宝德自强AI产品

信创BU

宝德计算机系统股份有限公司

AI作为新的通用目的技术，将深刻推动社会发展进程

人工智能是一种新的通用目的技术（GPT）

General purpose technology

公元前9000~公元1000 农耕时代	驯化植物 驯养动物 冶炼矿石 轮子	写作 青铜 铁 水轮
15~18世纪 蒸汽机时代	三帆帆船 印刷	工厂体系 蒸汽机
19世纪 电力时代	铁路 蒸汽船	内燃机 电力
20世纪 互联网时代	汽车 飞机 大规模生产 精益生产	**电脑** **互联网** 生物技术
21世纪 智能时代	商业虚拟化 纳米技术	**人工智能**

开始爆炸式增长

人工智能应用与生产力发展曲线

阶段1
技术与应用
局部探索

阶段2
技术发展与社会
环境相互碰撞

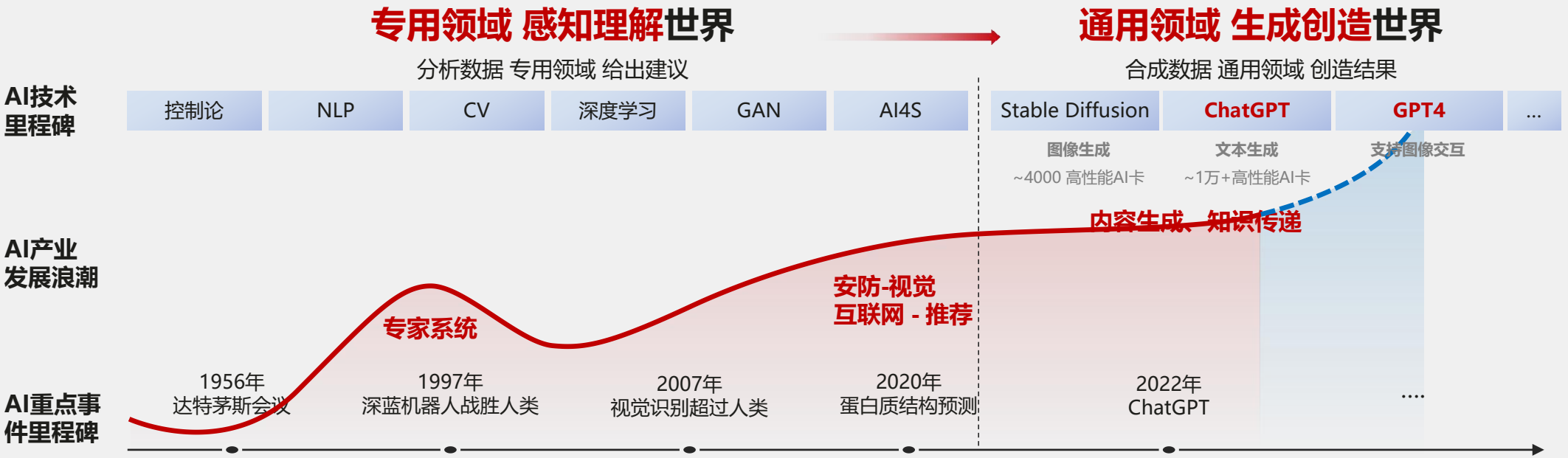
阶段3
技术发展与社会
环境相互促进

阶段4
稳定发展
期待新GPT

2025年AI产业市场规模
7.44万亿美金
未来五年年均增长30%+

2030年AI可使
全球GDP每年增长 1.2%
中国GDP每年增长 1.6%

AI发展跨越拐点， AIGC即将重塑X万亿元级产业格局



感知AI技术，撬动万亿级机会

7000亿元级

机器视觉

- 催生智能安防产业
- 海康、大华等千亿级企业

万亿元级

内容推荐

- 颠覆传媒/内容传播行业
- 字节跳动等7000亿级企业

认知AI技术，撬动X万亿级机会

X万亿元级

AIGC/AGI

- 重塑1.5万亿元级搜索市场
- 颠覆万亿元级数字传媒行业

大模型促进行业应用在2C规模落地、2B逐渐起步

2C类高价值应用快速落地
GPTs推动大模型商业模式趋于成熟

聊天类应用	智慧办公类	文生图应用	自动驾驶
			
ChatGPT月 活 用 户 数	微 软 Copilot付 费 用 户	Midjourney用 户 数	国内L3试点上路许可
1亿+	100万+	1500万+	宝马、奔驰、极狐、长安 阿维塔、智已、问界等

