC作为一种数据中心设计可为现实环境中的边缘计算提供动力支持。
- ◆ 微模块IDC边缘计算可以获得数据安全、减少带宽、降低延迟，可根据企业需求便捷部署。

图表30 微模块IDC拆解图



图表31 微模块IDC产业链情况

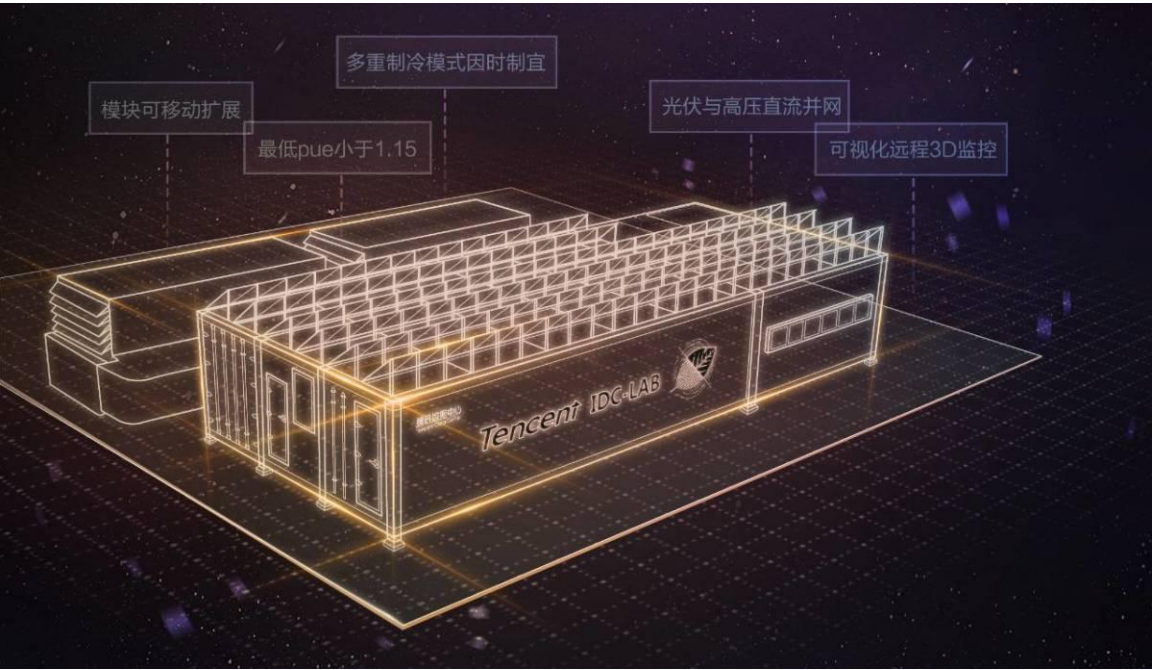
产业链名称	参与公司	公司产品或方案
机房空调	维谛技术	Liebert XD 系列空调
UPS	科华数据	KELONG MR 系列
精密温控	英维克	全链条液冷方案
芯片级散热	双鸿科技	NB/DT产品线
半导体制冷	富信科技	TES1系列产品
热管理	中石伟业	EMC解决方案



### 3：微模块idc边缘算力拆解

- ◆ 微模块数据中心单个机柜一般可以提供数百至数千的算力，通常由十余个机柜组成。
- ◆ 目前在智能交通、新零售、自动驾驶、工业互联网、智慧校园、线上医疗等大量边缘场景中得到大量应用，主要解决了应用场景中快速建设、灵活部署与扩容、设备可靠性高、因地制宜、投资保护性强等痛点问题。