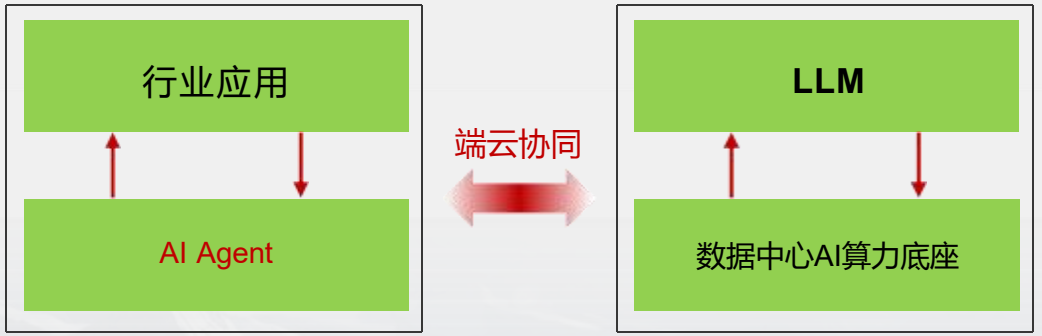
GPTs极低开发门槛，收入分成模式实现商业闭环

每分钟产生一个GPT应用， 3天全网达到 3000+

各企业均围绕自研大模型布局2B类应用
AI Agent推动大模型深入行业场景

阿里云	百度	智谱AI	Open AI
通义 大模型	文心一言 大模型	GLM 大模型	Sora 大模型
聚焦5大行业	聚焦8大行业	聚焦5大行业	聚焦5大行业
医疗、司法、金融 客服、写作	能源、金融、航天、制造 传媒、城市、社科、影视	金融、医疗、能源 制造、政务	影视制作、虚拟现实、 教育、娱乐、医疗

AI Agent实现高自动化执行和处理专业或繁复的工作任务



大模型能力快速提升，引爆AI应用，激发产业智能跃迁

大模型加速渗透垂直场景，引爆AI应用

聊天类应用	智慧办公类	文生图应用	自动驾驶
			
ChatGPT月活用户数	微软Copilot付费用户	Sora 横空出世，带来互联网短视频、电影工业、广告营销的产业变革	国内L3试点上路许可
1亿+	100万+		宝马、奔驰、极狐、长安阿维塔、智已、问界等

AI+对话

- OpenAI GPT
- DeepMind Gopher
- facebook OPT
- Hugging Face BLoom
- COHERENT. Cohere
- ...

AI+办公

- OpenAI GPT
- tabnine
- stability ai
- Microsoft Copilot
- ...

AI+图像视频

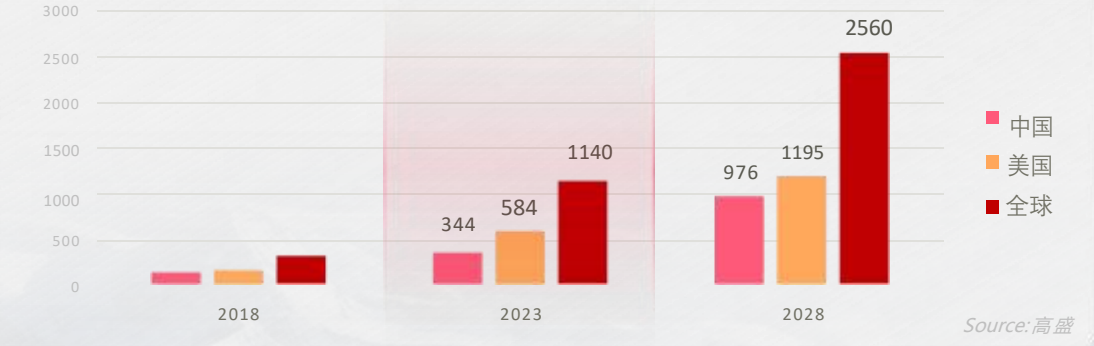
- OpenAI Dall-E 2
- Stable diffusion
- craiyon
- Microsoft X- CLIP
- Meta Make- A- Video
- ...

各企业均围绕自研大模型深入金融行业场景

工商银行	交通银行	北京银行	太平洋保险
客服	客服	数字员工	赋能培训
办公	运营	决策	投资助理
数据分析	风控	风控	承保核保

AI市场空间：未来5年全球投资超2500亿美元

单位：亿美金



头部金融机构将人工智能作为未来科技创新的重点方向

六大行

部署千卡集群打造多维应用场景

千亿参数 千卡集群	ICBC 中国工商银行	千亿参数 千卡集群	中国建设银行 China Construction Bank
全行智能助手		金融行业大模型	
千亿参数 千卡集群	邮储银行 POSTAL SAVINGS BANK	千亿参数 千卡集群	交通银行 BANK OF COMMUNICATIONS
邮储大脑		数字员工	

股份制及城商农信

百卡到千卡集群开源模型先行试点

千亿参数 千卡集群	招商银行 CHINA MERCHANTS BANK	百亿参数 百卡集群	广发银行 CGB
智能数字人		试点应用中	
十亿参数 百卡集群	北京银行 BANK OF BEIJING	百亿参数 百卡集群	中信银行 CHINA CITIC BANK
京智大脑		应用试点	

保险：百卡到千卡集群

聚焦降本增效场景

千亿参数 千卡集群	太平洋保险 CPIC
数字劳动力	
百亿参数 百卡集群	中国平安 PING AN
智能核保	

证券：百卡集群

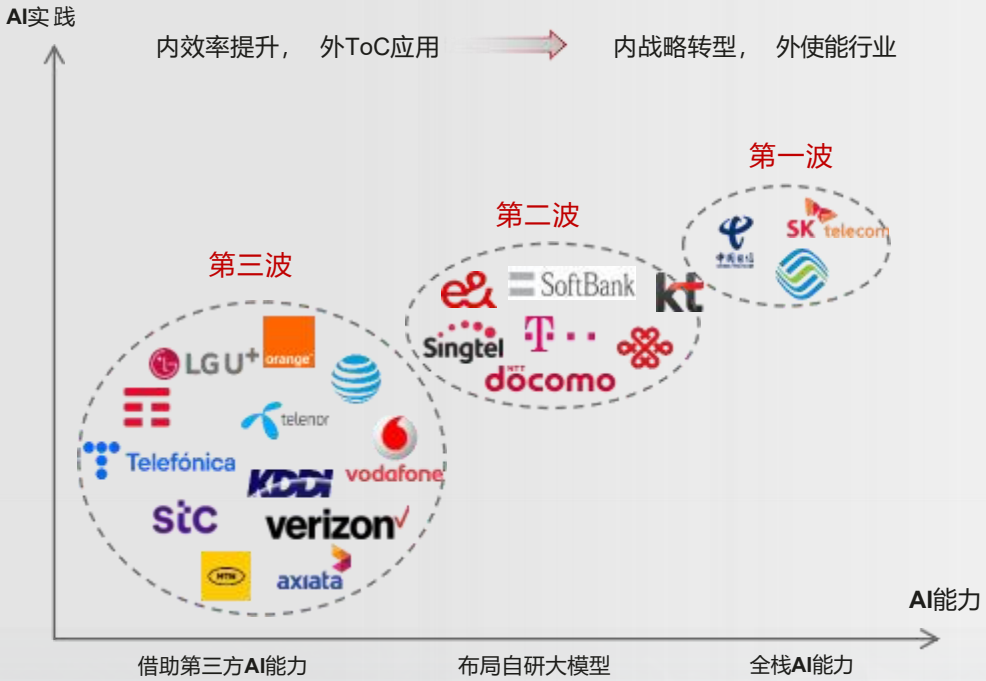
聚焦客服和代码

百亿参数 百卡集群	中金财富 CICC Wealth Management
投资者教育	
百亿参数 百卡集群	CMS 招商证券
智能问答	

智能风控	智能营销	智能投研	办公助手	智能客服	智慧运营
财务异常分析	营销话术生成	摘要生成	会议/日志	坐席辅助	网点助手
舞弊动机识别	营销物料生成	风险传导	智能助手	聊天机器人	数据分析助手
违约风险分析	实时互动数字人客服	观点提取	客服/培训	智能外呼	产品设计助手