图表32 腾讯T-block微模块数据中心产品图



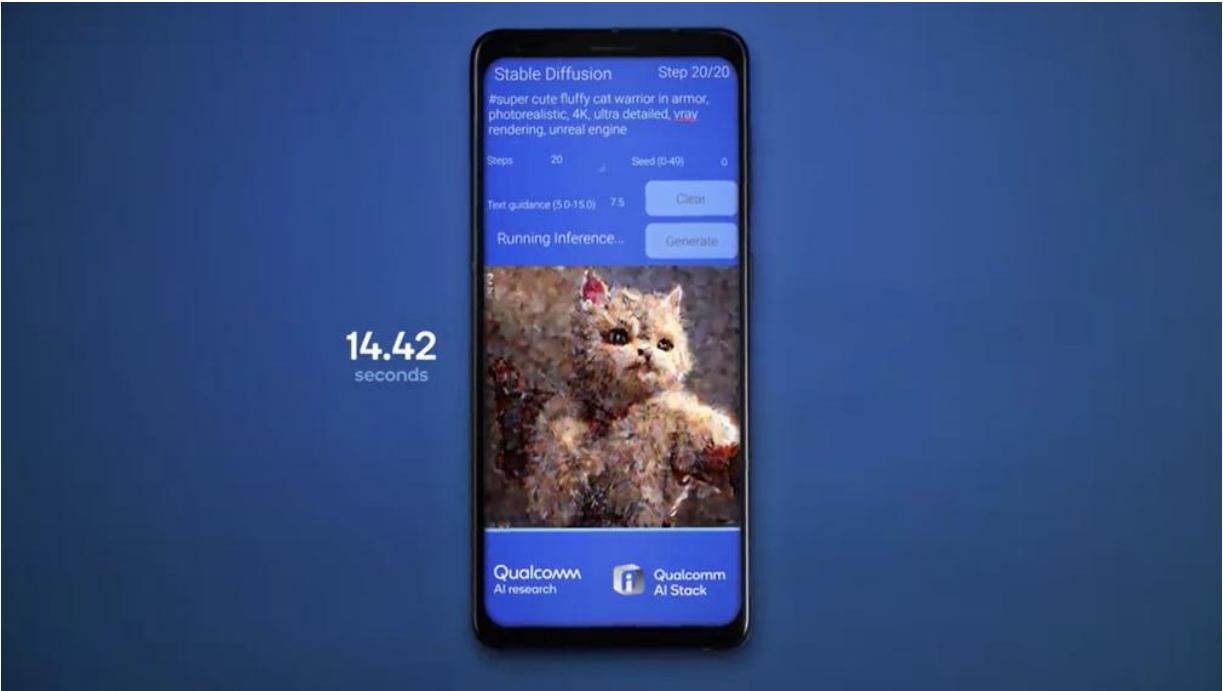
图表33 腾讯T-block参数情况

参数名称	参数情况
单机柜功耗	10kW~20kW
模块规格	20尺4机柜或40尺6-8个机柜
散热制冷模式	新风模式、喷淋蒸发模式、压缩机辅助制冷模式
液冷方式	间接蒸发冷技术、氟泵自然冷技术
PUE	1.0955
CLF制冷负载系数	0.0812
光伏使用系数	0.1982

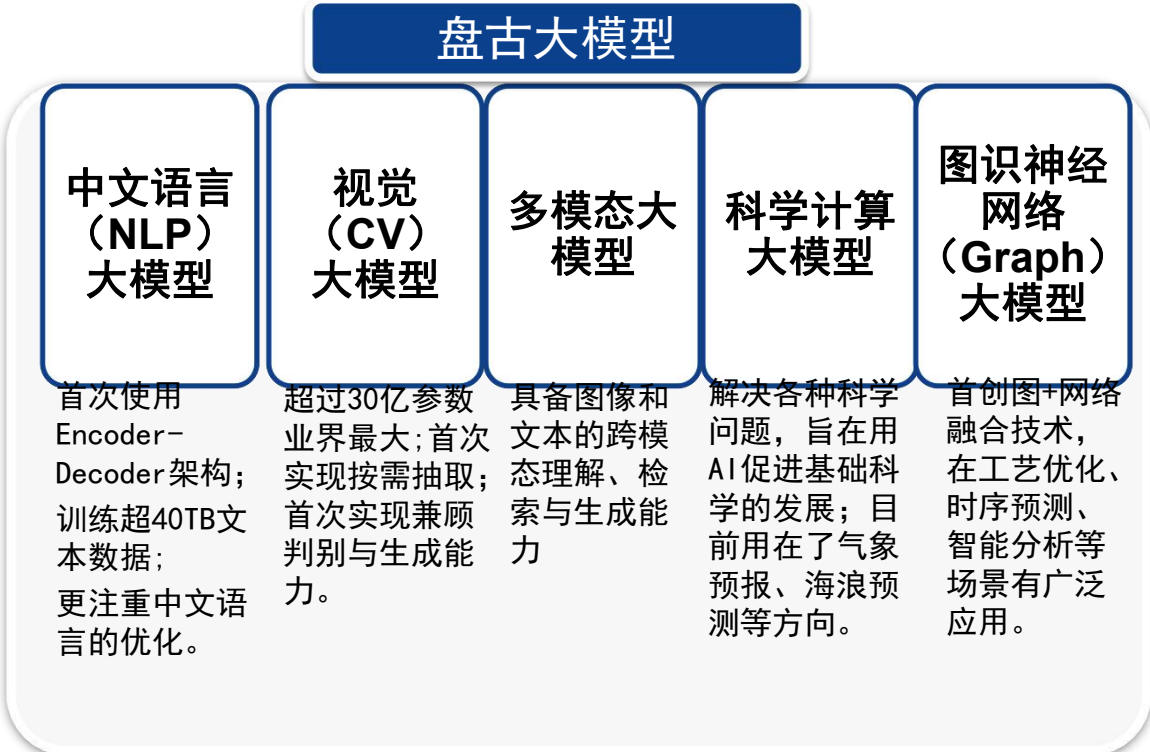
- 01 算力向推理和边缘扩散
- 02 产业链重构带来价值
- 03 车和垂直行业弹性爆发
- 04 投资逻辑和重点标的
- 05 投资建议与风险提示