运营商加大AI战略投入，在新产品、新方案持续创新

全球运营商形成三波AI阵营， 第一波正在构建全栈AI能力



尝试入局

借第三方能力尝试发展AI业务
（欧美大T，亚非领先运营商…）

积极布局

有战略规划； 连接+AI
（运营商AI联盟成员、日韩）

智化先锋

全栈AI能力； AI+千行百业
（电信、移动、 SKT）

* From Huawei Source

海外多家运营商引入AI，提升现有业务效率

AI战略 德电/SKT/KT

基于数据主权，做AI的Shaper
AI能力中心， 500+AI专家
独立自主开发电信大模型

AI金字塔战略，围绕AI Infra、AI转 型（AIX）、 AI应 用（A.）加速向全球AI公司转型

向AI MSP转型（模型&管理服务提供商）

其他

SKT、 e&、新电、DT、软银宣布成立打造电信行业大语言模型的合资企业，通过数字助理和聊天机器人改善客户互动

AI业务 美/欧/日/韩

通过AI降成本，机器人自动化RPA。 23年推出146个AI应用， GenAI工具提升效率

依托微软和Open AI，打造智能BSS和OSS系统，包括计费、客服、数据分析工具

基于基础大模型，进行调优，打造电信大模型，聚焦客服场景

同AWS建立伙伴关系，用AWS Bedrock为企业客户开发AIGC云服务

Microsoft 大模型， AT&T 知识体系，用于客服系、员工办公服务、优化软件代码

将Azure AI Studio 集成到Telefónica Kernel 2.0，GenAI应用于内部关键工作流程优化，增加客户体验

AI使能6G应用， 语音信息传递走向感官信息传递， Feel Tech互动（AI触觉/味觉模拟），探索多感知需求

AI改变生活，如用AI实现元宇宙和Web3的“αU”业务

AI终端 手机/PC/穿戴

AI + 手机

-> 内置大模型的手机成主流

小米14、三星Galaxy S24、荣耀Magic6…

->德电： 两种选择

1）同Brain. AI 推出AI运行在云端的T-Phone
2）内置高通骁龙8 Gen3的手机直接部署大模型

AI + PC

HP X360 14 内置AI引擎

联想CreatorZone AI交互

AI + 穿戴

三星 Galaxy Ring

OPPO AI眼镜连接大模型

昇腾持续打造极致性能、极简易用的全场景人工智能平台

Ascend

行业应用	运营商、互联网、能源、金融、交通、制造、医疗等行业应用				
应用使能	ModelArts		HiAI Service		第三方平台
	MindX 昇腾应用使能	MindX SDK 领域套件			
		MindX DL	MindX Edge	MindSpeed	MindIE
AI框架	[M]S 昇思 MindSpore			TensorFlow/PyTorch 等第三方框架	
异构计算架构	CANN				

全流程开发工具链

Mindstudio

Atlas系列硬件


Atlas 200I A2
AI加速模块


PI300T G2
智能小站


Atlas 300V


Atlas 300I/V Pro


Atlas 300I Duo
推理卡


PR420KI G2


PR205KI智能
边缘服务器


PR210KI系列
推理服务器


PRA100 AI 集群

宝德 PowerLeader

自强 ZIQIANG

宝德自强AI服务器产品整体规划

宝德自强人工智能全产品线

液冷智能计算集群

PRA100 PoD G1/G2

PR420KI G1/G2

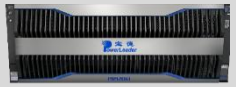
中心推理

PR2715E

PR2715W3

PR4908E/PR4910E

PR4910P2



PR410EI(新产品)