- ◆ 边缘终端嵌入大模型。高通在骁龙芯片上成功推理 stable diffusion 模型，华为推出基于手机算力的“智能搜图”功能。
- ◆ 智能模组是边缘算力的最佳承载模式。华为已自研盘古大模型，面向ToB/ToG的政企端客户打造多模态千亿级大模型，赋能多行业多场景。

图表34 高通在Android手机上运行Stable Diffusion



图表35 华为盘古系列大模型图解



# 高通：智能边缘计算全布局

- ◆ 高通正在从一家通信公司过渡到一家智能边缘计算公司，在智能手机业务的基础上，加大了汽车、网络、计算和可穿戴设备领域的增加；推出业内首个集成式汽车超算SOC，最大算力达2000TOPS； QCS8550处理器算力达48TOPS。
- ◆ 高通形成“边缘+算力”平台型公司，手机、射频领域对接终端厂商，汽车、物联网领域通过集成商、模组厂商间接对接以实现终端布局最大化。

图表36 高通业务领域布局

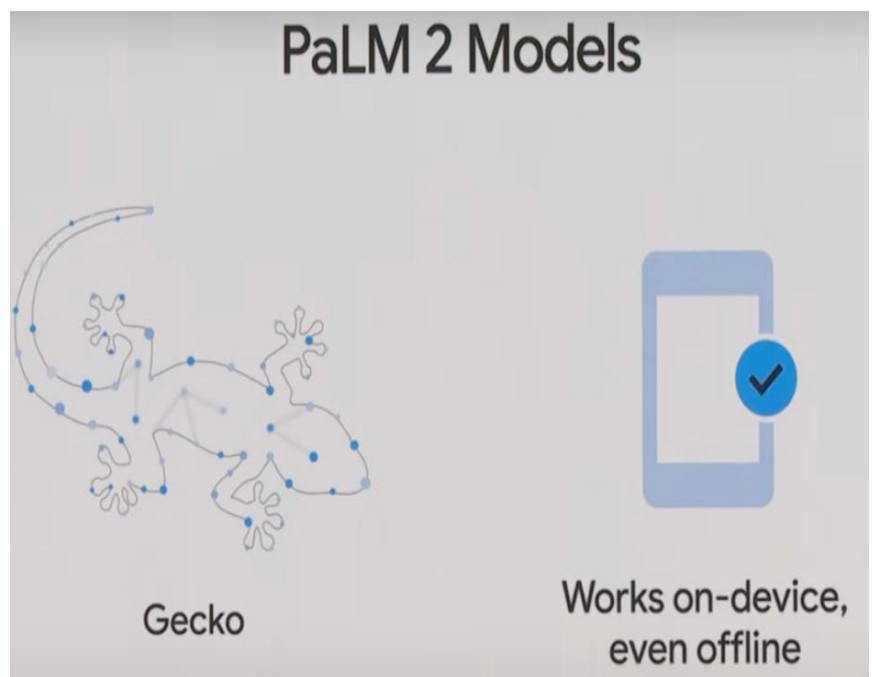
手机领域	高通骁龙手机平台：高通骁龙660、870、8Gen2处理器，低、中、高端手机全布局	骁龙8Gen2处理器：一颗Cortex 3. 19Ghz+4颗2. 8Ghz+三颗2. 0Ghz小核心	小米、OPPO、vivo、荣耀
汽车领域	骁龙汽车数字座舱平台：数字座舱、ADAS、自动驾驶、车联网	第四代骁龙座舱平台：5nm制程，AI 算力30TOPS，相对8155整体性能提升2倍、3D渲染性能提升3倍	40余家中国汽车品牌超过100款车型
物联网领域	QCS8550、QCM4490等处理器：聚焦物联网领域生态系统，包括边缘 AI 处理、高能效、超清晰视频和 5G 连接等	高通®QCS8550处理器：8核高通®Kryo™ CPU、综合AI算力高达48Tops、支持Wi-Fi 7等特性	美格智能、移远通信、广和通等模组厂商
射频前端	目前5G端到端唯一一站式供应商，包括天线开关调谐到数字化处理、信号滤波和功率放大	射频前端模组：支持WIFI7、支持高达30Gbps的网络数据吞吐量	主流手机厂商