PR210KI

PR410KI(新产品)

PR425KI G1/G2

边缘推理服务器

PR205KI

智能小站

PI300T

PI300T G2

宝德自强AI产品（PR420KI G2）

8 * NPU

主板贴片

PR420KI G2

上市信息 已上市　　**当前状态** **可销售**

应用场景 面向运营商/互联网/大模型等市场，适于AI大模型训练

运营商

互联网

大模型

关键特性	规格描述	
形态	4U机架服务器（175mm × 447mm × 790mm）	
CPU	4 * 鲲鹏920	
NPU	8 *AI处理器 ***（具体性能请联系宝德客户经理）***	
AI算力	半精度（FP16）**XXX** PFLOPS	单精度（FP32）**XXX**PFLOPS
内存	8 * 64G HBM；支持32个DDR4内存插槽	
内部拓扑	NPU HCCS全互联，互联带宽392GB/s	
网络接口	NPU直出8 * 200G RoCE	
散热	风冷散热	

宝德自强AI高密算力平台，支撑大模型系统工程建设

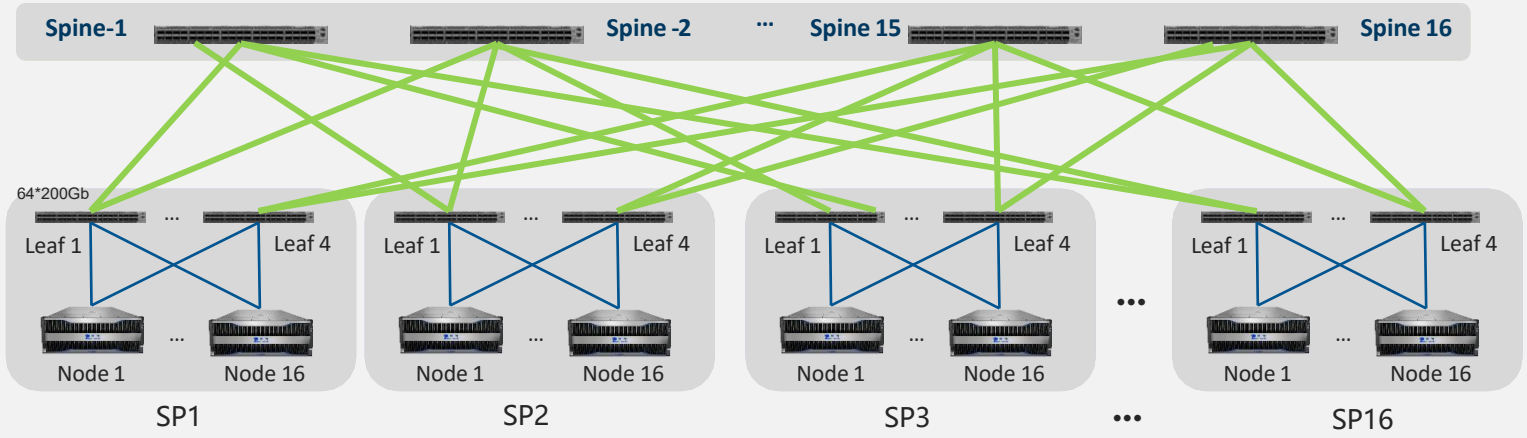
PR420KI G2

CPU	4 x Kunpeng920处理器
AI算力	（FP32）xxx FLOPS （FP16）xxx FLOPS
HBM	8*60G，1.6TB/s
节点内互联	392GB/s
跨节点互联	8*200G Roce

PR410EI

CPU	2 x Sapphire Rapids处理器
AI算力	（FP32）xxx FLOPS （FP16）xxx FLOPS
HBM	8*60G，1.6TB/s
节点内互联	392GB/s
跨节点互联	8*200G Roce

Pod集群组网



系统参数	性能
峰值AI算力	641PFlops FP16
跨结点带宽	跨POD任意两节点互联带宽200GB/s
扩展性	16个POD，2048颗NPU

以256台AI服务器为例，共计2048颗NPU芯片
分成16组，每组16台服务器，里面放置一个完整的模型，每组之间采用数据并行；每组16台服务器，采用模型并行，每台服务器放1/16的模型；

宝德自强中心推理集群产品（PR425KI G2）

8 * NPU

主板贴片

PR425KI G2

上市信息 已上市　　**当前状态** 可销售

关键特性	规格描述	
形态	4U机架服务器（175mm × 447mm × 790mm）	
CPU	4 * 鲲鹏920	
NPU	8 * AI处理器 ***（具体性能请联系宝德客户经理）***	
AI算力	半精度（FP16）XX PFLOPS	单精度（FP32）XX PFLOPS
显存	8*64G / 8*32GB HBM；支持32个DDR4内存插槽	
内部拓扑	NPU HCCS全互联，互联带宽392GB/s	
网络接口	NPU直出8 * 200G RoCE	
散热	风冷散热	

应用场景　面向互联网/大模型/科研教育等市场，全面适配大模型微调、集群推理；

互联网

科研教育

行业大模型

PR425KI G2主要面向大模型集群推理

同时也可以支持大模型集群的训练任务

宝德自强昇腾推理产品（PR210KI/PR215PI）

支持 1 ~ 8 张卡

Atlas 300I
Atlas 300I Pro
Atlas 300V Pro

支持 1 ~ 4 张卡

Atlas 300I DUO 推理卡

插卡式

自强 PR210KI 推理服务器

支持 1 ~ 7 张卡

Atlas 300I
Atlas 300I Pro
Atlas 300V
Atlas 300V Pro

插卡式

自强 PR215PI 推理服务器

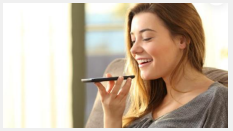
上市信息 已上市　　**当前状态** 可销售

应用场景 部署于数据中心机房中，使能AI中心推理

搜索推荐

金融大脑

语音识别



内容审核

关键特性	规格描述	
型号	PR210KI	PR215PI
形态	2U机架服务器 86.1mm × 447mm × 790mm	2U机架服务器 86.1mm × 447mm × 748/708mm
CPU	2 x 鲲鹏920	1/2个Intel® Xeon® SP Skylake 或 Cascade Lake处理器
内存	最多32个DDR4内存插槽	最多24个DDR4内存插槽
AI加速卡	最大支持 **8** 张Atlas 300I、300I Pro/300V Pro，或 **4** 张Atlas 300I DUO	最大支持 **7** 个Atlas 300I、300IPro/300V Pro
散热	风冷散热	

宝德自强昇腾边缘产品PR205KI：满足“短机柜”部署

Atlas 300I

Atlas 300I Pro

Atlas 300V

Atlas 300V Pro

支持 1~3 张卡

插卡式

*无Riser卡的情况

“短机柜”

室外电信设备机柜（III、IV）

典型尺寸深度 **475mm**

PR205KI 智能边缘服务器

关键特性	规格描述
形态	2U服务器，短机箱（86.1mm x 447mm x **475mm**）
AI加速卡	最大支持 **3张** Atlas 300I/V Pro 推理卡
CPU	1 x 鲲鹏920 （单路24核）
内存	4个DDR4内存插槽，最高3200 MT/s
AI算力	最大 **420** TOPS INT8 或 **384** 路1080P 30FPS视频解析（硬件解码能力）

应用场景

边缘侧独立部署，使能智能边缘

大型园区、电力、交通、商超等场景

智慧交通

湖南高速

变电站智能巡视

国网、南网

智慧加油站

加油站生产监测

1 稳定算力高

PR205KI VS 主流产品 **1.6X** ↑

TOPS (INT8)

2 能效比高

PR205KI VS 主流产品 **2.6X** ↑

TOPS/W (INT8)