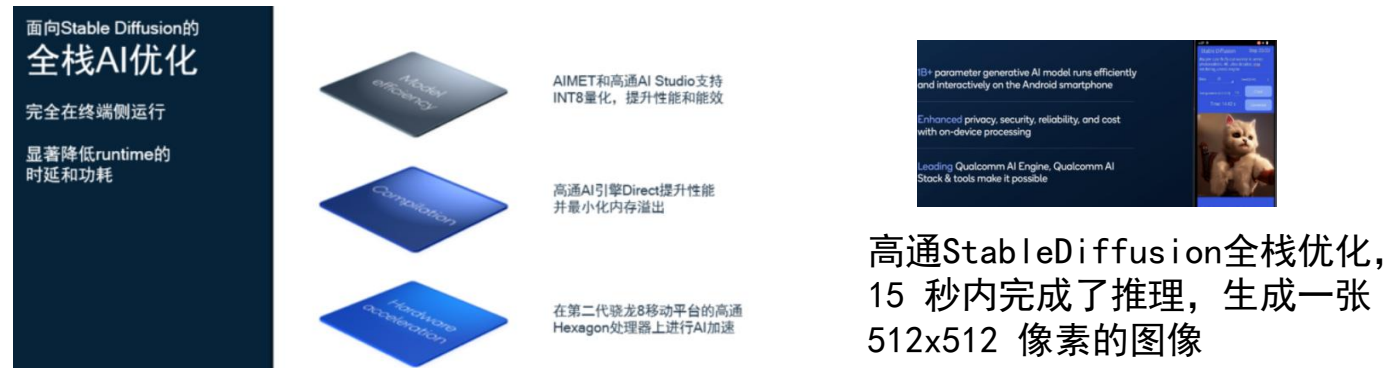
# 边缘终端嵌入（谷歌壁虎模型）：硬件优化+降低模型参数

- ◆ 我们认为多模态模型在边缘端应用需进行两方面改造，一是硬件优化；二是降低模型参数。目前谷歌最轻量级的Gecko（壁虎）可以直接在手机上离线使用，且每秒可以处理约20个token，对应16-17个单词，基本满足移动设备用户的需要。
- ◆ 硬件优化：利用高通AI软件栈执行全栈AI优化，首次在Android智能手机上部署Stable Diffusion。
- ◆ 降低模型参数：逐步蒸馏法训练大模型，实验表示，用7.7亿参数蒸馏可超过5400亿的大语言模型，最多可比原模型小2000倍。