3 视频解码能力强

PR205KI VS 主流产品 解码路数 **2.4X** ↑

4 典型模型性能好

1.9X ↑ **业界友商** images/s

1.3X ↑ **业界友商** sentences/s

Atlas 200I DK A2：开箱即用、参考丰富的开发者套件

Atlas 200I DK A2
开发者套件

上市信息 已上市　　**当前状态** 可销售

Atlas 200 DK 演进一代产品

早期小批量发货

关键特性	规格描述
形态	135mm x 120mm x 44mm
AI算力	整数精度（INT8）：**8** TOPS　半精度（FP16）：**4** TFLOPS
摄像头接口	2*MIPI-CSI 支持两个树莓派摄像头
USB接口	1*USB TypeC 仅支持从模式；　2*USB3.0 Type-A
以太网接口	2*RJ45千兆网口
编解码能力	内置DVPP预处理单元　图片JPEG/PNG编解码能力 视频 **16路** 1080P 30FPS
功耗	24 W
存储设备	1*NVMe/SATA M.2 SSD；　1*Micro SD卡
图形显示	2* 4K分辨率HDMI视频输出

应用场景 昇腾AI开发者上手学习，实践创新场景，提供配套软硬件

AI应用创新
机械人/智能小车...

高校教学
智能基座

行业算法验证
工业互联网等ISV

1 丰富代码示例

- 3大典型场景示例，覆盖开发者80%应用场景

智能车　机械臂　语音交互

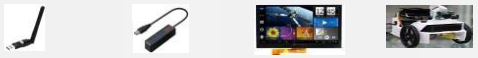
2 开源预训练模型库

- **ModelZoo**：900+高性能预训练模型，CV/NLP/语音等

3 专业认证硬件配件

认证配件，降低选型兼容难度

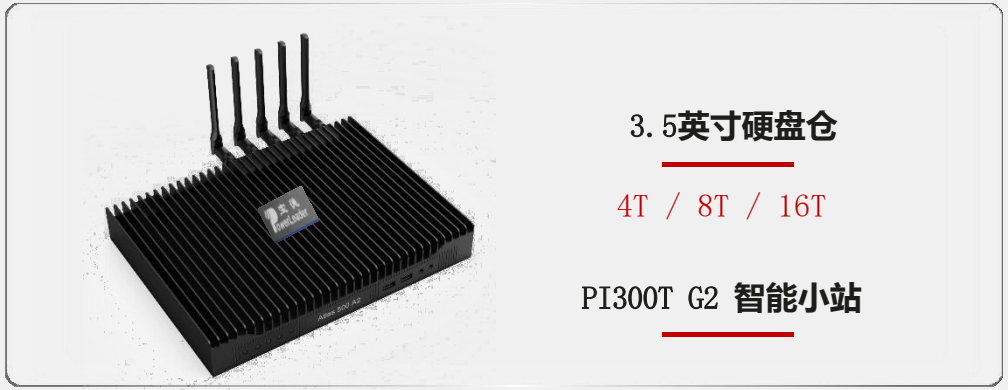
连接扩展　接口扩展　组件扩展　场景扩展



4 高效、易用开发工具

- **一键制卡工具**：一键安装镜像，30min环境搭建
- **模型适配工具**：Windows下的端侧模型适配工具，支持用户完成端侧模型适配开发全流程

PI300T G2：边缘宽温部署，视频智能分析利器



3.5英寸硬盘仓

4T / 8T / 16T

PI300T G2 智能小站

上市信息 已上市　　**当前状态** 可销售

PI300T 演进一代产品

应用场景 满足严苛的边缘部署场景设计，在智慧城市、交通、社区、园区、商场、超市等复杂环境区域应用

智慧网点	自由流收费	智慧园区	智慧加油站
招行、工行	全国ETC收费站	商汤	江苏石化



关键特性	规格描述
尺寸（长x宽x高）	无盘配置：290 mm x 220 mm x 44 mm 有盘配置：410 mm x 220 mm x 44 mm
内存	LPDDR4X，12GB / 4GB； 总带宽51.2 GB/s
AI算力	整数精度（INT8）：**20**TOPS　半精度（FP16）：**10**TFLOPS
编解码能力	内置DVPP预处理单元　图片JPEG/PNG编解码能力 视频 **40路** 1080P 30FPS
重要接口	2*USB3.0；**5***RJ45千兆网口；1* M.2 KEYB（可接**5G**模组）； 4*DI / 4*DO；（选配：1*MicroSD卡槽；1*M.2 NVMe SSD）
媒体	2路HDMI图片输出，满足现场结果直接显示
典型功耗	无盘 32.3W / 有盘 44.5 W
环境条件	-40℃～ +60℃