图表37 谷歌壁虎模型

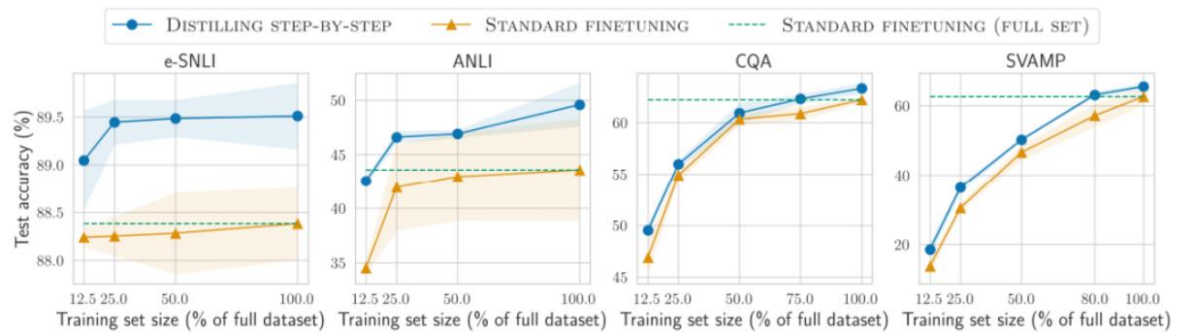


图表38 高通模型StableDiffusion模型优化



高通StableDiffusion全栈优化，15 秒内完成了推理，生成一张512x512 像素的图像

图表39 逐步蒸馏法效果对比





# 华为盘古大模型：推动行业AI开发从“作坊式”到“工厂式”升级

- ◆ 华为构建以矿鸿、工业承载网、云基础设施、数字平台和智能应用为核心的工业互联网架构智能矿山，基于盘古大模型覆盖矿山采、掘、机、运、通等主营业务开发和训练AI算法，目前正在掘进、综采、运输等16大类256个矿山应用场景展开科研攻关，并取得阶段性成果。
- ◆ 云边协同架构：集团侧高集群服务器进行训练开发，1000P算力性能；矿侧依托服务器+边缘小站，实时分析处理。
- ◆ 100亿参数预训练，短周期低成本开发：“手工作坊式” 单场景训练→”社会化工厂式” 开发，训练周期从6月→1月，分类精度58%→81%

图表40 华为矿山AI大模型架构图



- 集群（1000P算力）性能相当于50万台高性能PC
- 集团+矿侧2级架构，云边协同
- 数据不出集团，分析不出矿山

图表41 华为矿山大模型边缘算力图谱梳理



- 定制化开发、作坊式开发→“工厂式”
- 小样本学习强，分类精度58%→81%
- 机器视觉+AI实现主运输皮带堆煤、跑偏监测

智能应用

AI煤流检测

全景视频拼接

智能选洗煤

梅安森

云鼎科技

龙软科技

科达自控

山源科技

华夏天信

华洋通信

阳光三极

精英数智

江苏三恒

系统

矿鸿OS系统

数字平台

AI硬件

AI模块

智能小站

AI服务器

紫光股份

拓维信息

三旺通信

卓易信息

# 投资价值一：边缘预处理最先落地

◆ 通过边缘部署的算力，将用户的多样化需求进行本地的预处理后再上传至云端，可以节省网络资源、降低时延、减少成本。新兴的AIoT和工业物联网应用场景为众多边缘AI芯片设计公司带来更多机会，瑞芯微、全志、清微智能、酷芯微、亿智电子等国产公司加入赛道。

图表42 Token 概念图解



图表43 Token 云端及边缘侧特点对比



TOKEN发送至云端，能够最低成本的实现应用功能，加速商业化	面向小算力时，ARM架构更具成本优势，加速边缘小算力的渗透速度
成本	时效
AI模型推理Token费用较贵，边缘预处理可推进商用进程	边缘进行预处理后再至云端，可以节省网络资源，降低时延。

# 投资价值二：大模型百花齐放参数不一

- ◆ AI大模型参数一般从百亿起步。更多的参数意味着需要更多的计算资源，AI大模型相应的开启了算力军备赛。
- ◆ 国外目前AI大模型公司以OpenAI、谷歌、微软为第一梯队。
- ◆ 国内AI大模型井喷，百度、阿里、华为、浪潮信息 etc 公司都有大模型的发布，众多公司都在加紧研发布局。

图表44 主流大模型参数

公司	大模型名称	发布时间	大模型参数
 OpenAI	 GPT-3	2020. 6	17500亿
 Microsoft	 MT-NLG	2021. 10	5300亿
 Google	 Switch Transformer	2021. 1	16000亿
 商汤 sensetime	 INTERN	2021. 11	100亿
 华为云	 盘古大模型	2021. 4	1000亿

图表45 特斯拉自动驾驶大脑FSD拆解

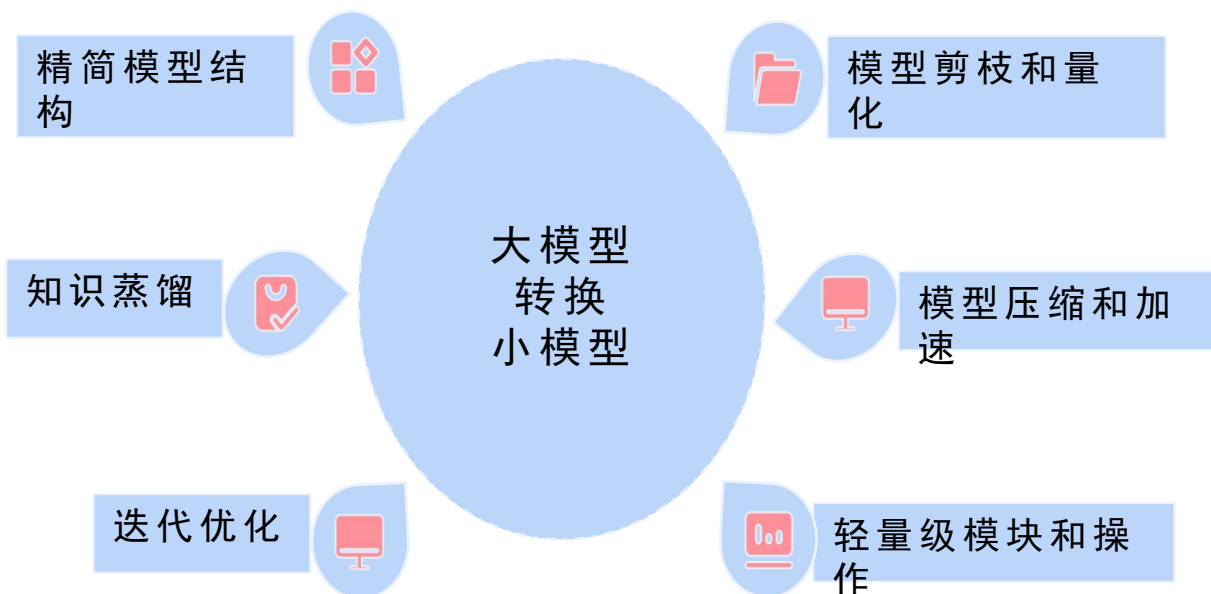


- ◆ 目前自动驾驶L3/L4瓶颈不在AI计算，而在储存带宽。以特斯拉FSD例子所示，每秒可加载5.12次权重模型。即使算力可达10万TOPs，每秒运算次数不会超过6次。

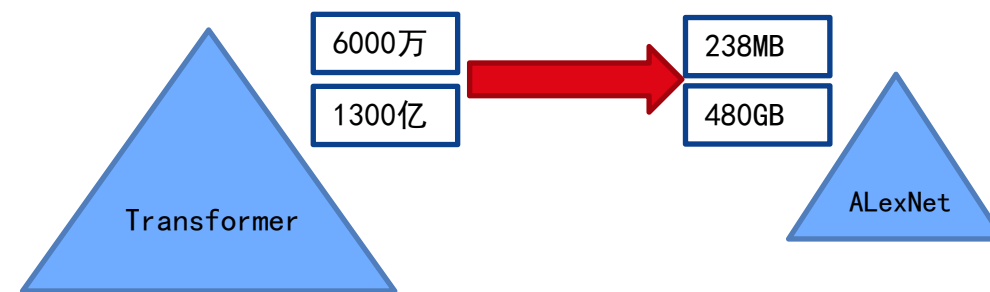
# 投资价值二：大模型转小模型是必须之路

- ◆ 深度学习的大模型通常具有数百万甚至数十亿个参数，需要大量的计算资源和时间来训练。在实际应用中，我们经常需要在资源受限的设备上运行小模型，因此我们常常采用方法将大模型转化为小模型。
- ◆ 未来将呈现从大模型转变为小模型的趋势。基于大模型开发的小模型针对性更强，可以迅速地应用到各行各业，给用户提供多种服务。

图表46 大模型转小模型常用方法



图表47 大模型转小模型例子



- ◆ AlexNet用于视觉识别神经网络，Transformer用于视觉应用模型。
- ◆ AlexNet在卷积层的参数量为34944，全连接层到卷积层参数量为37752832；Transformer在Embedding层参数量为31782912。
- ◆ 经计算可得6000万大模型可转换为238MB小模型，1300亿大模型可转换为480GB小模型。