优势

- 重点引导视频分析能力强、接口丰富适应更多边缘场景

1 视频分析能力强 1080P 30FPS

PI300T G2 VS 友商

最大40路 VS 最大32路

1.25X

2 接口更丰富

PI300T G2		友商
5路	千兆网口	1路
2路	HDMI	1路

宝德自强昇腾智算中心集群解决方案

Ascend

应用层

训练推理　远程协作　自然语言　数据分析　机器学习　图像处理

AI计算加速平台

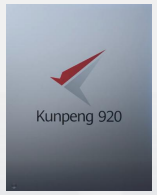
异构计算平台CANN

全流程开发工具链MindStudio　AI框架昇思MindSpore　昇腾推理引擎MindIE

AI资源管理调度平台

PLStack　分布/并行计算优化　异构资源调度与编排　自动化交付

AI基础设施平台

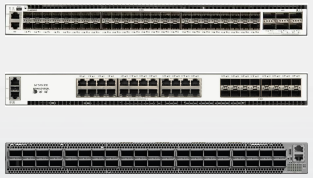


Kunpeng

Ascend NPU

宝德通用服务器

宝德昇腾服务器



宝德交换机

宝德 PowerLeader

自强 ZIQIANG

宝德智算平台 PLStack 特色功能 – 管理服务

平台管理

- **多租户**管理、多种计量计费模式；
- 集群资源管理：以**套餐方式**进行资源配额；
- 自动化运维工具：集群节点扩缩容；
- 集群监控与告警：安全管理和完善的日志审计功能。

GPU卡管理

- **GPU卡四种模式：**独享/共享/vGPU/MIG分配模式；
- **vGPU：**将GPU卡切分出多张小的vGPU卡，对其显存和算力进行限制，提高物理GPU卡利用率；
- **虚拟显存：**物理显存不足时，通过对显存做虚拟化处理，使得可用显存超过物理显存，从而支持大批量、大模型的训练任务。

多数据中心管理

- 支持将多个物理区域的GPU资源**统一纳管**；
- 统一对多个区域资源使用监控、计量计费等管理；
- 用户可选择不同区域的资源并调用；
- **优化成本**：降低对运维人员成本投入。

数据管理

- **在线标注工具**：使用标注数据进行模型开发、训练、预测；
- **存储管理**：持久化存储工作目录、可视化文件管理系统、共享存储；
- **存储性能、安全性**：分布式文件存储、将本地硬盘组建分布式存储。

宝德智算平台 PLStack 特色功能 – 模型开发训练

开发环境

- **一键式环境生成**，集成数十种集成主流AI框架，如TensorFlow、Pytorch、PaddlePaddle等，支持自定义框架镜像；
- **Mlab**交互式开发工具：兼容Jupyter；
- 定时快照/备份，数据快速回滚；
- 弹性伸缩，资源/框架/存储弹性变更。

模型训练

- 分布式训练：深度集成Horovod、Ray的分布式并行训练，支持**多机多卡，单机多卡**；
- 参数调优：内置Auto ML自动调优，提高模型训练效率；
- 训练可视化，训练过程中，实时输出资源的利用率和模型训练日志。

模型管理

- 集成行业预训练模型和行业数据集，降低用户模型开发难度；
- 多版本模型管理：训练模型和本地模型一键导入；
- **格式转换**：模型文件支持一键转换为ONNX格式；
- 云端服务：模型发布为云端服务能力，并对外提供http协议访问接口。

模型服务

- 模型在线推理应用：直接上传推理文件进行**在线模型推理**；
- **远程调用**：发布后的模型服务提供对外的Web接口和Token秘钥，实现模型的真正应用。
- 数据回传：算法应用后的数据资源可反哺数据存储中，方便后续持续训练更新；

宝德 PowerLeader

自强® ZIQIANG

谢谢

国之重器 强者自强