- ◆ 高通通过与其芯片配套的全栈AI优化方案，将模型从 FP32 压缩至 INT8，显著的降低了运行时延和能耗，从而实现了模型在手机算力上的安全高效推理。
- ◆ 谷歌发布全新的语言大模型 PaLM2，其中拥有众多版本和参数量的模型体系，谷歌计划将模型嵌入旗下核心产品，重新构想产品，包括搜索、地图、邮件、视频等。
- ◆ 中科创达将智能音箱与机器人进行融合，并通过中科创达魔方 Rubik 大模型的不断训练，已经实现了能够自由对话的智能销售机器人。

图表48 谷歌PaLM2模型介绍

模型名	性能参数	应用场景
壁虎（Gecko）	100亿	可以在移动设备上运行，并且速度足够快，即使在离线时也可以在设备本身上提供交互式应用程序。
水獭（Otter）	1000亿	自然语言处理、机器翻译、代码生成等
野牛（Bison）	10000亿	可以处理更复杂任务，例如生成逼真的图像和视频
独角兽（Unicorn）	100000亿	可用于处理最复杂的任务，比如编写创意文本

图表49 中科创达魔方大模型




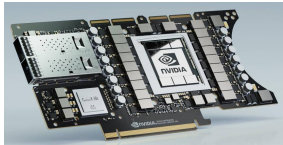

# 投资价值二：模型嵌入所需硬件支撑—算力/带宽/内存

- ◆ 目前面向行业的模型已实现边缘嵌入，如图像识别、人脸识别、简单交互，但这些算法都是基于CV算法、CNN算法，模型参数上限有限，非生成式transform架构；
- ◆ 未来大模型趋势将由作坊式向工业式转变，降低门槛。在医药研发、卫星遥感、灾害评估、自然生态监测等场景、医疗、金融、工业、教育等多个领域发挥作用。
- ◆ 目前实现大模型的运行需要具备三个要素：1、算力；2、带宽；3、内存。其中硬件是急需解决的瓶颈，在AI算力逐渐突破的背景下，所需硬件耗费将越来越多，难以实现对带宽和内存的支撑。

图表50 不同参数的大模型需求

模型参数	所需算力 (PFLOPS/day)	所需内存	所需带宽	所需硬件 (A100 GPU)
10亿	21	1GB	5.12GB/s	108张
30亿	63	3GB	15.36GB/s	325张
100亿	209	10GB	51.2GB/s	1083张
1750亿	3650	175GB	896GB/s	18953张

图表51 相关硬件价格

	骁龙8 Gen 2 CPU	1134元/颗
	英伟达A100 GPU	15万元/个
	美光GDDR6X 16G 内存	216元/颗

# 建议关注标的

证券代码	证券简称	总市值 (亿元)	EPS-2022A	EPS-2023E	EPS-2024E	PE-2022A	PE-2023E	PE-2024E
688080.SH	映翰通	34.22	1.06	1.41	1.89	48.65	33.22	24.06
300590.SZ	移为通信	55.22	0.36	0.49	0.65	33.38	23.98	18.24
688609.SH	九联科技	57.90	0.12	0.26	0.43	95.96	44.88	27.18
300353.SZ	东土科技	68.24	0.05	0.19	0.30	338.24	69.45	40.74
002881.SZ	美格智能	86.71	0.46	0.78	1.17	67.83	42.81	28.51
688018.SH	乐鑫科技	104.68	1.25	1.90	2.79	107.56	66.76	46.50
603236.SH	移远通信	150.68	1.38	3.20	4.22	24.19	18.09	13.22
300638.SZ	广和通	170.63	0.53	0.79	1.00	46.82	28.90	22.46
000810.SZ	创维数字	182.31	0.64	0.89	1.07	22.15	17.86	14.70
688521.SH	芯原股份	360.34	0.00	0.30	0.47	488.17	239.18	154.19
300496.SZ	中科创达	446.41	1.71	2.24	3.02	58.07	43.16	32.32

-  01 算力向推理和边缘扩散
-  02 产业链重构带来价值
-  03 车和垂直行业弹性爆发
-  04 投资逻辑和重点标的
-  05 投资建议与风险提示



- ◆ 边缘算力下，车和垂直行业弹性最具爆发性，我们认为，边缘预处理最先落地，大模型百花齐放参数不一、大模型转小模型是必须之路、模型嵌入所需硬件支撑——算力/带宽/内存；
- ◆ 建议重点关注：美格智能、移远通信、映翰通、东土科技、移为通信、三旺通信、广和通等。

- ◆ 边缘模型算法开发进展不及预期；随着chatGPT带动，生成式算法迭代加速，但主流大模型的模型参数达上千亿甚至万亿，如果想嵌入到边缘端，需要压缩至10亿参数以内，但会损失算法的稳定性甚至是精确度，开发难度大。
- ◆ 边缘应用场景需求不及预期；目前在边缘侧需要算法算力支持的场景主要有摄像头识别、自动货柜识别等等，还缺少如生成式模型落地的场景，需要不断拓展开发新需求。
- ◆ 芯片成本过高影响落地进度。目前高通、英伟达都推出面向边缘侧高算力的芯片，但价格偏高，导致最终产品售价过高，影响产品的普及性。

## 公司评级体系

### 收益评级：

- 买入 — 未来6个月的投资收益率领先沪深300指数15%以上；
- 增持 — 未来6个月的投资收益率领先沪深300指数5%至15%；
- 中性 — 未来6个月的投资收益率与沪深300指数的变动幅度相差-5%至5%；
- 减持 — 未来6个月的投资收益率落后沪深300指数5%至15%；
- 卖出 — 未来6个月的投资收益率落后沪深300指数15%以上。

### 风险评级：

- A — 正常风险，未来6个月投资收益率的波动小于等于沪深300指数波动；
- B — 较高风险，未来6个月投资收益率的波动大于沪深300指数波动。

## 行业评级体系

### 收益评级：

领先大市 — 未来6个月的投资收益率领先沪深300指数10%以上；

同步大市 — 未来6个月的投资收益率与沪深300指数的变动幅度相差-10%至10%；

落后大市 — 未来6个月的投资收益率落后沪深300指数10%以上；

### 风险评级：

A — 正常风险，未来6个月投资收益率的波动小于等于沪深300指数波动；

B — 较高风险，未来6个月投资收益率的波动大于沪深300指数波动。



## 分析师声明

李宏涛声明，本人具有中国证券业协会授予的证券投资咨询执业资格，勤勉尽责、诚实守信。本人对本报告的内容和观点负责，保证信息来源合法合规、研究方法专业审慎、研究观点独立公正、分析结论具有合理依据，特此声明。

## 本公司具备证券投资咨询业务资格的说明

华金证券股份有限公司（以下简称“本公司”）经中国证券监督管理委员会核准，取得证券投资咨询业务许可。本公司及其投资咨询人员可以为证券投资人或客户提供证券投资分析、预测或者建议等直接或间接的有偿咨询服务。发布证券研究报告，是证券投资咨询业务的一种基本形式，本公司可以对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向本公司的客户发布。

## 免责声明：

本报告仅供华金证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因为任何机构或个人接收到本报告而视其为本公司的当然客户。

本报告基于已公开的资料或信息撰写，但本公司不保证该等信息及资料的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映本公司于本报告发布当日的判断，本报告中的证券或投资标的价格、价值及投资带来的收入可能会波动。在不同时期，本公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，本公司将随时补充、更新和修订有关信息及资料，但不保证及时公开发布。同时，本公司有权对本报告所含信息在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以本公司向客户发布的本报告完整版本为准。

在法律许可的情况下，本公司及所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务，提请客户充分注意。客户不应将本报告为作出其投资决策的惟一参考因素，亦不应认为本报告可以取代客户自身的投资判断与决策。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议，无论是否已经明示或暗示，本报告不能作为道义的、责任的和法律的依据或者凭证。在任何情况下，本公司亦不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告版权仅为本公司所有，未经事先书面许可，任何机构和个人不得以任何形式翻版、复制、发表、转发、篡改或引用本报告的任何部分。如征得本公司同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“华金证券股份有限公司研究所”，且不得对本报告进行任何有悖原意的引用、删节和修改。

华金证券股份有限公司对本声明条款具有惟一修改权和最终解释权。

## 风险提示:

报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价或询价。投资者对其投资行为负完全责任，我公司及其雇员对使用本报告及其内容所引发的任何直接或间接损失概不负责。

华金证券股份有限公司

办公地址:

上海市浦东新区杨高南路759号陆家嘴世纪金融广场30层

北京市朝阳区建国路108号横琴人寿大厦17层

深圳市福田区益田路6001号太平金融大厦10楼05单元

电话: 021-20655588

网址: [www.huajinsc.cn](http://www.huajinsc.cn)